Online Diskriminanzanalyse für Datensituationen mit Concept Drift

DISSERTATION

zur Erlangung des akademischen Grades
eines *Doktors der Naturwissenschaften*
der Technischen Universität Dortmund

Der Fakultät Statistik
der Technischen Universität Dortmund
vorgelegt von

Sarah Anna Schnackenberg

Dortmund, Januar 2020

VORGELEGT
27. Januar 2020

TAG DER MÜNDLICHEN PRÜFUNG
10. Juni 2020

PRÜFUNGSKOMMISSION
1. Gutachter: Dr. Uwe Ligges
2. Gutachter: Prof. Dr. Claus Weihs
Kommissionsvorsitz: Prof. Dr. Jörg Rahnenführer

Inhaltsverzeichnis

Notationsverzeichnis

In der folgenden Tabelle sind Symbole und Parameter inklusive einer kurzen Beschreibung aufgelistet. Die Spalte „Erklärung/Erstmals“ verweist auf die entsprechende Seite, auf welcher das Symbol bzw. der Parameter das erste Mal verwendet bzw. eingeführt wird. Es folgen zunächst griechische, dann lateinische Buchstaben.

Symbol	Bedeutung	Erklärung/Erstmals
$\boldsymbol{\alpha}$	Diskriminanzkomponente in der Fisher LDA	S. 46
$\boldsymbol{\alpha}_j$	j-te Diskriminanzkomponente in der Fisher LDA	S. 49
$\alpha_{\max}$	oberer Schwellenwert für Schrittweite $\alpha_t^{(c)}$	S. 82
$\alpha_{\min}$	unterer Schwellenwert für Schrittweite $\alpha_t^{(c)}$	S. 82
$\alpha_t^{(c)}$	Schrittweite beim Gradientenabstiegsverfahren	S. 81
$\boldsymbol{\beta}_0^{(c)},\ \boldsymbol{\beta}_1^{(c)}$	Parameter des linearen Trends der Erwartungswertvektoren in Klasse c	S. 138
$\boldsymbol{\beta}_{0t}^{(c)},\ \boldsymbol{\beta}_{1t}^{(c)}$	Parameter des lokalen linearen Regressionsmodells der Mittelwertvektoren aus Klasse c	S. 162
$\hat{\boldsymbol{\beta}}_{0t}^{(c)},\ \hat{\boldsymbol{\beta}}_{1t}^{(c)}$	entsprechende KQ-Schätzer	S. 165/169
$\boldsymbol{\beta}_j^{(c)}$	Parameter des linearen Modells für eine einzelne Dimension j	S. 162
$\hat{\boldsymbol{\beta}}_j^{(c)}$	entsprechender KQ-Schätzer	S. 168
$\gamma_j,\ j=1,\ldots,p$	Eigenwerte der Kovarianzmatrix $\boldsymbol{\Sigma}$	S. 48
δ	Faktor für Initialisierung bei *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 96
$\Delta_{t,L}$	Fehlerdifferenz zur Adaption der Lernrate bei *OLDC*	S. 75
$\boldsymbol{\epsilon}_i^{(c)}=(\epsilon_{i1}^{(c)},\ldots,\epsilon_{ip}^{(c)})^T$	Fehler des lokalen linearen Regressionsmodells	S. 162
$\boldsymbol{\epsilon}_j^{(c)}$	Fehler des linearen Modells für eine einzelne Dimension j	S. 162
$\tilde{\boldsymbol{\theta}}_t^{(0)}$	Vektor aller Größen für den Algorithmus *M-AF*	S. 83
$\tilde{\boldsymbol{\theta}}_t^{(c)}$	Vektor aller Größen aus Klasse c für den Algorithmus *G-AF*	S. 82
$\tilde{\boldsymbol{\theta}}_t^{(P)}$	Vektor mit Größen für den Algorithmus *LDA-AF*	S. 86/88
λ	Lernrate bei *OLDC*	S. 72
$\lambda_j,\ j=1,\ldots,r$	Eigenwerte von $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ bzw. $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ bei Fisher LDA	S. 48
λ_-	unterer Schwellenwert für Faktor bei *QDA-AF*/*LDA-AF*	S. 81
λ_+	oberer Schwellenwert für Faktor bei *QDA-AF*/*LDA-AF*	S. 81
$\lambda_j^{(0)}$	Faktor (für Gewichtung) für Likelihood Term bei *M-AF*	S. 83
$\lambda_{(j)}^{(c)}$	Faktor (für Gewichtung) in Klasse c für Likelihood Term bei *G-AF*	S. 78
$\lambda_t^{(P)}$	Faktor bei zusätzlichem *G-AF* Durchlauf für *LDA-AF*	S. 87
λ_{start}	initiale Lernrate bei *OLDC* mit adaptiver Lernrate	S. 226
$\boldsymbol{\mu}^{(c)}$	Erwartungswertvektor von Klasse c	S. 32
$\bar{\boldsymbol{\mu}}$	Mittelwert der Erwartungswertvektoren über alle Klassen	S. 46
$\hat{\boldsymbol{\mu}}^{(c)}$	Schätzer für Erwartungswertvektor von Klasse c	S. 42
$\hat{\bar{\boldsymbol{\mu}}}$, $\hat{\bar{\boldsymbol{\mu}}}_{\text{gew}}$, $\hat{\bar{\boldsymbol{\mu}}}_{\text{gew},2}$	Schätzer für den Mittelwert $\bar{\boldsymbol{\mu}}$ der Erwartungswertvektoren über alle Klassen	S. 52
$\boldsymbol{\mu}_i^{(c)}$	Erwartungswertvektor von Klasse c zum Zeitpunkt i	S. 59
$\boldsymbol{\mu}_{\mathcal{Y}}^{(c)}$	Erwartungswert der Transformation von $\boldsymbol{X}^{(c)}$ bei Fisher LDA	S. 46

Fortsetzung auf der nächsten Seite

Symbol	Bedeutung	Erklärung/Erstmals
$\bar{\boldsymbol{\mu}}_{\mathcal{Y}}$	Mittelwert von $\boldsymbol{\mu}_{\mathcal{Y}}^{(c)}$ über alle Klassen	S. 46
$\bar{\boldsymbol{\mu}}_{\text{gew}}$	gewichtete Variante des Erwartungswertvektors	S. 47
$\bar{\boldsymbol{\mu}}_{\mathcal{Y}_{\text{gew}}}$	Transformation der gewichteten Variante des Erwartungswertvektors bei Fisher LDA	S. 47
$\boldsymbol{\nu}_j,\ j = 1, \ldots, r$	normalisierte Eigenvektoren von $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ bei Fisher LDA	S. 49
$\boldsymbol{\xi}_t$	Differenz aus Beobachtung und Mittelwertvektor: $\boldsymbol{x}_t - \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}$ bei *LDA-AF*	S. 87
$\hat{\boldsymbol{\Pi}}_t^{(c)}$	Teil des ML-Schätzers für Kovarianzmatrix in Klasse c bei der *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 78
$\tilde{\boldsymbol{\Pi}}_t^{(c)}$	Teil des Schätzers für Kovarianzmatrix in Klasse c bei der *Online Diskriminanzanalyse mit adaptivem Vergessen* im Falle von exponentiellem Vergessen	S. 78
$\tilde{\boldsymbol{\Pi}}_t^{(P)}$	Parameter bei *LDA-AF*	S. 88
$\boldsymbol{\Sigma}$	Kovarianzmatrix	S. 38
$\boldsymbol{\Sigma}_0$	Parameter bei *LDA-AF*	S. 86
$\hat{\boldsymbol{\Sigma}}$	Schätzer für $\boldsymbol{\Sigma}$ (gepoolte Kovarianzmatrix)	S. 42
$\boldsymbol{\Sigma}_{\text{gew}}$	gewichtete Kovarianzmatrix innerhalb der Gruppen	S. 47
$\hat{\boldsymbol{\Sigma}}_{\text{gew}}$	entsprechender Schätzer	S. 42
$\boldsymbol{\Sigma}^{(c)}$	Kovarianzmatrix in Klasse c	S. 32
$\hat{\boldsymbol{\Sigma}}^{(c)}$	Schätzer für Kovarianzmatrix $\boldsymbol{\Sigma}^{(c)}$ aus Klasse c	S. 42
$\boldsymbol{\Sigma}_i^{(c)}$	Kovarianzmatrix in Klasse c zum Zeitpunkt i	S. 59
$\hat{\boldsymbol{\Sigma}}_t^{(c)}$	ML-Schätzer für Kovarianzmatrix in Klasse c bei der *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 78
$\tilde{\boldsymbol{\Sigma}}_t^{(c)}$	Schätzer für Kovarianzmatrix in Klasse c bei der *Online Diskriminanzanalyse mit adaptivem Vergessen* im Falle von exponentiellem Vergessen	S. 79
$\tilde{\boldsymbol{\Sigma}}_t^{(P)}$	Parameter bei *LDA-AF*	S. 86/88
τ	Anzahl Chunks bei Erweiterung der *Diskriminanzanalyse mit adaptivem Vergessen*	S. 121
$\boldsymbol{\Psi}$	Kovarianzstruktur der Fehler im Regressionsmodell	S. 166
$\boldsymbol{A}$	Matrix der Diskriminanzkomponenten bei Fisher LDA	S. 46
A	Hilfsvariable in Beweis	S. 194
A_i	Hilfsvariable in Beweis	S. 186
$\boldsymbol{B}$	Zwischen-den-Klassen Kovarianzmatrix	S. 47
$\hat{\boldsymbol{B}}$	Schätzer für Zwischen-den-Klassen Kovarianzmatrix	S. 52
$\boldsymbol{B}_{\text{gew}}$	gewichtete Zwischen-den-Klassen Kovarianzmatrix	S. 47
$\hat{\boldsymbol{B}}^{*}$, $\hat{\boldsymbol{B}}^{*}_{\text{gew}}$, $\hat{\boldsymbol{B}}_{\text{gew},2}$, $\hat{\boldsymbol{B}}_{\text{gew}}$	weitere Schätzer für Zwischen-den-Klassen Kovarianzmatrix	S. 52/53
B_i	Kurzform für $\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}}$	S. 140
$\boldsymbol{B}_t$	Zwischen-den-Klassen Kovarianzmatrix zum Zeitpunkt t	S. 61
$\tilde{\boldsymbol{B}}_t$	$\boldsymbol{B}_t$ ohne Vorfaktor	S. 61
$c = 1, \ldots, M$	Anzahl an Klassen	S. 9
c_i	Klassenausprägung von i-ter Beobachtung	S. 59/82
$\tilde{c}_t$	prognostizierte Klasse durch *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 86
$C(.,.)$	Kostenfunktion	S. 34
d	Teil der Dimension von $\boldsymbol{U}$ und $\boldsymbol{V}$ bei Chunk Erweiterung von *OLDC* (Rang-d-Update)	S. 98
d_0	Parameter bei *LDA-AF*	S. 87
$d_t^{(c)}$	Parameter bei *G-AF* (und somit *QDA-AF* und *LDA-AF*)	S. 85
$d_t^{(P)}$	Parameter bei *LDA-AF*	S. 87
$d(.,.)$	euklidischer Abstand in Simulationsstudie	S. 233
$\boldsymbol{D}^{(c)}$, $\boldsymbol{E}^{(c)}$, $\boldsymbol{F}^{(c)}$	Teile der aktualisierten Kovarianzmatrix bei *Chunk ILDA*	S. 68
$\boldsymbol{e}_1$	Einheitsvektor mit 1 in erster Dimension	S. 222
err_{Bayes}	Bayesfehler	S. 36

Fortsetzung auf der nächsten Seite

Symbol	Bedeutung	Erklärung/Erstmals
$f^{(c)}(\boldsymbol{x})$	Dichtefunktion von Klasse c	S. 32
$f_i^{(c)}$	Dichtefunktion der klassenbedingten Verteilung der Zufallsvektoren in Klasse c zum Zeitpunkt i	S. 59
$f_{\boldsymbol{X}}(\boldsymbol{x})$	Dichtefunktion der Verteilung von $\boldsymbol{X}$	S. 9
$f_{\boldsymbol{X},Y}(\boldsymbol{x}, y_c)$	Dichtefunktion der gemeinsamen Verteilung von $\boldsymbol{X}$ und Y	S. 9
$f_{\boldsymbol{X}\|Y=y_c}(\boldsymbol{x}\|y_c)$	klassenbedingte Wahrscheinlichkeitsdichtefunktion	S. 9
$\text{fix}_{\boldsymbol{m}}$	logischer Wert bei Algorithmus *LDA-AF*	S. 82
$\text{fix}_{\boldsymbol{\Sigma}}$	logischer Wert bei Algorithmus *LDA-AF*	S. 82
F_c	Verteilung von Klasse c	S. 32
$g(.)$	Funktion, die beschreibt, welche der M Klassen auftritt	S. 32
G	Klassifikationsregel	S. 35
$\boldsymbol{G}_0$	Parameter bei *LDA-AF*	S. 87
G_{Bayes}	Bayesregel	S. 34
G_{ML}	Maximum-Likelihood-Regel	S. 33
$\boldsymbol{G}_t^{(c)}$	Parameter bei *G-AF* (und somit *QDA-AF* und *LDA-AF*)	S. 85
$\boldsymbol{G}_t^{(P)}$	Parameter bei *LDA-AF*	S. 87
$h_F^{(c)}$, $h_{F,2}^{(c)}$	Diskriminanzfunktion bei Fisher LDA	S. 50/51
$h_L^{(c)}$, $h_{L,2}^{(c)}$	Diskriminanzfunktion bei Kanonischer LDA	S. 40/41
$\hat{h}_{L,t}^{(c)}(\boldsymbol{x})$	Schätzer für Diskriminanzfunktion (*OLDC*)	S. 75
$h_Q^{(c)}$, $h_{Q,2}^{(c)}$	Diskriminanzfunktion bei QDA	S. 44
I	Intervall für lokale lineare Regressionsmodelle	S. 162
$\boldsymbol{I}_{p\times p}$	Einheitsmatrix der Dimension p	S. 56
$J_t^{(0)}$	negative log-Likelihood der folgenden Klassenausprägung bei der *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 84
$J_t^{(c)}$	negative log-Likelihood der folgenden Beobachtung bei der *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 80
K	Parameter der Kostenfunktion	S. 34
L	Fensterbreite bei adaptiver Lernrate bei *OLDC*	S. 75
$\mathcal{L}$	negative log-Likelihood	S. 77
m	Anzahl neuer Klassen in Chunk	S. 102
$\boldsymbol{m}_0$	Parameter bei *LDA-AF*	S. 87
$\boldsymbol{m}_j^{(c)}$	linke Seite bei linearem Modell für eine einzelne Dimension j	S. 162
$\boldsymbol{m}_{n_t}$	Mittelwertvektor zum Zeitpunkt t basierend auf n_t Beob.	S. 60
$\boldsymbol{m}_{n_t^{(c)}}^{(c)}$	ML-Schätzer für Erwartungswertvektor $\boldsymbol{\mu}_t^{(c)}$ basierend auf $n_t^{(c)}$ Beobachtungen (Mittelwertvektor)	S. 60
$\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}$	Schätzer für Erwartungswertvektor von Klasse c bei der *Online Diskriminanzanalyse mit adaptivem Vergessen* im Falle von exponentiellem Vergessen	S. 78
$\tilde{\boldsymbol{m}}_{n_t}^{(P)}$	Parameter bei *LDA-AF*	S. 87
$\boldsymbol{m}_{n_{t_1}}/\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)}$	Mittelwertvektoren zum Zeitpunkt t_1 (bei Chunks)	S. 66
$\boldsymbol{m}_{n_{t_1:t_2}}/\boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}$	Mittelwertvektoren für Chunk aus Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$	S. 66
$\bar{\boldsymbol{m}}_t^{(c)}$	Mittelwert aller Mittelwertvektoren aus Klasse c ($i \in I$) zum Zeitpunkt t	S. 169
M	Anzahl Klassen	S. 9/31
$Mult$	Multinomialverteilung	S. 31
n_μ	Anzahl der Mittelwerte für Initialisierung des lokalen linearen Regressionsmodells in Simulationsstudie	S. 226
n_{init}	Anzahl der Beobachtungen für Initialisierung der Diskriminanzanalyse in Simulationsstudie	S. 226
$n_{\text{trend}}^{(c)}$	Anzahl aktualisierter Mittelwertvektoren aus Klasse c in N_{trend} Aktualisierungsschritten der Diskriminanzanalyse	S. 162
$n_{\text{Situation}}$	Anzahl Beobachtungen in jeweiliger Datensituation der Simulationsstudie	S. 232
n	Anzahl Beobachtungen in Batch Variante der Diskriminanzanalyse	S. 31

Fortsetzung auf der nächsten Seite

Symbol	Bedeutung	Erklärung/Erstmals
$n^{(c)}$	Anzahl Beobachtungen in Klasse c in Batch Variante der Diskriminanzanalyse	S. 31
n_t	Gesamtanzahl der Beobachtungen zum Zeitpunkt t	S. 60
$n_t^{(c)}$	Anzahl an Beobachtungen in Klasse c zum Zeitpunkt t	S. 60
$n_{t_1}/n_{t_1}^{(c)}$	entsprechende Anzahl an Beobachtungen zum Zeitpunkt t_1 (bei Chunks)	S. 62
$n_{t_1:t_2}/n_{t_1:t_2}^{(c)}$	entsprechende Anzahl an Beobachtungen in Chunk (Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$)	S. 62
$\boldsymbol{N}$	Matrix von Eigenvektoren	S. 49
$\mathcal{N}$	Normalverteilung	S. 15/32
$N_t^{(0)}$	Normierungskonstante für die Gewichte $v_i^{(0)}$	S. 84
$N_t^{(c)}$	Normierungskonstante für die Gewichte $v_i^{(c)}$	S. 79
$N_t^{(P)}$	entsprechende Normierungskonstante bei *LDA-AF*	S. 88
N_{trend}	Anzahl der betrachteten Aktualisierungsschritte der Diskriminanzanalyse für Anpassung des lokalen linearen Regressionsmodells	S. 162
p	Anzahl an Variablen (Dimension)	S. 31
$p^{(c)} := P(Y = y_c)$	a-priori Wahrscheinlichkeit von Klasse c	S. 9
$\hat{p}^{(c)}$	Schätzer für die a-priori Wahrscheinlichkeit von Klasse c	S. 43
$\boldsymbol{P}$	Transformationsmatrix	S. 167
$P_t^{(c)}$	Schätzer für die a-priori Wahrscheinlichkeit von Klasse c zum Zeitpunkt t	S. 63
$\tilde{P}_t^{(c)}$	rekursiver Schätzer für die a-priori Wahrscheinlichkeit von Klasse c bei *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 83
$P(\boldsymbol{X}, Y)$	gemeinsame Verteilung von $\boldsymbol{X}$ und Y	S. 9
$P(\boldsymbol{X} = \boldsymbol{x} \mid Y = y_c)$	bedingte Verteilung	S. 9
$P(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$	bedingte Verteilung	S. 9
$\boldsymbol{Q}_t^{(c)}$	Teil von $\boldsymbol{S}_t^{(c)}$, $\boldsymbol{S}_t$ und $\tilde{\boldsymbol{S}}_t$	S. 60
r	Anzahl an Variablen bei Fisher LDA ($r < p$)	S. 45
R_c	Region von $\boldsymbol{x}$ mit maximaler a-posteriori Wahrscheinlichkeit für Klasse c	S. 36
$\boldsymbol{S}_t$	gepoolte Kovarianzmatrix innerhalb der Klassen zum Zeitpunkt t	S. 61
$\tilde{\boldsymbol{S}}_t$	$\boldsymbol{S}_t$ ohne Vorfaktor	S. 61
$\boldsymbol{S}_t^{(c)}$	empirische Kovarianzmatrix der Klasse c zum Zeitpunkt t	S. 60
t	fester Zeitpunkt	S. 59
t_1, t_2	zwei verschiedene Zeitpunkte in Kapitel 2	S. 10
t_j	Zeitpunkte bei Chunks	S. 61
T	Menge der Zeitpunkte eines Chunks: $T := \{t_1 + 1, \ldots, t_2\}$	S. 62
$T_1^{(c),A}$	Schätzfunktion für $\boldsymbol{\mu}^{(c)}$ bei *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 155
$T_1^{(c),K}$	Schätzfunktion für $\boldsymbol{\mu}^{(c)}$ bei *OLDC*	S. 147
$T_1^{(c),P}$	Schätzfunktion für $\boldsymbol{\mu}^{(c)}$ bei *Sequential ILDA*	S. 139
$T_2^{(c)}$	Schätzfunktion für $\boldsymbol{\mu}_{t+1}^{(c)}$ der erweiterten Methoden	S. 179
$T_2^{(c),A}$	Schätzfunktion für $\boldsymbol{\mu}_{t+1}^{(c)}$ bei Erweiterung von *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 204
$T_2^{(c),K}$	Schätzfunktion für $\boldsymbol{\mu}_{t+1}^{(c)}$ bei Erweiterung von *OLDC*	S. 198
$T_2^{(c),P}$	Schätzfunktion für $\boldsymbol{\mu}_{t+1}^{(c)}$ bei Erweiterung von *Sequential ILDA*	S. 182
$\mathcal{U}$	Rechteckverteilung/stetige Gleichverteilung	S. 213
$\boldsymbol{U}$, $\boldsymbol{V}$	Matrizen für Chunk Erweiterung von *OLDC*	S. 98
$v_i^{(0)}$	Gewichte (Produkt der Faktoren) bei Verteilung von Klasse bei der *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 83
$v_i^{(c)}$	Gewicht in Klasse c (Produkt der Faktoren) bei der *Online Diskriminanzanalyse mit adaptivem Vergessen*	S. 78
$\boldsymbol{v}_j$, $j = 1, \ldots, p$	normalisierte Eigenvektoren von $\boldsymbol{\Sigma}$	S. 48
$\boldsymbol{W}$	Zufallsvektor für Verteilung der Klassen	S. 31

Fortsetzung auf der nächsten Seite

Symbol	Bedeutung	Erklärung/Erstmals
$W_{t,L}$	Fenster zur Adaption der Lernrate bei *OLDC*	S. 75
$\boldsymbol{x}$	Ausprägung des Zufallsvektors $\boldsymbol{X}$	S. 9
$\boldsymbol{x}^*$	beliebige neue Beobachtung	S. 40
$\boldsymbol{x}_i$	Realisation des Zufallsvektors $\boldsymbol{X}_i$	S. 31/59
$\boldsymbol{x}_{(i)}^{(c)}$	i-te in Klasse c realisierte Beobachtung	S. 32
$\boldsymbol{X}$	Zufallsvektor der Einflussvariablen	S. 9
$\boldsymbol{X}^{(c)}$	Zufallsvektor für Klasse c	S. 32
$\boldsymbol{X}_i$	Zufallsvektor zum Zeitpunkt i bzw. für Beobachtung i	S. 31/59
$\boldsymbol{X}_i^{(c)}$	Zufallsvektor zum Zeitpunkt i für Klasse c	S. 59
y_c	Ausprägung der Zielvariable Y (Klassenausprägung)	S. 9
$\boldsymbol{y}_i^{(c)}$	Schätzer für Erwartungswertvektor aus Klasse c zum Zeitpunkt i für lokales lineares Regressionsmodell	S. 179/180
$\bar{\boldsymbol{y}}_t^{(c)}$	Mittelwert über alle $n_{\text{trend}}^{(c)}$ Schätzwerte $\boldsymbol{y}_i^{(c)}$ $(i \in I)$	S. 179/180
Y	Zielvariable	S. 9
$\boldsymbol{\mathcal{Y}}^{(c)}$	transformierter Zufallsvektor von $\boldsymbol{X}^{(c)}$ bei Fisher LDA	S. 46
$\boldsymbol{Y}_i^{(c)}$	entsprechende Schätzfunktion zu $\boldsymbol{y}_i^{(c)}$	S. 179/180
$\bar{\boldsymbol{Y}}_t^{(c)}$	entsprechende Schätzfunktion zu $\bar{\boldsymbol{y}}_t^{(c)}$	S. 179/180
$\boldsymbol{z}$	Differenz aus neuer Beobachtung und bisherigem Mittelwertvektor bei *OLDC*: $\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_t^{(k)}}^{(k)}$	S. 71
$\boldsymbol{z}^*$	Differenz bei *OLDC* bei Betrachtung mit Lernrate	S. 73
$\boldsymbol{z}_{\text{chunks}}$	Differenz bei Chunk Version von *OLDC*	S. 106
$\boldsymbol{z}_{\text{chunks}}^*$	Differenz bei Chunk Version von *OLDC* mit Lernrate	S. 119
$z_i^{(c)}$	verschobener Zeitpunkt des Zeitpunktes i: $z^{(c)}(i)$	S. 162/164
$\bar{z}_t^{(c)}$	Mittelwert aller verschobenen Zeitpunkte $z_i^{(c)}$ für $i \in I$ zum Zeitpunkt t	S. 169
$Z_i^{(c)}$	entsprechende Zufallsvariable zu $z_i^{(c)}$	S. 179/180
$\boldsymbol{Z}_j^{(c)}$	Designmatrix bei linearem Modell für eine einzelne Dimension j	S. 162
$\bar{Z}_t^{(c)}$	entsprechende Zufallsvariable zu $\bar{z}_t^{(c)}$	S. 179/180
$\mathbb{1}_A = \mathbb{1}(A)$	Indikatorfunktion mit Ereignis A	S. 42
$\mathbf{1}_p$	p-dimensionaler Einsenvektor	S. 220

1 Einleitung

„Das Leben gehört dem Lebendigen an, und wer lebt, muß auf Wechsel gefasst sein.“

Johann Wolfgang von Goethe,
Wilhelm Meisters Wanderjahre

Laut einer der International Data Corporation[1] (IDC) als Whitepaper „Data Age 2025: The Digitization of the World. From Edge to Core.“ (Reinsel et al., 2018) veröffentlichten und von Seagate[2] gesponserten Studie wird es einen prognostizierten Zuwachs der globalen Datenmenge von im Jahr 2018 „noch“ 33 Zettabytes auf bis zu 175 Zettabytes im Jahr 2025 geben. Zudem wird prognostiziert, dass 2025 bereits annähernd 30 % der Datenmenge als Echtzeit-Daten (real-time) vorliegen wird.

Vor diesem Hintergrund gewinnen Online-Algorithmen zur Analyse von Daten in allen möglichen praktischen Anwendungsgebieten sowohl der Industrie als auch der Forschung immer mehr an Bedeutung. Die Entwicklung von Methoden zur kontinuierlichen Datenproduktion und -speicherung alleine liefert keinen Mehrwert, wenn die Information der zusätzlichen Daten nicht ausgewertet und genutzt werden kann. Daher spielen statistische Methoden zur Aufbereitung und Analyse von Massendaten, die in Form von *Datenströmen* auftreten können, neben sogenannten Batch-Methoden, bei denen alle Daten zur Anpassung eines Modells gleichzeitig zur Verfügung stehen und verarbeitet werden können, eine immer stärkere Rolle.

Im Kontext mit Datenzuwachs und insbesondere Datenströmen kann durch den Zeitfaktor ein weiteres Problem in Form von *concept drift* auftreten. Dies bedeutet, dass sich im Laufe der Zeit die den Daten zugrunde liegende Verteilung ändern kann, sodass die Annahme einer identischen Verteilung, aus der die Beobachtungen realisiert werden, verletzt ist. Die Online-Fähigkeit bestehender statistischer Modelle wird durch Update-Algorithmen erreicht, die somit eine weitere Adaption an veränderliche Verteilungen erfordern.

Die Dissertation befasst sich mit *Klassifikationsmethoden* (insbesondere der *Diskriminanzanalyse*). Das Flussdiagramm in Abbildung 1.1 stellt eine Übersicht über die Thematik dar

[1] International Data Corporation (IDC) ist laut eigener Aussage das weltweit führende Marktforschungs- und Beratungsunternehmen sowie Veranstalter auf dem Gebiet der Informationstechnologie, Telekommunikation und Verbrauchertechnologie. `https://www.idc.com`.

[2] Seagate Technology ist ein Hersteller von Festplatten und Bandlaufwerken. `https://www.seagate.com`.

und veranschaulicht die Zusammenhänge der einzelnen Themen, welche in den folgenden Kapiteln und Abschnitten behandelt werden. Vertikal von oben nach unten vergrößert sich die Problemkomplexität. Auf der rechten Seite des Pfeils ist jeweils eine Übersicht über Methoden zum Umgang mit der (links beschriebenen) Problemstellung aufgeführt, welche im Folgenden behandelt werden. Farbig hervorgehoben sind dabei erarbeitete Erweiterungen und daraus resultierende Erkenntnisse, die das thematische Feld der Klassifikationsverfahren und insbesondere der Diskriminanzanalyse unter concept drift erweitern.

Kann zur Analyse einer Problemstellung (insbes. bei kategorialer Zielvariable) eine Klassifikationsmethode herangezogen werden, ist eine bekannte und verbreitete Methode die Diskriminanzanalyse (1. „Block“ in Abbildung 1.1). Als Überbegriff wird damit zum einen die Lineare Diskriminanzanalyse nach Fisher (*Fisher LDA*) und zum anderen die Kanonische (Lineare) Diskriminanzanalyse bezeichnet. Diese unterscheiden sich in der Herleitung, führen jedoch zu denselben Klassifikationsgrenzen. Bei der Kanonischen Diskriminanzanalyse besteht zudem die Möglichkeit der Anpassung nicht-linearer Klassifikationsgrenzen durch Betrachtung verschiedener Kovarianzmatrizen in den Klassen (*Kanonische QDA*).

Die Komplexität von Klassifikationsproblemen wird erhöht, wenn die Daten nicht als Batch vorliegen und alle Beobachtungen gleichzeitig zur Anpassung eines Modells herangezogen werden können, sondern wenn Datenströme betrachtet werden (2. „Block“ in Abbildung 1.1). Aufgrund von zeitlichen oder auch Datenspeicher-technischen Einschränkungen ist es häufig nicht möglich das gesamte Modell auf allen Beobachtungen neu anzupassen. Die kontinuierlich auftretenden Beobachtungen können daher eine Aktualisierung des bisherigen Klassifikationsmodells ermöglichen oder sogar erfordern. Pang et al. (2005a,b) schlagen eine Adaption der Fisher LDA für Datenströme vor. Mithilfe von Standard Update-Formeln für Mittelwertvektoren und Kovarianzmatrizen können die nötigen Größen zur Konstruktion der Klassifikationsregel mit einer oder mehreren neuen Beobachtungen aktualisiert werden. Eine Online Variante der Kanonischen LDA führen Kuncheva und Plumpton (2008) durch *OLDC* ohne Lernrate ein. Die Ergebnisse beider Methoden sind identisch mit denen der jeweiligen Batch Variante bei Anpassung auf allen Beobachtungen gleichzeitig. Es erfolgt demnach keine Anpassung an einen concept drift.

Insbesondere zum Umgang mit einem solchen concept drift wurde aufgrund der aktuellen Relevanz besonders in den letzten Jahren sehr viel Forschungsarbeit betrieben (3. „Block“ in Abbildung 1.1). Concept drift erfordert eine Adaption der Update-Methoden, um weiterhin gute Klassifikationsergebnisse zu garantieren. Als Beispiel führen Kuncheva und Plumpton (2008) (feste oder adaptive) Gewichtungen in Form einer Lernrate λ bei den Aktualisierungsformeln der nötigen Größen zur Bildung der Klassifikationsregel ein. Anagnostopoulos et al. (2012) nutzen die Idee des adaptiven exponentiellen Vergessens und entwickeln so eine Online Variante der Diskriminanzanalyse, welche sich adaptiv an eine Veränderung der zugrunde liegenden Verteilung der Beobachtungen im Datenstrom anpassen soll. Die Anpassungsgüte der resultierenden Klassifikatoren zum Zeitpunkt der Aktualisierung kann somit stark verbessert werden.

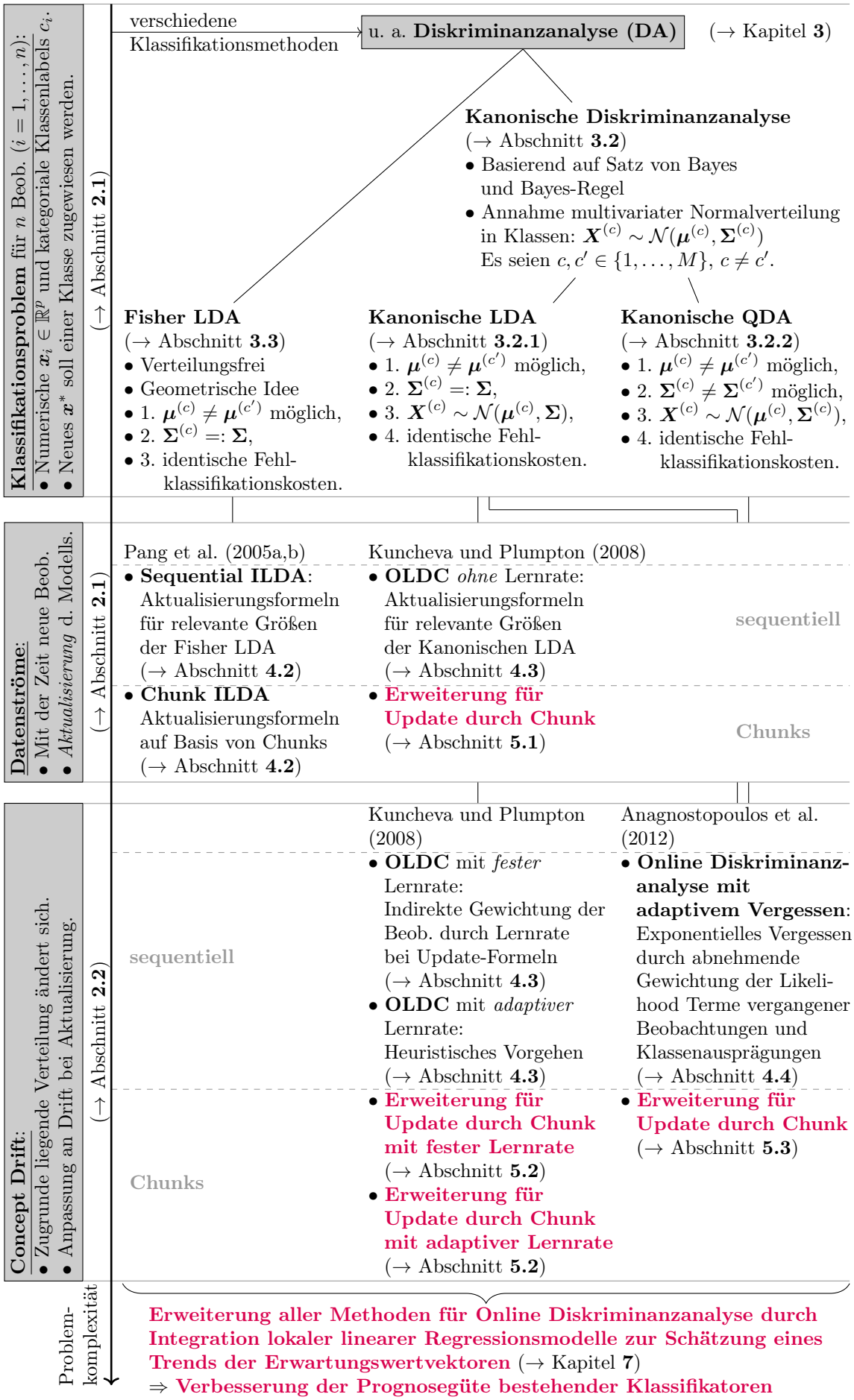


Abbildung 1.1: Übersicht über die Thematik und Zusammenhänge. **Farbig markiert**: Neu entwickelte Methoden und Erweiterungen.

Ein möglicher fortschreitender Trend der zugrunde liegenden Verteilung wird jedoch meistens nicht beachtet. Im Falle eines (linearen) Trends der Erwartungswertvektoren fließen „zeitverzögerte“ Schätzer für die Erwartungswertvektoren in die Klassifikationsregel der Diskriminanzanalyse ein, wodurch die Prognosegüte bestehender Klassifikatoren für kommende Zeitpunkte nicht optimal ist. Da Klassifikationsmethoden jedoch häufig herangezogen werden, um Prognosen für zukünftige Beobachtungen tätigen zu können, liegt der Fokus der Dissertation auf einer Erweiterung der Methoden für Online Diskriminanzanalyse zur Verbesserung der Prognosegüte. Als Optimierungsschritt in Situationen mit concept drift werden lokale lineare Regressionsmodelle zur Modellierung und Prognose eines Trends der Erwartungswerte in die Update-Algorithmen integriert. Dadurch können verbesserte Schätzer für die Erwartungswertvektoren in die jeweiligen Klassifikatoren einfließen.

Im Folgenden beleuchtet Kapitel 2 zum einen allgemein das Thema *Datenströme* (Abschnitt 2.1). Zum anderen wird der Begriff *concept drift* eingeführt und formalisiert (Abschnitt 2.2).

Kapitel 3 befasst sich mit den verschiedenen *Diskriminanzanalyseverfahren*. Als Folgerung aus der *Bayes-Regel* (Abschnitt 3.1) wird in Abschnitt 3.2 sowohl die Lineare (LDA) als auch Quadratische *Kanonische Diskriminanzanalyse* (QDA) inklusive Schätzung der Parameter erläutert. Die entsprechende *Fisher LDA* als Resultat einer geometrischen Überlegung der Trennung von Daten wird in Abschnitt 3.3 beschrieben. Abschnitt 3.4 befasst sich mit einem Vergleich von Kanonischer LDA und Fisher LDA. Anlehnend an van Meegen (2015) und van Meegen et al. (2019) wird gezeigt, dass beide Varianten in der Theorie unabhängig von der Anzahl an Klassen, Anzahl an Variablen und der Form der a-priori Wahrscheinlichkeiten dieselben Ergebnisse liefern.

In Kapitel 4 werden als Beispiele für Methoden für Online Diskriminanzanalyse die oben bereits erwähnten Methoden *Sequential ILDA* oder *Chunk ILDA* von Pang et al. (2005a,b) (Abschnitt 4.2), *Online Linear Discriminant Classifier (OLDC)* von Kuncheva und Plumpton (2008) (Abschnitt 4.3) und *Online Diskriminanzanalyse mit adaptivem Vergessen* von Anagnostopoulos et al. (2012) (Abschnitt 4.4) vorgestellt.

Da die Algorithmen von Kuncheva und Plumpton (2008) und Anagnostopoulos et al. (2012) nur die Aktualisierung der Klassifikationsregel durch eine einzelne neue Beobachtung behandeln, befasst sich Kapitel 5 mit einer Erweiterung der beiden Methoden, sodass mehrere Beobachtungen gleichzeitig in die Aktualisierungsformeln einfließen können.

In Kapitel 6 werden die Schätzfunktionen für die Erwartungswertvektoren der betrachteten Methoden untersucht. Der Fokus liegt dabei auf der Analyse der Erwartungstreue der Schätzfunktionen für die Erwartungswertvektoren der Prognose, also der Verteilungen des kommenden Zeitpunktes. In vielen Fällen sind die Schätzfunktionen erwartungstreu für den Erwartungswertvektor der Prognose, falls stabile Verteilungen über die Zeit vorliegen. Wird jedoch ein linearer Trend der Erwartungswertvektoren in den Klassen unterstellt, so sind die Schätzfunktionen verzerrt. Trotz Methodik zum Umgang mit concept drift wird

kein fortschreitender Trend beachtet. Im Falle eines linearen Trends fließen „zeitverzögerte“ Schätzer für die Erwartungswertvektoren in die Klassifikationsregel der Diskriminanzanalyse ein, wodurch die Prognosegüte verschlechtert wird (Abschnitt 7.1).

Vor dem Hintergrund von concept drift in Datenströmen wird daher in Kapitel 7 eine Methodik zur Verbesserung der Prognosegüte unter der Annahme eines linearen Trends der Erwartungswertvektoren der Klassen (Abschnitt 7.2) entwickelt. Der Trend wird dazu mithilfe von lokalen linearen Regressionsmodellen auf den kontinuierlich aktualisierten Schätzern der Erwartungswertvektoren modelliert (Abschnitt 7.3). Mithilfe dieser Regressionsmodelle können die Erwartungswertvektoren der Klassen des kommenden Zeitpunktes prognostiziert werden (Abschnitt 7.4), sodass verbesserte Schätzer in die Klassifikatoren einfließen können und die Prognosegüte verbessert werden kann (Abschnitte 7.5–7.7).

Kapitel 8 befasst sich mit einer Untersuchung der Erwartungstreue der erweiterten Schätzfunktionen. Es werden Unterschiede und Verbesserungen zu den bisherigen Schätzfunktionen für die Erwartungswertvektoren in bestimmten Situationen herausgestellt.

Kapitel 9 umfasst eine umfangreiche Simulationsstudie zur Untermauerung und Erweiterung der (auf Spezialfällen bewiesenen) theoretischen Ergebnisse aus Kapitel 8. Zunächst werden einige typische simulierte Datensituationen, die sich in der Literatur zum Umgang mit concept drift etabliert haben, beschrieben (Abschnitt 9.1). Zudem werden weitere Datensituationen und Arten von concept drift definiert (Abschnitt 9.2). Nach einer Beschreibung der Durchführung der Simulationsstudie (Abschnitt 9.3), der Wahl der Parametereinstellungen für die einzelnen Methoden (Abschnitt 9.4) und der Implementierung (Abschnitt 9.5) folgt in Abschnitt 9.6 eine Darstellung, Diskussion und ein Fazit der Ergebnisse der gesamten Simulationsstudie.

Kapitel 10 fasst die Ergebnisse der Analyse vor dem Hintergrund der Problemstellung der Verbesserung der Prognosegüte von Methoden für Online Diskriminanzanalyse zusammen. Des Weiteren folgt ein Ausblick zu über die Analyse hinausgehenden Problemstellungen. Es werden weitere Ideen zusammengefasst, die im thematischen Kontext mit concept drift und Diskriminanzanalyse interessant sind und eine Grundlage für weitere Forschungsarbeit darstellen oder zu einer Erweiterung bisheriger Forschungsarbeit beitragen können.

2 Datenströme und Concept Drift

2.1 Datenströme

Die Häufigkeit von Datenströmen im Gegensatz zu festen Datensätzen („Batch“) hat in den letzten Jahren aufgrund diverser Weiterentwicklungen zur Generierung von immer mehr Daten kontinuierlich zugenommen. Wares et al. (2019) fassen in ihrem Review Paper „Data stream mining: methods and challenges for handling concept drift“ mit Verweis auf Babcock et al. (2002) Charakteristiken von Datenströmen zusammen:

- Daten kommen im Datenstrom *online* an.
- Datenströme können eine beliebige (unbeschränkte) Größe annehmen.
- Die Beobachtungen kommen in beliebiger Reihenfolge an bzw. die Reihenfolge kann nicht kontrolliert werden.
- Die einzelnen Beobachtungen sind häufig nur eine Zeit lang verfügbar und werden nicht dauerhaft gespeichert bzw. können nicht dauerhaft gespeichert werden.

All diese Charakteristiken bereiten neue Herausforderungen zur Verarbeitung von Daten in Datenströmen im Vergleich zur Betrachtung von „Batch“ Daten (Wares et al., 2019). In festen (und beschränkten) Datensätzen können alle Beobachtungen gleichzeitig verarbeitet werden. Zudem können die einzelnen Beobachtungen immer wieder neu aufgerufen werden und es besteht die Möglichkeit einer zufälligen Auswahl bzw. Reihenfolge der Beobachtungen für die Anwendung von Algorithmen. In Datenströmen hingegen können die einzelnen Beobachtungen aufgrund der Online-Generierung mit einer sehr schnellen Geschwindigkeit ankommen. Die Beobachtungen werden nur vorübergehend gespeichert und können daher nur sequentiell (in der festen Reihenfolge) und eine begrenzte Zeit lang bzw. nicht beliebig häufig verarbeitet werden. Eine zufällige Auswahl von Beobachtungen oder Bestimmung der Reihenfolge zur Verarbeitung in Algorithmen ist nicht möglich. Zudem können anders als bei festen Datensätzen unvorhersehbare Charakteristiken der Beobachtungen auftreten. So kann sich insbesondere vor dem Hintergrund des Zeitfaktors die zugrunde liegende Verteilung der Daten ändern (Aggarwal, 2007, S. 2).

Solch eine Veränderung der zugrunde liegenden Verteilung wird auch als *concept drift* bezeichnet und im nächsten Abschnitt genauer charakterisiert und formalisiert. Aggarwal (2007, S. 2) weist darauf hin, dass als Folge einer solchen Veränderung einfache Adaptionen von one-pass Algorithmen für Datenströme, mithilfe derer Modelle aktualisiert werden

können, nicht ausreichend sind. Vielmehr müssen Online-Algorithmen zusätzlich in der Lage sein eine Veränderung der zugrunde liegenden Verteilung geeignet einzubeziehen.

Die Forschung auf diesem Gebiet der Adaption von Algorithmen für Datenströme sowie Aufdeckung und Behandlung von concept drift hat in den letzten Jahren aufgrund der praktischen Relevanz sehr stark zugenommen. Wenn concept drift im Datenstrom vermutet wird oder aufgedeckt wurde, müssen Algorithmen in geeigneter Weise den Fokus auf den Umgang mit solch einem Drift legen können. Hoens et al. (2012, S. 93) unterteilen die Methoden zum Umgang mit concept drift in drei verschiedene Kategorien und liefern zusätzlich Beispiele mit Literaturverweisen: *adaptive base learners* (dt. adaptive Lerner), *learners which modify the training set* (dt. Modifikation des Trainingsdatensatzes) und *ensemble techniques* (dt. Ensemble Techniken).

Adaptive Lerner sind in der Lage sich adaptiv an eine Veränderung der zugrunde liegenden Verteilung anzupassen. Das konkrete Vorgehen hängt dabei von dem speziellen Lerner ab. Die grundlegende Idee besteht darin, dass der Raum der verwendeten Daten zur Erzeugung der Klassifikationsregel eingeschränkt oder erweitert wird (Hoens et al., 2012, S. 94 f.).

Modifikation des Trainingsdatensatzes kann hingegen unabhängig von dem Lerner (zum Beispiel der Klassifikationsmethode) mithilfe von *Fenstertechniken* oder *Gewichtung* der einzelnen Beobachtungen erfolgen (Hoens et al., 2012, S. 95 ff.). Die grundlegende Idee besteht darin, dass entweder nur ein Fenster der neuesten Beobachtungen im Datenstrom betrachtet wird und diese zur Konstruktion des Klassifikators herangezogen werden. Oder aber die Beobachtungen im Datenstrom werden entsprechend gewichtet, sodass beispielsweise neuere Beobachtungen einen stärkeren Einfluss haben. Die Schwierigkeit besteht dann in der geeigneten Wahl der Fensterbreite oder der Gewichte.

Bei *Ensemble Methoden* werden verschiedene schwache Lerner auf ähnlichen Datensätzen kombiniert. Im Zusammenhang mit concept drift kann dies insbesondere einen Vorteil bei reoccurring concepts (siehe Seite 26) bieten (Hoens et al., 2012, S. 97 ff.).

Einen Überblick über Methoden zur Aufdeckung von concept drift, welche auch als *change detection*, *concept drift detection* oder *anomaly detection* bezeichnet werden, liefern unter anderen Aggarwal (2007), Gama (2010), Gama et al. (2014) und Wares et al. (2019). Gaber et al. (2005), Gama (2010), Hoens et al. (2012), Gama et al. (2014) und Kolajo et al. (2019) bieten unter anderen einen Literaturüberblick über Algorithmen für Datenströme und den Umgang mit concept drift für verschiedene statistische Methoden wie Cluster-, Klassifikationsverfahren oder Zeitreihenanalyse.

2.2 Concept und Concept Drift

Bevor *concept drift* charakterisiert und formalisiert werden kann, muss zunächst definiert werden, was in der Literatur üblicherweise als *concept* (dt. Konzept) bezeichnet wird.

Zur Veranschaulichung werden hier zunächst vereinfachend die Bezeichnungen Y für die Zielvariable und $\boldsymbol{X}$ für den Zufallsvektor der Einflussvariablen verwendet. Die entsprechenden Ausprägungen werden mit y_c und $\boldsymbol{x}$ bezeichnet. Da die Problematik im Kontext von Klassifikationsverfahren betrachtet wird, ist Y eine diskret verteilte Zufallsvariable. Als Einflussvariablen werden vereinfachend nur stetig verteilte Einflussgrößen betrachtet. Die speziellen Verteilungen (wie Multinomialverteilung der Klassen) und daraus resultierenden theoretischen Eigenschaften werden in Kapitel 3 formal eingeführt und hier zunächst außer Acht gelassen.

Insgesamt wird daher die Verteilung von $\boldsymbol{X}$ anhand der Dichtefunktion $f_{\boldsymbol{X}}(\boldsymbol{x})$ charakterisiert, die Verteilung von Y anhand der Wahrscheinlichkeitsfunktion $P(Y = y_c)$. Anders als in den meisten Veröffentlichungen, die im Folgenden erwähnt und gegenübergestellt werden, wird anstelle von $P(\boldsymbol{X}, Y)$ die gemeinsame Dichtefunktion der Einfluss- und Zielvariablen $f_{\boldsymbol{X},Y}(\boldsymbol{x}, y_c)$ herangezogen, um die gemeinsame Verteilung zu charakterisieren.

2.2.1 Concept

In jüngerer Literatur, welche sich mit der Definition und dem Umgang von *concept drift* in Datenströmen und Klassifikationsproblemen beschäftigt, werden wahrscheinlichkeitstheoretische Definitionen von *concept* betrachtet (Webb et al., 2016, S. 968; Kuncheva, 2004, S. 2).

Basierend darauf, dass eine Klassifizierung mithilfe des Satzes von Bayes (vgl. Satz 1 auf Seite 11 und Abschnitt 3.1) beschrieben werden kann, kann ein *concept* durch Wahrscheinlichkeitsaussagen definiert werden. Žliobaitė (2010, S. 3) definiert ein concept als eine Menge aus den a-priori Klassenwahrscheinlichkeiten $p^{(c)} := P(Y = y_c),\ c = 1, \ldots, M$, sowie den klassenbedingten Wahrscheinlichkeitsdichtefunktionen $f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c),\ c = 1, \ldots, M$. Aufgrund des Zusammenhangs $P(\boldsymbol{X}, Y) = P(\boldsymbol{X} = \boldsymbol{x}|Y = y_c)P(Y = y_c)$ charakterisieren Gama et al. (2014, S. 4) und Webb et al. (2016, S. 968) ein concept lediglich durch diese gemeinsame Verteilung der Einfluss- und Zielvariablen. Da die Einflussvariablen jedoch stetig verteilt sein können, ist es aus theoretischer Sicht sinnvoller die Verteilungen anhand der Dichtefunktionen $f_{\boldsymbol{X},Y}(\boldsymbol{x}, y_c)$ zu charakterisieren, da die Wahrscheinlichkeit einer stetigen Zufallsvariablen in einem einzelnen Punkt als Null definiert ist und nur Intervalle betrachtet werden können.

Hoens et al. (2012, S. 91) formulieren im weiteren Sinne, dass ein concept durch die datengenerierende Funktion charakterisiert wird. Diese datengenerierende Funktion bzw. der datengenerierende Prozess hängt stark mit der zugrunde liegenden Verteilung der Beobachtungen zusammen. Die zugrunde liegende Verteilung kann durch die verschiedenen Wahrscheinlichkeiten oder eben (im stetigen Fall) die Dichtefunktionen repräsentiert werden.

Im Folgenden wird *concept* äquivalent für den *datengenerierenden Prozess*/ die *datengenerierende Funktion*, die Menge $\{P(Y = y_c), f_{\boldsymbol{X}}(\boldsymbol{x}), f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c), P(Y = y_c|\boldsymbol{X} = \boldsymbol{x})\}$,

$c = 1, \ldots, M$, oder die *zugrunde liegende Verteilung* verwendet. Die englischen Begriffe *concept* und *concept drift* bzw. die Übersetzungen *Konzept* und *Drift* werden äquivalent verwendet.

2.2.2 Concept Drift

Der Begriff *concept drift* wurde als Erstes von Schlimmer und Granger (1986) eingeführt (Widmer und Kubat, 1996, S. 70; Žliobaitė, 2010, S. 9). Er lässt sich als Analogon zum *dataset shift* betrachten, wenn inkrementelle Lernverfahren in Datenströmen statt klassische Lernverfahren auf einem festen Datensatz betrachtet werden (Webb et al., 2016, S. 968). Nach Quiñonero-Candela et al. (2009, S. xi) beschreibt *dataset shift* den Zustand, dass die gemeinsame Verteilung in Trainings- und Testdaten eines Datensatzes jeweils unterschiedlich ist. Im Unterschied zum *concept drift* ist die Veränderung beim *dataset shift* somit nicht zwangsläufig zeitgebunden. Wenn jedoch im Datenstrom neue Beobachtungen als Testdaten charakterisiert werden, Trainings- und Testdaten also zeitlich aufeinander folgen, so lassen sich sowohl mit *concept drift* als auch *dataset shift* Veränderungen in der Struktur der Daten im Laufe der Zeit beschreiben. Im Folgenden wird nicht zwischen diesen beiden Ausgangssituationen unterschieden. Es werden immer – auch bei *dataset shift* und beispielsweise den von Moreno-Torres et al. (2012) beschriebenen verschiedenen Typen (s. folgende Tabelle 2.1) – aufeinanderfolgende Zeitpunkte betrachtet, sodass eine zeitabhängige Veränderung der Struktur der Daten unterstellt wird.

Ganz allgemein betrachtet lässt sich *concept drift* in Hinblick auf die Definition eines *concepts* im vorangegangenen Abschnitt so erklären, dass sich die datengenerierende Funktion bzw. die zugrunde liegende Verteilung der Daten in einem Datenstrom über die Zeit aufgrund irgendwelcher Ereignisse ändert. Die Beobachtungen folgen demnach nicht mehr einer gemeinsamen zugrunde liegenden Verteilung. Eine recht allgemeine formale Definition für *concept drift* liefern Webb et al. (2016, S. 968):

$$P_{t_1}(\boldsymbol{X}, Y) \neq P_{t_2}(\boldsymbol{X}, Y).$$

Analog lässt sich dies durch veränderte gemeinsame Dichtefunktionen von Zeitpunkt t_1 zu Zeitpunkt t_2

$$f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$$

beschreiben, falls Dichtefunktionen zur Charakterisierung der Verteilung betrachtet werden.

In der Literatur, welche sich mit concept drift Thematik beschäftigt, gibt es bis dato jedoch keine einheitliche Definition für den Begriff *concept drift*. Dies äußert sich darin, dass verschiedene AutorInnen teilweise zwischen unterschiedlichen Arten und Formen von concept drift differenzieren, mit denselben Bezeichnungen unterschiedliche Veränderungen der

Verteilungen charakterisiert werden, oder aber identische Veränderungen unterschiedlich bezeichnet werden.

Einige AutorInnen haben sich bereits mit diesen Uneindeutigkeiten auseinandergesetzt. So zitieren Moreno-Torres et al. (2012, S. 522 ff.) verschiedene Definitionen und Beschreibungen anderer AutorInnen bezüglich ihrer Unterscheidungen zwischen den drei verschiedenen Formen von concept drift: *covariate shift*, *prior probability shift* und *concept shift* (vgl. Tabelle 2.1). Ebenso stellen Gama et al. (2014, S. 4 f.) verschiedene Definitionen und Beschreibungen gegenüber.

In Tabelle 2.1 sind die Bezeichnungen unterschiedlicher Arten bzw. „Auslöser“ von concept drift inklusive formaler Definition oder qualitativer Beschreibung verschiedenster Veröffentlichungen zur concept drift Thematik (ohne Anspruch auf Vollständigkeit) zusammengefasst. Die Definitionen basieren dabei alle auf Veränderungen einer oder mehrerer Größen $f_{\boldsymbol{X},Y}(\boldsymbol{x}, y_c)$, $P(Y = y_c) =: p^{(c)}$, $f_{\boldsymbol{X}}(\boldsymbol{x})$, $f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c)$, $P(Y = y_c|\boldsymbol{X} = \boldsymbol{x})$, auf Basis derer concept definiert wurde. Mit einem Sternchen (*) sind jene Veröffentlichungen gekennzeichnet, die nicht einzelne spezielle Verfahren zum Umgang mit concept drift behandeln oder lediglich concept drift kurz beschreiben oder definieren. Stattdessen handelt es sich um Übersichtspaper zur concept drift Thematik oder es werden Veröffentlichungen damit gekennzeichnet, die sich explizit mit der Unterscheidung und Differenzierung bzw. Formalisierung verschiedener Typen und Ausprägungen von concept drift befassen.

Im Zusammenhang mit Klassifikationsproblemen stellt sich die Frage, wann eine Anpassung des Modells aufgrund eines Drifts nötig ist. Aus theoretischer Sicht kann es verschiedene „Auslöser“ für einen Drift geben, weswegen in den meisten Veröffentlichungen auch zwischen unterschiedlichen Veränderungen differenziert wird.

„Auslöser“ des Drifts Ein weit verbreiteter Ansatz ist das Heranziehen des Satzes von Bayes (vgl. Mood et al. (1974, S. 36) mit Übertragung auf Dichtefunktionen):

Satz 1. *Satz von Bayes*

$$P(Y = y_c|\boldsymbol{X} = \boldsymbol{x}) = \frac{f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c)P(Y = y_c)}{\sum_{y_c} f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c)P(Y = y_c)} = \frac{f_{\boldsymbol{X},Y}(\boldsymbol{x}, y_c)}{f_{\boldsymbol{X}}(\boldsymbol{x})}$$

Aufgrund des Zusammenhangs der (bedingten) Verteilungen bzw. Dichtefunktionen der Einfluss- und Zielvariablen in Klassifikationsproblemen erwähnen mehrere AutorInnen (Kelly et al., 1999, S. 368; Kuncheva, 2004, S. 2; Žliobaitė, 2010, S. 4; Hoens et al., 2012, S. 91; Gama et al., 2014, S. 3 f.), dass eine Veränderung einer oder mehrerer der Größen $f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c)$, $P(Y = y_c|\boldsymbol{X} = \boldsymbol{x})$ und $P(Y = y_c)$ einen Drift auslösen kann. Interessant ist hier der kausale Zusammenhang, d. h., was letztendlich den Drift auslöst, da die Verteilungen aufgrund von Satz 1 zusammenhängen.

Tabelle 2.1: Zusammenfassung von Bezeichnungen und Definitionen (bzw. Beschreibungen) verschiedener Arten von concept drift in der Literatur inklusive Verweise auf veranschaulichende Beispiele. Mit * sind jene Veröffentlichungen gekennzeichnet, bei denen es sich um Übersichtspaper zur concept drift Problematik handelt oder die ausführlich verschiedene Arten und Ausprägungen von concept drift formalisieren und gegenüberstellen.

Quelle	Bezeichnung	Änderung	Abbildung
Alaiz-Rodríguez und Japkowicz (2008)	*Changing Environments*	„The fundamental assumption of supervised learning is that the joint probability distribution $[P(\boldsymbol{X}, Y)]$ will remain unchanged between training and testing. There are, however, some mismatches that are likely to appear in practice [...].“: $f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$	2.1, 2.2, 2.3, 2.4, 2.5
	change in class distribution	$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $f_{t_1}(\boldsymbol{x}\|y_c) = f_{t_2}(\boldsymbol{x}\|y_c)$	2.2
	class definition change	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$	2.3
	population drift	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
Bickel et al. (2009)	*covariate shift*	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
Cieslak und Chawla (2009)	*covariate shift*	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
Delany et al. (2005)	*concept drift*	„Changes in the hidden context can induce changes in the target concept, which is generally known as concept drift“: $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
	virtual concept drift	„Hidden changes [...] may also cause a change in the underlying data distribution. Even if the target concept remains the same but the data distribution changes, a model rebuild may be necessary as the model's error may no longer be acceptable.“: $f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
*Gama et al. (2014)	*concept drift*	$f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$	2.1, 2.2, 2.3, 2.4, 2.5
	real concept drift	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$ oder $f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$	2.1, 2.2, 2.3
	virtual drift	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ (und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2
		oder $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$)	2.4, 2.5
Gao et al. (2007)	*conditional change*	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$	2.3
	dual change	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2
	feature change	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
Hand (2006)	*population drift*	„A fundamental assumption of the classical paradigm is that the various distributions involved do not change over time. In fact, in many applications this is unrealistic and the population distributions are nonstationary.“	
	concept drift	„[...] changes to the definitions of the classes“	

Fortsetzung auf der nächsten Seite

Quelle	Bezeichnung	Änderung	Abbildung
Hoens et al. (2012)	*concept drift*	„Concept drift is said to occur when the underlying function [...] changes over time.“:	
		$f_{t_1}(\boldsymbol{x}\|y_c) \neq f_{t_2}(\boldsymbol{x}\|y_c)$	2.1, 2.3, 2.4, 2.5
		$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
		$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$	2.2, 2.3, 2.5
	real concept drift	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
	virtual concept drift	$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.5
		$f_{t_1}(\boldsymbol{x}\|y_c) \neq f_{t_2}(\boldsymbol{x}\|y_c)$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
Huang et al. (2007)	*sample selection bias/ covariate shift*	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
Kelly et al. (1999)	*population drift*	„When the population distribution can change over time we say it is subject to population drift.“:	
		$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
		$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$	2.2, 2.3, 2.5
		$f_{t_1}(\boldsymbol{x}\|y_c) \neq f_{t_2}(\boldsymbol{x}\|y_c)$	2.1, 2.3, 2.4, 2.5
*Moreno-Torres et al. (2012)	*dataset shift*	$f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$	2.1, 2.2, 2.3, 2.4, 2.5
	concept shift	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$ und $\boldsymbol{X} \to Y$	2.3
		$f_{t_1}(\boldsymbol{x}\|y_c) \neq f_{t_2}(\boldsymbol{x}\|y_c)$ und $P_{t_1}(Y = y_c) = P_{t_2}(Y = y_c)$ und $Y \to \boldsymbol{X}$	2.1, 2.4
	covariate shift	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $\boldsymbol{X} \to Y$	2.4, 2.5
	prior probability shift	$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $f_{t_1}(\boldsymbol{x}\|y_c) = f_{t_2}(\boldsymbol{x}\|y_c)$ und $Y \to \boldsymbol{X}$	2.2
Moreno-Torres et al. (2013)	*fracture between the data*	„we have data from one laboratory (dataset A), and derive a classifier from it that can predict its category accurately. We are then presented with data from a second laboratory (dataset B). This second dataset is not accurately predicted by the classifier we had previously built due to a fracture between the data of both laboratories.“	
Salganicoff (1997)	*concept shift*	„Several aspects of the learning problem can vary, including the mapping to be learned [...]“	
	sampling shift	„Several aspects of the learning problem can vary, [...] and the sampling distribution that governs the input-space location of exemplars that make up the learning set.“	
Schlimmer und Granger (1986)	*concept drift/concept change*	„Frequently, however, a derived description of a useful concept is disrupted by some change which requires its revision.“	
Shimodaira (2000)	*covariate shift*	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$	2.1, 2.2, 2.4, 2.5

Fortsetzung auf der nächsten Seite

Quelle	Bezeichnung	Änderung	Abbildung
Storkey (2009)	*covariate shift*	„[...] is when only the distributions of covariates $\boldsymbol{x}$ change and everything else is the same.“:	
		$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$ und $\boldsymbol{X} \rightarrow Y$	2.4, 2.5
	dataset shift	„[...] deals with the business of relating information in (usually) two closely related environments to help with the prediction in one given the data in the other(s)„	
	prior probability shift	„[...] is when only the distribution over y changes and everything else stays the same.“:	
		$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $f_{t_1}(\boldsymbol{x} \vert y_c) = f_{t_2}(\boldsymbol{x} \vert y_c)$ und $Y \rightarrow \boldsymbol{X}$	2.2
Tsymbal (2004)	*real concept drift*	„Hidden changes in context may [...] be a cause of a change of target concept [...]“:	
		$P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
	virtual concept drift	„Even if the target concept remains the same, and it is only the data distribution that changes, this may often lead to the necessity of revising the current model [...]. The necessity in the change of current model due to the change of data distribution [...]“	
*Webb et al. (2016)	*concept drift*	$f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$	2.1, 2.2, 2.3, 2.4, 2.5
	class drift	$P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
	covariate drift	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$	2.1, 2.2, 2.4, 2.5
	pure class drift	$P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$	2.3
	pure covariate drift	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
	neue Klasse	$P_{t_1}(Y = y_c) = 0$ und $P_{t_2}(Y = y_c) > 0$	
Widmer und Kubat (1993)	*real concept drift*	„Real concept drift reflects real changes in the world [...]“	
	virtual concept drift	„Virtual concept drift [...] does not occur in reality but, rather, in the computer model reflecting this reality. In a practical setting, this kind of effect can emerge when the representation language is poor and fails to identify all relevant features, or when the order of training examples for learning is skewed, so that different types of instances are not evenly distributed over the training sequence.“	
Widmer und Kubat (1996)	*concept drift*	„[...] changes in the target concepts [...]“	
Yamazaki et al. (2007)	*covariate shift*	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
	class prior change	$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$	2.2, 2.3, 2.5
	functional relation change	$P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$	2.3
Žliobaitė (2010)	*concept drift*	$f_{t_1}(\boldsymbol{x} \vert y_c) \neq f_{t_2}(\boldsymbol{x} \vert y_c)$	2.1, 2.3, 2.4, 2.5
		$P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
		$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$	2.2, 2.3, 2.5

Kelly et al. (1999, S. 368) argumentieren, dass eine Veränderung der bedingten Verteilung der Klassenzugehörigkeit entscheidend ist, da diese zwangsläufig eine Verschiebung der Klassifikationsgrenzen nach sich zieht. Allerdings können auch Veränderungen der anderen (bedingten) Verteilungen und Wahrscheinlichkeiten ein Problem darstellen. Die Fehlerrate des Modells kann anwachsen, wodurch eine Anpassung des Klassifikationsmodells erforderlich wird. Eine starke Veränderung der a-priori Wahrscheinlichkeiten $p^{(c)} := P(Y = y_c)$ kann zudem in einem Problem von „unbalancierten“ Klassen (imbalance) enden. Daher unterscheiden beispielsweise Hoens et al. (2012, S. 91) nicht zwischen den verschiedenen Auslösern. Zudem machen Kelly et al. (1999, S. 368) auch darauf aufmerksam, dass in den meisten praktischen Anwendungen die „wahren“ Verteilungen und Dichten geschätzt werden müssen und dass sich somit die Güte des Klassifikationsmodells aufgrund von Ungenauigkeiten in den Schätzungen negativ ändern kann, auch wenn theoretisch $P(Y = y_c|\boldsymbol{X} = \boldsymbol{x})$ von einem Drift unberührt bleibt.

Ändert sich etwas in der Verteilung von $\boldsymbol{X}$ oder Y oder in den bedingten Verteilungen, so hat dies wegen des Satzes 1 zwangsläufig auch eine Veränderung einer der anderen oder aller anderen Größen zur Folge. Liegt der Fokus auf der linken Seite der Umformungen des Satzes von Bayes, so lassen sich synthetische Beispiele finden, bei denen zwar eine Veränderung der Verteilungen erfolgt, die bedingte Verteilung der Klassenzugehörigkeit jedoch unberührt bleibt, z. B. durch eine symmetrische Verschiebung der klassenbedingten Verteilungen in entgegengesetzte Richtungen (Žliobaitė, 2010, S. 4) oder gegenseitige Aufhebung von Veränderungen auf der rechten Seite der Gleichung. Das Auftreten solcher Spezialfälle ist in der Praxis jedoch recht unwahrscheinlich und für praktische Anwendungen daher nahezu irrelevant. Es kann davon ausgegangen werden, dass in dem Großteil der Fälle, in denen concept drift in der Praxis auftritt, unabhängig davon, was der Auslöser ist, eine Veränderung der a-posteriori Wahrscheinlichkeiten der Klassenzugehörigkeit resultiert.

Die Veränderung der einzelnen (bedingten) Verteilungen und ihre Auswirkung auf die a-posteriori Wahrscheinlichkeiten der Klassenzugehörigkeiten bzw. auf die Entscheidungsgrenze ist anhand eines Beispiels eines eindimensionalen Datensatzes in den Abbildungen 2.1 bis 2.5 veranschaulichend dargestellt. Es werden zwei Klassen betrachtet, wobei zunächst zum Zeitpunkt t_1 die Beobachtungen jeweils einer Normalverteilung mit Varianz 0.5 und unterschiedlichen Erwartungswerten folgen: $X|(Y = y_1) \sim \mathcal{N}(1.5, 0.5)$ und $X|(Y = y_2) \sim \mathcal{N}(-1.5, 0.5)$. Die a-priori Wahrscheinlichkeiten der Klassen betragen $p^{(1)} = p^{(2)} = 0.5$. Der Trennpunkt erfolgt (in Hinblick auf die Lineare Diskriminanzanalyse) am Schnittpunkt der Produkte aus Dichtefunktionen und a-priori Klassenwahrscheinlichkeiten. Zum Zeitpunkt t_2 verändern sich

1. $f_{X|Y}(x|y_1)$ aufgrund einer Verschiebung des Erwartungswertes in der ersten Klasse: $X|(Y = y_1) \sim \mathcal{N}(0.5, 0.5)$ (Abbildung 2.1),
2. die a-priori Klassenwahrscheinlichkeiten: $p^{(1)} = 0.8$, $p^{(2)} = 0.2$ (Abbildung 2.2),

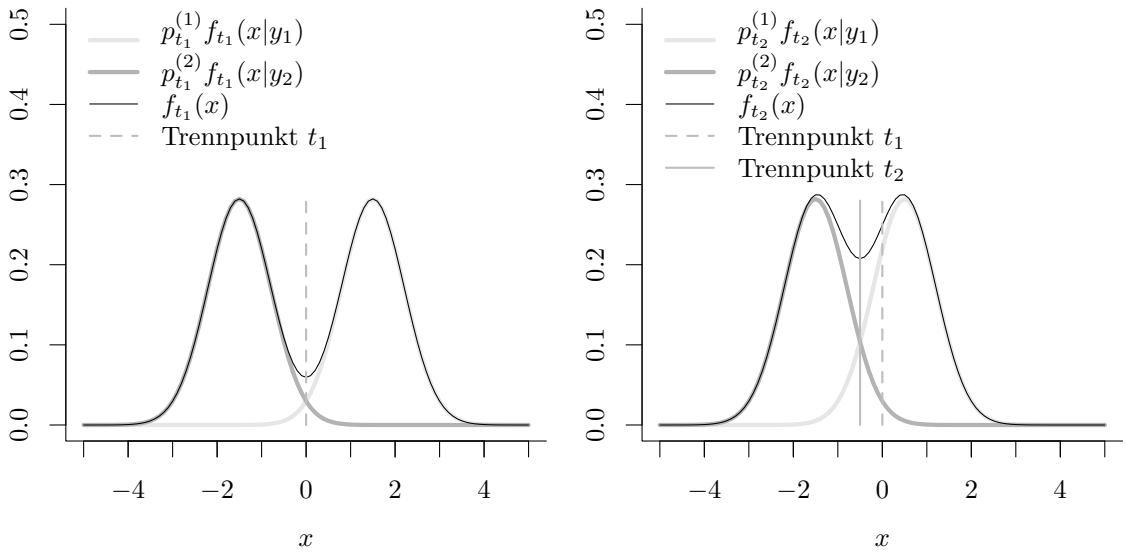


Abbildung 2.1: Eindimensionales Beispiel für *covariate shift* (in Anlehnung an Moreno-Torres et al. (2012, S. 525)): Die klassenbedingte Verteilung der ersten Klasse ändert sich von $X|(Y = y_1) \sim \mathcal{N}(1.5, 0.5)$ zu $X|(Y = y_1) \sim \mathcal{N}(0.5, 0.5)$, was eine Veränderung der Verteilung von X nach sich zieht. Als Konsequenz ändern sich die a-posteriori Wahrscheinlichkeiten der Klassenzugehörigkeit und demnach die Entscheidungsgrenze.

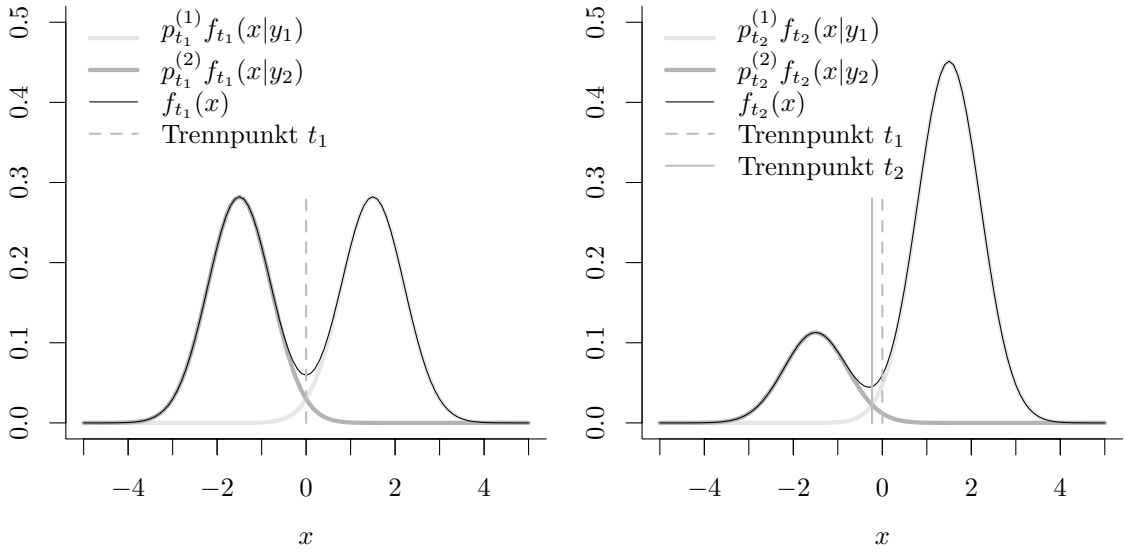


Abbildung 2.2: Eindimensionales Beispiel für *prior probability shift* (in Anlehnung an Moreno-Torres et al. (2012, S. 525 f.)): Die a-priori Klassenwahrscheinlichkeiten ändern sich von $p_{t_1}^{(1)} = p_{t_1}^{(2)} = 0.5$ zu $p_{t_2}^{(1)} = 0.8$ und $p_{t_2}^{(2)} = 0.2$. Als Konsequenz ändern sich die Verteilung von X, die a-posteriori Wahrscheinlichkeiten der Klassenzugehörigkeit und demnach die Entscheidungsgrenze.

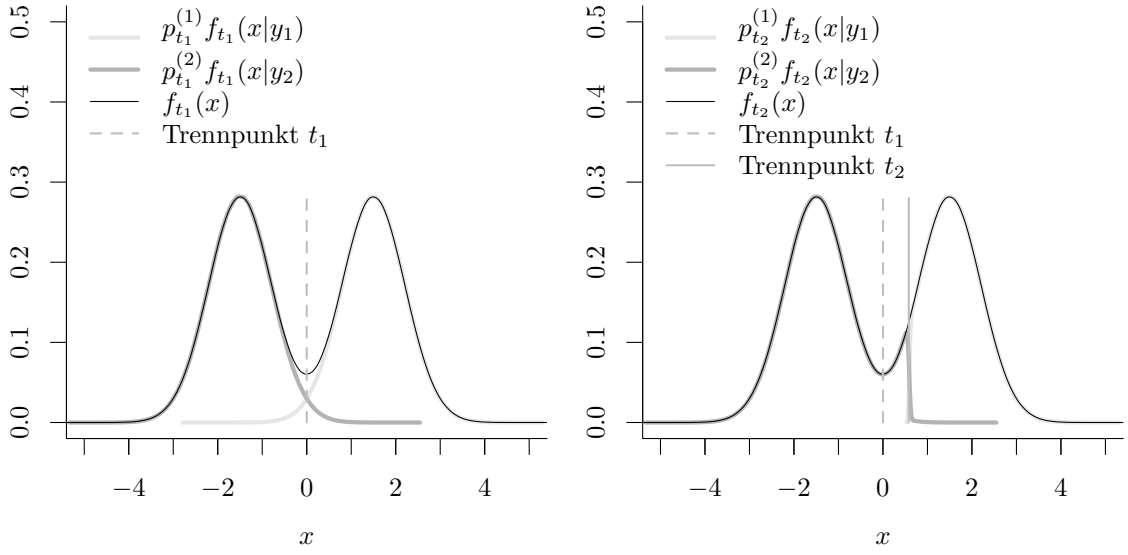


Abbildung 2.3: Eindimensionales Beispiel für $P_{t_1}(Y = y_c|X = x) \neq P_{t_2}(Y = y_c|X = x)$: Zum Zeitpunkt t_1 gilt: $X|(Y = y_1) \sim \mathcal{N}(1.5, 0.5)$ und $X|(Y = y_2) \sim \mathcal{N}(-1.5, 0.5)$. Zum Zeitpunkt t_2 ändert sich die Klassenzugehörigkeit der 5 % kleinsten Werte der Verteilung der ersten Klasse, d. h. die Entscheidungsgrenze verschiebt sich. Entsprechend der Anteile in jeder Klasse verändern sich die a-priori Klassenwahrscheinlichkeiten von $p_{t_1}^{(1)} = p_{t_1}^{(2)} = 0.5$ zu $p_{t_2}^{(1)} = 0.45$ und $p_{t_2}^{(2)} = 0.55$. Die gesamte Verteilung von X bleibt dadurch unberührt, trotz allem ändert sich die Entscheidungsgrenze.

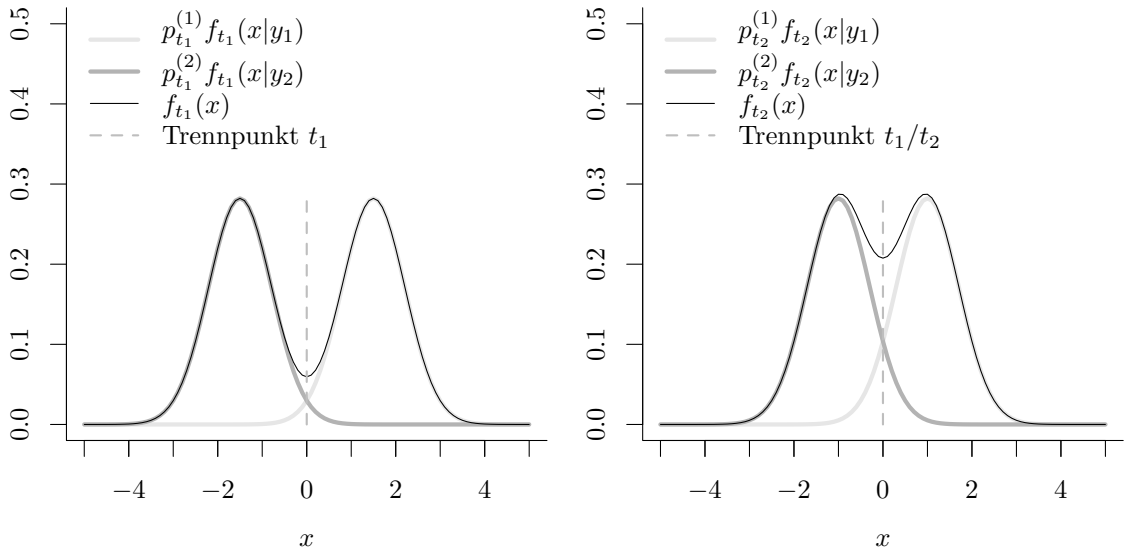


Abbildung 2.4: Eindimensionales Beispiel für $P_{t_1}(Y = y_c|X = x) = P_{t_2}(Y = y_c|X = x)$ (in Anlehnung an Moreno-Torres et al. (2012, S. 523)): Die klassenbedingten Verteilungen ändern sich von Zeitpunkt t_1 zu t_2 von $X|(Y = y_1) \sim \mathcal{N}(1.5, 0.5)$ und $X|(Y = y_2) \sim \mathcal{N}(-1.5, 0.5)$ zu $X|(Y = y_1) \sim \mathcal{N}(1, 0.5)$ und $X|(Y = y_2) \sim \mathcal{N}(-1, 0.5)$. Aufgrund der symmetrischen Verschiebung der Erwartungswerte bleibt jedoch die Trenngerade unverändert.

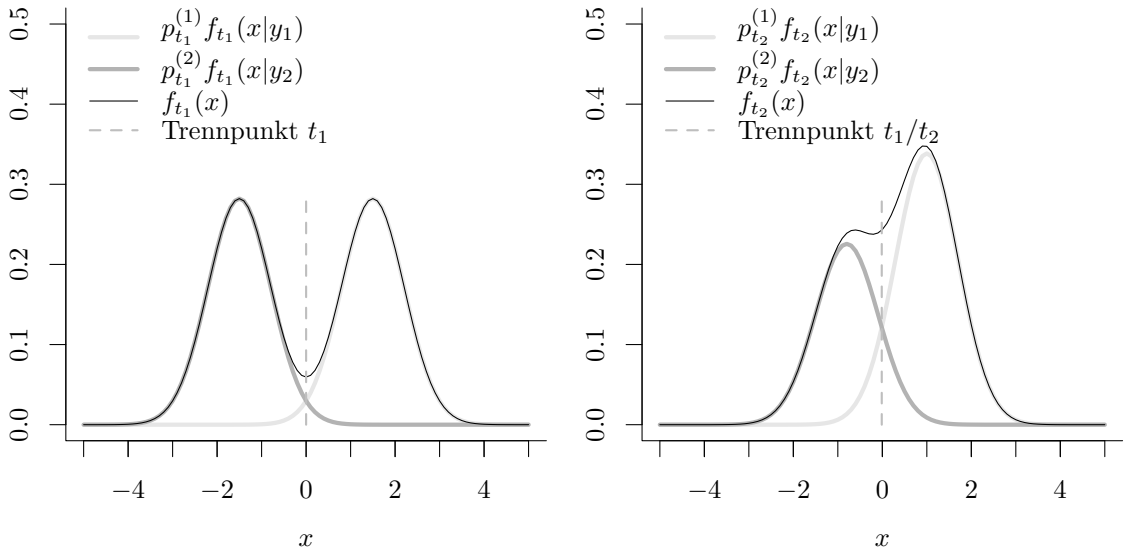


Abbildung 2.5: Eindimensionales Beispiel für $P_{t_1}(Y = y_c|X = x) = P_{t_2}(Y = y_c|X = x)$: Die a-priori Klassenwahrscheinlichkeiten ändern sich von $p_{t_1}^{(1)} = p_{t_1}^{(2)} = 0.5$ zu $p_{t_2}^{(1)} = 0.6$ und $p_{t_2}^{(2)} = 0.4$. Gleichzeitig ändern sich die klassenbedingten Verteilungen von Zeitpunkt t_1 zu t_2 von $X|(Y = y_1) \sim \mathcal{N}(1.5, 0.5)$ und $X|(Y = y_2) \sim \mathcal{N}(-1.5, 0.5)$ zu $X|(Y = y_1) \sim \mathcal{N}(1, 0.5)$ und $X|(Y = y_2) \sim \mathcal{N}(-0.8, 0.5)$. Die a-posteriori Wahrscheinlichkeiten der Klassenzugehörigkeit und demnach auch die Trenngerade bleiben jedoch unverändert.

3. $f_{X|Y}(x|y_1)$ und $f_{X|Y}(x|y_2)$ aufgrund einer Veränderung der Klassenzuordnung und gleichzeitig der a-priori Klassenwahrscheinlichkeiten: $p^{(1)} = 0.45$, $p^{(2)} = 0.55$ (Abbildung 2.3),

4. $f_{X|Y}(x|y_1)$ und $f_{X|Y}(x|y_2)$ aufgrund einer Verschiebung der Erwartungswerte: $X|(Y = y_1) \sim \mathcal{N}(1, 0.5)$, $X|(Y = y_2) \sim \mathcal{N}(-1, 0.5)$ (Abbildung 2.4),

5. $f_{X|Y}(x|y_1)$ und $f_{X|Y}(x|y_2)$ aufgrund gleichzeitiger Verschiebung der Erwartungswerte $X|(Y = y_1) \sim \mathcal{N}(1, 0.5)$, $X|(Y = y_2) \sim \mathcal{N}(-0.8, 0.5)$ und der a-priori Klassenwahrscheinlichkeiten: $p^{(1)} = 0.6$, $p^{(2)} = 0.4$ (Abbildung 2.5).

In den ersten drei Beispielen (Abbildungen 2.1–2.3) verändert sich die Entscheidungsgrenze als Konsequenz veränderter Verteilungen der Einflussvariablen und/oder a-priori Klassenwahrscheinlichkeiten (oder direkt aufgrund sich ändernder Klassen). Die letzten beiden Beispiele (Abbildungen 2.4 und 2.5) zeigen, dass nicht zwangsläufig eine Veränderung der Klassifikationsgrenze resultieren muss. Wie oben bereits erwähnt, sind solche Beispiele jedoch eher theoretisch und in Datensituationen in der Praxis recht unwahrscheinlich.

Ebenfalls ist das Beispiel aus Abbildung 2.3 recht synthetisch. Im Falle von normalverteilten Einflussgrößen wird es in der Praxis nicht passieren, dass sich $P(Y = y_c|\boldsymbol{X} = \boldsymbol{x})$, aber nicht $f_{\boldsymbol{X}}(\boldsymbol{x})$ ändert. Dieses Beispiel wurde lediglich aus Vergleichbarkeit zu den anderen Beispielen gewählt. In der Praxis wäre ein *class definition change*/*conditional change*/*functional*

relation change/pure class drift (vgl. Tabelle 2.2) jedoch zum Beispiel bei kategorialen Variablen denkbar.

In Tabelle 2.2 sind einige in der Literatur definierten verschiedenen Änderungen des concepts inklusive Bezeichnung zusammengefasst und gegenübergestellt. Mit Tabelle 2.3 folgt zudem eine Zusammenfassung der verschiedenen Bezeichnungen für verschiedene Arten von concept drift inklusive variierender Definition in der Literatur. Auch anhand dieser Tabellen (neben Tabelle 2.1) werden die Abweichungen und Überschneidungen von Definitionsansätzen der verschiedenen AutorInnen deutlich.

Zusammenfassend gibt es einige Überschneidungen der Definition von *concept drift* in der Literatur. Letztendlich führen jedoch alle AutorInnen eigene Definitionen und Bezeichnungen ein, um concept drift zu beschreiben. In einigen Veröffentlichungen wird zwischen verschiedenen „Auslösern“ unterschieden, in anderen werden allgemeinere und weitreichende Definitionen eingeführt (vgl. Tabelle 2.1).

In der Praxis sind in den allermeisten Fällen die a-posteriori Wahrscheinlichkeiten der Klassenzugehörigkeit von einem Drift betroffen, unabhängig davon, ob als Konsequenz oder direkt. Dies zieht zwangsläufig eine Verschlechterung der Güte des betrachteten Klassifikationsmodells nach sich. Daher wird im Folgenden jegliche Änderung des *concepts* als *concept drift* bezeichnet. Genauer also jeder Fall, in dem sich die zugrunde liegende Verteilung bestehend aus einer oder mehrerer der Größen $f_{\boldsymbol{X},Y}(\boldsymbol{x}, y_c)$, $P(Y = y_c)$, $f_{\boldsymbol{X}}(\boldsymbol{x})$, $f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c)$, $P(Y = y_c|\boldsymbol{X} = \boldsymbol{x})$ bzw. die datengenerierende Funktion ändert.

Tabelle 2.2: Zusammenfassung von in der Literatur definierten (bzw. beschriebenen) verschiedenen Änderungen des concepts inklusive Bezeichnung und Verweise auf veranschaulichende Beispiele.

Änderung	Abbildung	Bezeichnung	Quelle
$f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$	2.1, 2.2, 2.3, 2.4, 2.5	*concept drift*	Gama et al. (2014)
			Webb et al. (2016)
		dataset shift	Moreno-Torres et al. (2012)
$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$	2.1, 2.2, 2.4, 2.5	*covariate drift*	Webb et al. (2016)
		covariate shift	Shimodaira (2000)
$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2	*dual change*	Gao et al. (2007)
		virtual drift	Gama et al. (2014)
$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5	*covariate shift*	Bickel et al. (2009)
			Yamazaki et al. (2007)
		feature change	Gao et al. (2007)
		population drift	Alaiz-Rodríguez und Japkowicz (2008)
		pure covariate drift	Webb et al. (2016)
		sample selection bias/ covariate shift	Huang et al. (2007)
		virtual concept drift	Delany et al. (2005)
		virtual drift	Gama et al. (2014)
$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$ und $\boldsymbol{X} \rightarrow Y$	2.4, 2.5	*covariate shift*	Moreno-Torres et al. (2012)
			Storkey (2009)
$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ (und $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2	*virtual drift*	Gama et al. (2014)
oder $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$)	2.4, 2.5		
$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$	2.2, 2.3, 2.5	*class prior change*	Yamazaki et al. (2007)
		concept drift	Hoens et al. (2012)
			Žliobaitė (2010)
		population drift	Kelly et al. (1999)
$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $f_{t_1}(\boldsymbol{x} \vert y_c) = f_{t_2}(\boldsymbol{x} \vert y_c)$	2.2	*change in class distribution*	Alaiz-Rodríguez und Japkowicz (2008)
$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.5	*virtual concept drift*	Hoens et al. (2012)
$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $f_{t_1}(\boldsymbol{x} \vert y_c) = f_{t_2}(\boldsymbol{x} \vert y_c)$ und $Y \rightarrow \boldsymbol{X}$	2.2	*prior probability shift*	Moreno-Torres et al. (2012)
			Storkey (2009)
$f_{t_1}(\boldsymbol{x} \vert y_c) \neq f_{t_2}(\boldsymbol{x} \vert y_c)$	2.1, 2.3, 2.4, 2.5	*concept drift*	Hoens et al. (2012)
			Žliobaitė (2010)
		population drift	Kelly et al. (1999)
$f_{t_1}(\boldsymbol{x} \vert y_c) \neq f_{t_2}(\boldsymbol{x} \vert y_c)$ und $P_{t_1}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \vert \boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5	*virtual concept drift*	Hoens et al. (2012)

Fortsetzung auf der nächsten Seite

Änderung	Abbildung	Bezeichnung	Quelle
$f_{t_1}(\boldsymbol{x}\|y_c) \neq f_{t_2}(\boldsymbol{x}\|y_c)$ und $P_{t_1}(Y = y_c) = P_{t_2}(Y = y_c)$ und $Y \rightarrow \boldsymbol{X}$	2.1, 2.4	*concept shift*	Moreno-Torres et al. (2012)
$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3	*class drift*	Webb et al. (2016)
		concept drift	Delany et al. (2005) Hoens et al. (2012) Žliobaitė (2010)
		covariate shift	Cieslak und Chawla (2009)
		population drift	Kelly et al. (1999)
		real concept drift	Hoens et al. (2012) Tsymbal (2004)
$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$	2.3	*class definition change*	Alaiz-Rodríguez und Japkowicz (2008)
		conditional change	Gao et al. (2007)
		functional relation change	Yamazaki et al. (2007)
		pure class drift	Webb et al. (2016)
$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$ und $\boldsymbol{X} \rightarrow Y$	2.3	*concept shift*	Moreno-Torres et al. (2012)
		functional relation change	Yamazaki et al. (2007)
$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$ oder $f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$	2.1, 2.2, 2.3	*real concept drift*	Gama et al. (2014)
$P_{t_1}(Y = y_c) = 0$ und $P_{t_2}(Y = y_c) > 0$		neue Klasse	Webb et al. (2016)

Tabelle 2.3: Zusammenfassung verschiedener Bezeichnungen für verschiedene Arten von concept drift inklusive variierender Definition (bzw. Beschreibung) in der Literatur und Verweise auf veranschaulichende Beispiele. Die unterstrichenen Bezeichnungen werden auf Seite 25 gegenübergestellt.

Bezeichnung	Quelle	Änderung	Abbildung
change in class distribution	Alaiz-Rodríguez und Japkowicz (2008)	$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $f_{t_1}(\boldsymbol{x}\|y_c) = f_{t_2}(\boldsymbol{x}\|y_c)$	2.2
Changing Environments	Alaiz-Rodríguez und Japkowicz (2008)	„The fundamental assumption of supervised learning is that the joint probability distribution $[P(\boldsymbol{X}, Y)]$ will remain unchanged between training and testing. There are, however, some mismatches that are likely to appear in practice [...].“: $f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$	2.1, 2.2, 2.3, 2.4, 2.5
class definition change	Alaiz-Rodríguez und Japkowicz (2008)	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$	2.3
class drift	Webb et al. (2016)	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
class prior change	Yamazaki et al. (2007)	$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$	2.2, 2.3, 2.5
concept change	Schlimmer und Granger (1986)	„Frequently, however, a derived description of a useful concept is disrupted by some change which requires its revision.“	
concept drift	Delany et al. (2005)	„Changes in the hidden context can induce changes in the target concept, which is generally known as concept drift“: $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
	Gama et al. (2014)	$f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$	2.1, 2.2, 2.3, 2.4, 2.5
	Hand (2006)	„[...] changes to the definitions of the classes“	
	Hoens et al. (2012)	„Concept drift is said to occur when the underlying function [...] changes over time.“:	
		$f_{t_1}(\boldsymbol{x}\|y_c) \neq f_{t_2}(\boldsymbol{x}\|y_c)$	2.1, 2.3, 2.4, 2.5
		$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
		$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$	2.2, 2.3, 2.5
	Schlimmer und Granger (1986)	„Frequently, however, a derived description of a useful concept is disrupted by some change which requires its revision.“	
	Webb et al. (2016)	$f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$	2.1, 2.2, 2.3, 2.4, 2.5
	Widmer und Kubat (1996)	„[...] changes in the target concepts [...]“	
	Žliobaitė (2010)	$f_{t_1}(\boldsymbol{x}\|y_c) \neq f_{t_2}(\boldsymbol{x}\|y_c)$	2.1, 2.3, 2.4, 2.5
		$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
		$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$	2.2, 2.3, 2.5
concept shift	Moreno-Torres et al. (2012)	$P_{t_1}(Y = y_c\|\boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$ und $\boldsymbol{X} \to Y$	2.3
		$f_{t_1}(\boldsymbol{x}\|y_c) \neq f_{t_2}(\boldsymbol{x}\|y_c)$ und $P_{t_1}(Y = y_c) = P_{t_2}(Y = y_c)$ und $Y \to \boldsymbol{X}$	2.1, 2.4
	Salganicoff (1997)	„Several aspects of the learning problem can vary, including the mapping to be learned [...]“	

Fortsetzung auf der nächsten Seite

Bezeichnung	Quelle	Änderung	Abbildung
conditional change	Gao et al. (2007)	$P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$	2.3
covariate drift	Webb et al. (2016)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$	2.1, 2.2, 2.4, 2.5
covariate shift	Bickel et al. (2009)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
	Cieslak und Chawla (2009)	$P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
	Huang et al. (2007)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
	Moreno-Torres et al. (2012)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$ und $\boldsymbol{X} \rightarrow Y$	2.4, 2.5
	Shimodaira (2000)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$	2.1, 2.2, 2.4, 2.5
	Storkey (2009)	„[...] is when only the distributions of covariates $\boldsymbol{x}$ change and everything else is the same.“:	
		$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$ und $\boldsymbol{X} \rightarrow Y$	2.4, 2.5
	Yamazaki et al. (2007)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
dataset shift	Moreno-Torres et al. (2012)	$f_{t_1}(\boldsymbol{x}, y_c) \neq f_{t_2}(\boldsymbol{x}, y_c)$	2.1, 2.2, 2.3, 2.4, 2.5
	Storkey (2009)	„[...] deals with the business of relating information in (usually) two closely related environments to help with the prediction in one given the data in the other(s).“	
dual change	Gao et al. (2007)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2
feature change	Gao et al. (2007)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
fracture between the data	Moreno-Torres et al. (2013)	„we have data from one laboratory (dataset A), and derive a classifier from it that can predict its category accurately. We are then presented with data from a second laboratory (dataset B). This second dataset is not accurately predicted by the classifier we had previously built due to a fracture between the data of both laboratories.“	
functional relation change	Yamazaki et al. (2007)	$P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$	2.3
neue Klasse	Webb et al. (2016)	$P_{t_1}(Y = y_c) = 0$ und $P_{t_2}(Y = y_c) > 0$	
population drift	Alaiz-Rodríguez und Japkowicz (2008)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
	Hand (2006)	„A fundamental assumption of the classical paradigm is that the various distributions involved do not change over time. In fact, in many applications this is unrealistic and the population distributions are nonstationary.“	
	Kelly et al. (1999)	„When the population distribution can change over time we say it is subject to population drift.“:	
		$P_{t_1}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c \mid \boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
		$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$	2.2, 2.3, 2.5
		$f_{t_1}(\boldsymbol{x} \mid y_c) \neq f_{t_2}(\boldsymbol{x} \mid y_c)$	2.1, 2.3, 2.4, 2.5

Fortsetzung auf der nächsten Seite

Bezeichnung	Quelle	Änderung	Abbildung
prior probability shift	Moreno-Torres et al. (2012)	$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $f_{t_1}(\boldsymbol{x}\|y_c) = f_{t_2}(\boldsymbol{x}\|y_c)$ und $Y \rightarrow \boldsymbol{X}$	2.2
	Storkey (2009)	„[...] is when only the distribution over y changes and everything else stays the same.“:	
		$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $f_{t_1}(\boldsymbol{x}\|y_c) = f_{t_2}(\boldsymbol{x}\|y_c)$ und $Y \rightarrow \boldsymbol{X}$	2.2
pure class drift	Webb et al. (2016)	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$	2.3
pure covariate drift	Webb et al. (2016)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
real concept drift	Gama et al. (2014)	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$ und $f_{t_1}(\boldsymbol{x}) = f_{t_2}(\boldsymbol{x})$ oder $f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$	2.1, 2.2, 2.3
	Hoens et al. (2012)	$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
	Tsymbal (2004)	„Hidden changes in context may [...] be a cause of a change of target concept [...]“:	
		$P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2, 2.3
	Widmer und Kubat (1993)	„Real concept drift reflects real changes in the world [...]“	
sample selection bias	Huang et al. (2007)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
sampling shift	Salganicoff (1997)	„Several aspects of the learning problem can vary, [...] and the sampling distribution that governs the input-space location of exemplars that make up the learning set.“	
virtual concept drift	Delany et al. (2005)	„Hidden changes [...] may also cause a change in the underlying data distribution. Even if the target concept remains the same but the data distribution changes, a model rebuild may be necessary as the model's error may no longer be acceptable.“:	
		$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
	Hoens et al. (2012)	$P_{t_1}(Y = y_c) \neq P_{t_2}(Y = y_c)$ und $P_{t_1}(Y = y_c\|\boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{x})$	2.5
		$f_{t_1}(\boldsymbol{x}\|y_c) \neq f_{t_2}(\boldsymbol{x}\|y_c)$ und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.4, 2.5
	Tsymbal (2004)	„Even if the target concept remains the same, and it is only the data distribution that changes, this may often lead to the necessity of revising the current model [...]. The necessity in the change of current model due to the change of data distribution“	
	Widmer und Kubat (1993)	„Virtual concept drift [...] does not occur in reality but, rather, in the computer model reflecting this reality. In a practical setting, this kind of effect can emerge when the representation language is poor and fails to identify all relevant features, or when the order of training examples for learning is skewed, so that different types of instances are not evenly distributed over the training sequence.“	
virtual drift	Gama et al. (2014)	$f_{t_1}(\boldsymbol{x}) \neq f_{t_2}(\boldsymbol{x})$ (und $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$	2.1, 2.2
		oder $P_{t_1}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c\|\boldsymbol{X} = \boldsymbol{x})$)	2.4, 2.5

real vs. virtual drift In der Literatur werden zudem häufig die beiden Begriffe *real concept drift* und *virtual concept drift* (bzw. auch *virtual drift*) zur Differenzierung aufgeführt (vgl. Tabelle 2.3). Einigkeit der AutorInnen bezüglich der Verwendung dieser Begriffe besteht darin, dass mit *real concept drift* Situationen bezeichnet werden, in denen sich die Entscheidungsgrenzen der Klassifikationsverfahren aufgrund veränderter a-posteriori Wahrscheinlichkeiten der Klassenzugehörigkeiten verschieben: $P_{t_1}(Y = y_c|\boldsymbol{X} = \boldsymbol{x}) \neq P_{t_2}(Y = y_c|\boldsymbol{X} = \boldsymbol{x})$. In diesen Fällen ist demnach zwingend eine Anpassung des Modells notwendig, falls sich die Güte des betrachteten Klassifikationsmodells nicht deutlich verschlechtern soll.

Die Definition von *virtual concept drift* ist nicht so eindeutig und es herrscht Uneinigkeit bezüglich der Bedeutung. In Tabelle 2.3 sind einige in der Literatur existierende Definitionen zusammengefasst. Es kann differenziert werden zwischen jenen AutorInnen, die den Fokus auf die Verteilung von $Y|\boldsymbol{X}$ und jenen, die den Fokus auf die Verteilung von $\boldsymbol{X}|Y$ oder $\boldsymbol{X}$ richten. Im ersten Fall wird dieser Drift als eine Art Gegensatz zum *real concept drift* beschrieben in dem Sinne, dass sich die a-priori Wahrscheinlichkeiten der Klassen oder die klassenbedingten Wahrscheinlichkeiten (Hoens et al., 2012, S. 91 f.) und demnach auch die Verteilung der Einflussvariablen (Delany et al., 2005, S. 188) ändern können, dies jedoch keine Veränderung der Entscheidungsgrenzen nach sich zieht: $P_{t_1}(Y = y_c|\boldsymbol{X} = \boldsymbol{x}) = P_{t_2}(Y = y_c|\boldsymbol{X} = \boldsymbol{x})$. Andere AutorInnen wie Tsymbal (2004, S. 2) und Gama et al. (2014, S. 5) legen den Fokus auf eine Änderung der gesamten Verteilung von $\boldsymbol{X}$.

Die Entscheidungsgrenzen können sich beim *virtual drift* ändern (bei gleichzeitigem Auftreten von *real drift*), müssen es aber nicht zwangsläufig (Gama et al., 2014, S. 5).

Gama et al. (2014, S. 5) fassen zusammen, dass dieser *virtual drift* von anderen AutorInnen unter anderem auch als *temporary drift*, *sampling shift* und *feature change* (s. Tabelle 2.1 und Tabelle 2.2) bezeichnet wird. Dadurch ergeben sich noch weitere Uneindeutigkeiten in der Definition.

Tsymbal (2004, S. 2) erwähnt, dass virtual und real concept drift zusammen auftreten können, dass mit virtual drift jedoch ebenfalls Situationen bezeichnet werden, in denen eine Anpassung des Modells nötig ist und es daher praktisch irrelevant ist, ob es sich um einen virtual oder einen real concept drift handelt.

Žliobaitė (2010, S. 4) argumentiert, dass sie im Gegensatz zu vielen anderen AutorInnen nicht zwischen *real* und *virtual drift* unterscheidet, da aufgrund von Satz 1 die Größen $f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c)$ und $P(Y = y_c|\boldsymbol{X} = \boldsymbol{x})$ zusammenhängen. Diese Ansicht wird im Folgenden ebenfalls verfolgt.

Art des Drifts Neben der Differenzierung verschiedener „Auslöser“ eines Drifts (vgl. Seite 11 ff.) teilen einige AutorInnen concept drift zudem in verschiedene „Arten“ ein. Die Definitionen sind dabei meist qualitativer Natur und unterscheiden sich teilweise in der

Hinsicht, dass mit demselben Begriff unterschiedliche Arten beschrieben werden und die verschiedenen Arten teilweise von den AutorInnen unterschiedlich bezeichnet werden.

Hoens et al. (2012, S. 92 f.) führen die Differenzierung dabei unter dem Gesichtspunkt der *Geschwindigkeit des concept drifts* (speed of drift) ein. Sie heben hervor, dass ein concept drift entweder plötzlich (*sudden concept drift*) oder allmählich (*gradual concept drift*) erfolgt. Eine plötzliche Veränderung kennzeichnet sich dadurch, dass bis zu einem bestimmten Zeitpunkt t die Beobachtungen aus einer datengenerierenden Funktion stammen und sich ab diesem Zeitpunkt die datengenerierende Funktion bzw. zugrunde liegende Verteilung plötzlich ändert und demnach ein Konzept das andere vollständig ersetzt. Daher bezeichnen die Autoren diese Art auch als *concept change*. Beim *gradual concept drift* erfolgt ein weicherer Übergang von einer datengenerierenden Funktion (bzw. einem Konzept) zur anderen.

Gama et al. (2014, S. 5 f.) differenzieren noch weiter in *gradual drift* und *incremental drift*, wobei sich bei letzterem viele sich nur leicht unterscheidende zugrunde liegende Verteilungen ersetzen. Die AutorInnen erläutern dies anhand des Beispiels eines Sensors, der sich mit der Zeit immer mehr abnutzt und die Ergebnisse daher ungenauer werden. Bereits zuvor wurde von Žliobaitė (2010, S. 6) angemerkt, dass der Begriff gradual drift, wie er in der Literatur verwendet wird, als Oberbegriff zu sehen sei, und der Unterschied zwischen den zwei genannten Arten hervorgehoben. Im Gegensatz zu einem incremental drift gibt es bei einem gradual drift zwei datengenerierende Funktionen, die sich eine Zeit lang überschneiden, sodass Beobachtungen aus beiden Verteilungen generiert werden. Nach und nach wird die Erste vollständig durch die Zweite ersetzt. Dies bedeutet, dass zunächst das erste concept dominiert. Mit der Zeit steigt jedoch die Wahrscheinlichkeit, mit der Beobachtungen der zweiten Verteilung folgen, bis die erste Verteilung vollständig ersetzt wurde. Als Vergleich zum ersten Beispiel würde dies bedeuten, dass der Sensor durch einen anderen ersetzt wird, eine Zeit lang jedoch beide gleichzeitig in Betrieb sind und die Wahrscheinlichkeit, dass der zweite Sensor herangezogen wird mit der Zeit immer mehr zunimmt.

Ein weiteres Muster stellen *reoccurring concepts* (Hoens et al., 2012, S. 92)/*reoccurring contexts* (Žliobaitė, 2010, S. 6) dar. Bei diesen treten zugrunde liegende Verteilungen nach einer Weile erneut auf, nachdem sie von anderen Verteilungen ersetzt wurden. Es kann somit eine beschränkte Menge an datengenerierenden Funktionen geben, aus denen Beobachtungen im Datenstrom generiert oder beobachtet werden, und die sich untereinander als dominante Funktionen abwechseln. Diese Art von concept drift kann auch übergeordnet betrachtet werden, da die einzelnen Wechsel plötzlich oder allmählich erfolgen können. Die Muster können dabei zufällig sein. Es müssen keine festen Zeitpunkte oder regelmäßige Zeitabschnitte vorliegen, in denen ein Wechsel erfolgt.

Die verschiedenen erläuterten Arten sind in Abbildung 2.6 anhand eines eindimensionalen Beispiels, bei welchem der concept drift auf die Veränderung des Erwartungswertes zurückzuführen ist, veranschaulicht.

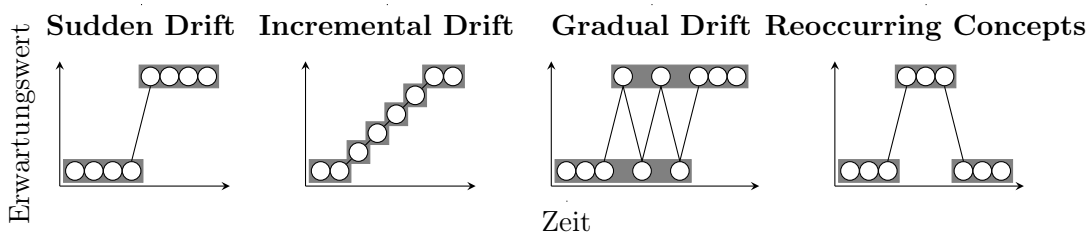


Abbildung 2.6: Verschiedene Arten von concept drift (in Anlehnung an Gama et al. (2014, S. 6) und Žliobaitė (2010, S. 7)): Betrachtung einer eindimensionalen Verteilung. Es kann sich beispielsweise der Erwartungswert (als Teil des *concepts*) der Verteilung ändern.

Als Ergänzung der qualitativen Beschreibungen der Arten des Drifts schlagen Webb et al. (2016, S. 970 ff.) formale Definitionen vor, um zwischen den vorgestellten Arten zu differenzieren. Dazu müssen zunächst einige Größen definiert werden:

Definition 1.

- $D(t, t+s)$ sei eine Distanzfunktion, welche einen positiven numerischen Wert für die Stärke des Unterschieds der Verteilungen zum Zeitpunkt t und $t+s$ liefert (Webb et al., 2016, S. 969). Als Beispiele erwähnen die Autoren die Kullback-Leibler-Divergenz (Kullback und Leibler, 1951) oder den Hellingerabstand (Hellinger, 1909; Hoens et al., 2011).
- Der Zeitraum $[t, u]$ eines stabilen Konzepts sei damit definiert durch (Webb et al., 2016, S. 971):

$$u \geq t + \phi \text{ und } \forall s \in (0, u - t] : D(t, t + s) = 0,$$

  wobei $\phi \geq 0$ eine positive Konstante ist, die definiert, wie lange der Zeitraum sein muss, in welchem die zugrunde liegende Verteilung unverändert bleibt, damit von einem stabilen Konzept gesprochen werden kann. Diese Konstante ist je nach Anwendung festzulegen.
- Start- und Endzeitpunkt eines solchen a-ten stabilen Zeitraums (Konzepts) in einem Datenstrom seien mit S_a und E_a bezeichnet (Webb et al., 2016, S. 971):

$$S_a = \begin{cases} \min\{t \,|\, \forall s \in (0, \phi] : D(t, t + s) = 0\}, & a = 1, \\ \min\{t \,|\, t > E_{a-1} \wedge \forall s \in (0, \phi] : D(t, t + s) = 0\}, & a > 1, \end{cases}$$
$$E_a = \max\{t \,|\, \forall s \in (0, t - S_a] : D(S_a, S_a + s) = 0\}.$$

Mit diesen Vorüberlegungen tritt ein *abrupt drift*/*sudden drift* zwischen dem a-ten und $(a+1)$-ten stabilen Konzept auf, falls (Webb et al., 2016, S. 975)

$$S_{a+1} - E_a \leq \rho,$$

wobei $\rho \in \mathbb{N}$ eine positive Konstante ist, welche definiert wie lange der Zeitraum höchstens sein darf, damit von einem sudden drift gesprochen werden kann. Zwischen dem a-ten und $(a+1)$-ten stabilen Konzept können beliebig viele Zeitpunkte vergehen, innerhalb derer ein Drift erfolgen kann (abhängig von ρ).

Eine Veränderung zwischen concept a und concept $a+1$ bezeichnen Webb et al. (2016, S. 977) hingegen als *gradual drift*, falls die Distanzen der Verteilungen innerhalb aller Intervalle der Länge κ (gleitende Fenster) in $[E_a, S_{a+1}]$ maximal ζ betragen:

$$\forall t \in [E_a, S_{a+1} - \kappa] : D(t, t+\kappa) \leq \zeta.$$

Die Konstante $\zeta > 0$ beschreibt dabei, wie groß diese Distanz zwischen den Verteilungen sein darf, damit die Veränderung als gradual bezeichnet werden kann.

Incremental drift wird hingegen folgendermaßen definiert (Webb et al., 2016, S. 977):

$$\forall t \in (E_a, S_{a+1})\, \forall u \in (t, S_{a+1}) : D(E_a, t) \leq D(E_a, u) \wedge D(t, S_{a+1}) \geq D(u, S_{a+1}).$$

Anschaulich bedeutet dies, dass zu jedem Zeitpunkt t im Intervall (E_a, S_{a+1}) (Zeitpunkte zwischen den stabilen Konzepten a und $a+1$) die Distanz der Verteilung des Zeitpunktes t zu jener des Konzepts a zunimmt und gleichzeitig zu jener des Konzepts $a+1$ abnimmt.

Den zuvor beschriebenen, von Žliobaitė (2010, S. 6) benannten *gradual drift* hingegen bezeichnen Webb et al. (2016, S. 977) als *probabilistic drift*. Die Veränderung kann dabei anhand von Wahrscheinlichkeiten ausgedrückt werden. Bezeichne f_t mit $f_0 = 0$ und $f_{S_{a+1}-E_a} = 1$ eine monoton wachsende Funktion, welche die Wahrscheinlichkeit beschreibt, mit der Beobachtungen aus der neuen Verteilung realisiert und beobachtet werden. Damit gilt formal (Webb et al., 2016, S. 977):

$$\forall t \in [0, S_{a+1} - E_a]\ \forall o : P_{E_a+t}(\mathcal{X} = o) = (1 - f_t) P_{E_a}(\mathcal{X} = o) + f_t P_{S_{a+1}}(\mathcal{X} = o).$$

Die Verteilung zum Zeitpunkt t lässt sich demnach als Linearkombination aus der Verteilung des a-ten und $(a+1)$-ten stabilen Konzepts darstellen. Abhängig davon wie f aussieht, kann der probabilistic drift ein gradual oder incremental drift sein. Webb et al. (2016, S. 977 f.) stellen damit den formalen Unterschied zu incremental drift (oder je nach Definition gradual drift) heraus, den die meisten anderen AutorInnen (u. a. Žliobaitė, 2010, S. 6; Huang et al., 2013, S. 75) außer Acht lassen. Sie erläutern, dass ein incremental drift ein probabilistic drift sein kann und umgekehrt, dass aber formal zwischen beiden Arten unterschieden werden sollte.

Die letzte vorgestellte Art waren *reoccurring concepts*. Webb et al. (2016, S. 978) definieren diese Art für zwei Verteilungen (Konzepte) a und b ebenfalls auf Basis der Distanzfunktion:

$$\exists a\ \exists b : a \neq b \wedge D(S_a, S_b) = 0.$$

Die Autoren beschreiben weitere (Unter-)Arten des Überbegriffs *concept drift* und liefern formale Definitionen zur Differenzierung der verschiedenen Arten. Dazu sei auf Webb et al. (2016) verwiesen. Zudem wird die Vereinheitlichung der Terminologie von Aschersleben (2016, S. 10 ff.) zusammengefasst.

Da im Folgenden die expliziten formalen Unterschiede nicht im Fokus stehen, werden die von Žliobaitė (2010) vorgeschlagenen Begriffe verwendet. Es werden die vier Arten unterschieden, welche in Abbildung 2.6 anhand eines Beispiels veranschaulicht wurden.

3 Diskriminanzanalyseverfahren

Die *Diskriminanzanalyse* ist ein Klassifikationsverfahren, mithilfe dessen auf einer Stichprobe mit bekannter Klasseneinteilung eine Klassifikationsregel gelernt wird, mit welcher eine neue Beobachtung einer dieser bereits vorliegenden Klassen zugeordnet werden kann. Es handelt sich somit um ein Verfahren des *überwachten Lernens*.

Als Grundlage seien p verschiedene (quantitative) Merkmale und $M \geq 2$ Klassen betrachtet. Als Datengrundlage für die Anpassung der Klassifikationsregel liegt eine Stichprobe von n Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ vor, wobei jede Beobachtung $\boldsymbol{x}_i := (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p,\ i = 1, \ldots, n$, die Ausprägungen der p Merkmale auffasst. Die entsprechenden Zufallsvektoren der Realisationen $\boldsymbol{x}_i$ seien mit $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ bezeichnet.

Zunächst wird jedoch unterstellt, dass alle Beobachtungen aus der Stichprobe einer gemeinsamen zugrunde liegenden Verteilung entstammen, weshalb folgender Zufallsvektor betrachtet wird:

$$\boldsymbol{X} := \boldsymbol{X}_i, \quad i = 1, \ldots, n. \tag{3.1}$$

Die Verteilung der Klassen folgt einer Multinomialverteilung mit den Parametern n (Größe der Stichprobe) und a-priori Wahrscheinlichkeiten (Wahrscheinlichkeit, dass ein Element aus der Grundgesamtheit zu Gruppe c gehört) $p^{(c)} \in [0, 1]$ für die einzelnen Klassen $c = 1, \ldots, M$, und wird durch den Zufallsvektor $\boldsymbol{W}$ beschrieben. Genauer (Hartung et al., 1995, S. 209 f.):

$$\boldsymbol{W} = \left(W^{(1)}, \ldots, W^{(M)}\right) \sim Mult\left(n, p^{(1)}, \ldots, p^{(M)}\right). \tag{3.2}$$

Die einzelne Zufallsvariable $W^{(c)}$ beschreibt die Anzahl der vorliegenden Beobachtungen aus Klasse c in der Stichprobe. Die Wahrscheinlichkeit, dass genau $n^{(1)}$ Beobachtungen aus Klasse 1, $n^{(2)}$ Beobachtungen aus Klasse 2, usw. vorliegen, lässt sich durch

$$P(W^{(1)} = n^{(1)}, \ldots, W^{(M)} = n^{(M)}) = \frac{n!}{n^{(1)}! \cdots n^{(M)}!} \cdot \left(p^{(1)}\right)^{n^{(1)}} \cdots \left(p^{(M)}\right)^{n^{(M)}}$$

beschreiben (Hartung et al., 1995, S. 210; Fahrmeir et al., 1996a, S. 33 f.), wobei $n^{(c)} \in \{0, 1, \ldots, n\}$ für $c = 1, \ldots, M$ und

$$\sum_{c=1}^{M} n^{(c)} = n \quad \text{sowie} \quad \sum_{c=1}^{M} p^{(c)} = 1.$$

Es wird angenommen, dass alle $n^{(c)}$ Beobachtungen in einer Klasse c einer gemeinsamen identischen Verteilung, und zwar einer multivariaten Normalverteilung folgen. Die klassenbedingte Verteilung der Zufallsvektoren

$$\boldsymbol{X}|(W^{(c)} = n^{(c)},\ W^{(j)} = 0,\ j \neq c) =: \boldsymbol{X}^{(c)} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}\right) \tag{3.3}$$

besitzt demnach den Erwartungswertvektor $\boldsymbol{\mu}^{(c)}$ sowie die Kovarianzmatrix $\boldsymbol{\Sigma}^{(c)}$. Die Erwartungswertvektoren unterscheiden sich für die verschiedenen Klassen, sodass $\boldsymbol{\mu}^{(k)} \neq \boldsymbol{\mu}^{(j)}$ für $k \neq j$, $k, j \in \{1, \ldots, M\}$.

In den Kapiteln 6 und 8 wird die Tatsache, dass die Verteilung aus Klasse c (vgl. (3.3)) betrachtet wird, in den dort betrachteten Schätzfunktionen auch mit $\boldsymbol{X} \sim F_c$ bezeichnet.

Die zugehörige Dichtefunktion zu (3.3) sei mit $f^{(c)}(\boldsymbol{x})$ bezeichnet:

$$\begin{aligned} f^{(c)}(\boldsymbol{x}) :&= f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c) = f^{(c)}(\boldsymbol{x}; \boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}) \\ &= \frac{1}{\sqrt{(2\pi)^p}} \cdot \left|\boldsymbol{\Sigma}^{(c)}\right|^{-1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(c)}\right)^T \left(\boldsymbol{\Sigma}^{(c)}\right)^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}^{(c)}\right)\right). \end{aligned} \tag{3.4}$$

Es wird $\boldsymbol{X}^{(c)}$ betrachtet, falls eine Beobachtung aus Klasse c realisiert wird, also $g(\boldsymbol{x}) = c$ ist. Die Funktion g beschreibt, welche Klasse auftritt:

$$g : \mathbb{R}^p \to \{1, \ldots, M\}, \quad \boldsymbol{x} \mapsto g(\boldsymbol{x}). \tag{3.5}$$

In der Stichprobe liegen jeweils $n^{(c)}$ Beobachtungen in Klasse c vor, sodass $\sum_{c=1}^{M} n^{(c)} = n$. Diese $n^{(c)}$ Beobachtungen aus $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ seien mit $\boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n^{(c)})}^{(c)}$ bezeichnet.

Die Zufallsvariable Y beschreibt im Folgenden das Auftreten einer der M Klassen. $Y = y_c$ ist daher äquivalent zu dem Ereignis $W^{(c)} = 1 \wedge W^{(j)} = 0,\ j \neq c,\ j, c \in \{1, \ldots, M\}$, bei einer Beobachtung ($n = 1$). Daher entspricht der Wahrscheinlichkeit dieses Ereignisses $P(Y = y_c) = p^{(c)}$ die a-priori Wahrscheinlichkeit für Klasse c. Für den Zusammenhang mit der Zufallsvariable Y wird die Ausprägung y_c als Notation eingeführt. Diese ist jedoch äquivalent zur Klasse c.

Im Folgenden werden in Abschnitt 3.1 zunächst die Maximum-Likelihood Regel und die Bayes-Regel erläutert. Letztere bildet die Grundlage für die *Kanonische Diskriminanzanalyse* (Abschnitt 3.2), sofern als Verteilung der Merkmale die multivariate Normalverteilung bzw. Gauß'sche Verteilung unterstellt wird. Abschnitt 3.2.1 befasst sich mit der *Kanonischen Linearen Diskriminanzanalyse*, welche auf der Annahme identischer Kovarianzmatrizen in den Klassen beruht. In Abschnitt 3.2.2 wird mit der *Quadratischen Diskriminanzanalyse* ein weiteres Diskriminanzanalyseverfahren vorgestellt, welches etwas flexibler ist, da nicht nur lineare Trennungen möglich sind.

Abschnitt 3.3 befasst sich mit der *Diskriminanzanalyse nach Fisher* und damit einer alternativen Herleitung. Fisher (1936) entwickelt ein verteilungsfreies Diskriminanzanalyseverfahren, bei welchem anders als bei der Kanonischen Linearen Diskriminanzanalyse keine Normalverteilungsannahme unterstellt wird. Ein Vergleich beider Varianten folgt in Abschnitt 3.4.

3.1 Maximum-Likelihood Regel und Bayes-Regel

Eine recht intuitive Klassifikationsregel ergibt sich aus dem Maximum-Likelihood Prinzip, nach welchem eine Beobachtung $\boldsymbol{x}$ genau der Klasse c zugeordnet wird, aus welcher sie „mit größter Wahrscheinlichkeit“ stammt (Huberty, 1994, S. 45).

Falls die Zufallsvektoren $\boldsymbol{X}^{(c)}$ alle einer Verteilung mit Dichtefunktion $f^{(c)}$ folgen, wobei die Dichtefunktionen für alle Klassen dieselbe Form haben und sich lediglich in den Parametern unterscheiden, lässt sich die *Maximum-Likelihood Regel* zur Klassifikation einer Beobachtung $\boldsymbol{x}$ betrachten (Huberty, 1994, S. 46):

Ordne $\boldsymbol{x}$ der Klasse $k \in \{1, \ldots, M\}$ zu, für die gilt:

$$f^{(k)}(\boldsymbol{x}) > f^{(j)}(\boldsymbol{x}), \quad j \in \{1, \ldots, M\},\ k \neq j \quad \Leftrightarrow \quad k = \underset{c=1,\ldots,M}{\arg\max}\, f^{(c)}(\boldsymbol{x}) =: G_{\mathrm{ML}}(\boldsymbol{x}). \tag{3.6}$$

Eine Verallgemeinerung dieser Regel ist die *Bayes-Regel*, welche sich aus dem Satz von Bayes nach Thomas Bayes (1702–1761) (Huberty, 1994, S. 48) (vgl. Satz 1 auf Seite 11)

$$P(Y = y_k | \boldsymbol{X} = \boldsymbol{x}) = \frac{p^{(k)} P(\boldsymbol{X} = \boldsymbol{x} | Y = y_k)}{\sum_{c=1}^{M} p^{(c)} P(\boldsymbol{X} = \boldsymbol{x} | Y = y_c)} \tag{3.7}$$

herleiten lässt: Anstelle der Dichtefunktionen werden die a-posteriori Wahrscheinlichkeiten der Klassenzugehörigkeiten betrachtet (Huberty, 1994, S. 48). Eine Beobachtung $\boldsymbol{x}$ wird der Klasse $k \in \{1, \ldots, M\}$ zugeordnet, für die für alle $j \in \{1, \ldots, M\},\ j \neq k$, gilt:

$$\begin{aligned} & P(Y = y_k | \boldsymbol{X} = \boldsymbol{x}) > P(Y = y_j | \boldsymbol{X} = \boldsymbol{x}) \\ \Leftrightarrow\ & \frac{p^{(k)} P(\boldsymbol{X} = \boldsymbol{x} | Y = y_k)}{\sum_{c=1}^{M} p^{(c)} P(\boldsymbol{X} = \boldsymbol{x} | Y = y_c)} > \frac{p^{(j)} P(\boldsymbol{X} = \boldsymbol{x} | Y = y_j)}{\sum_{c=1}^{M} p^{(c)} P(\boldsymbol{X} = \boldsymbol{x} | Y = y_c)}. \end{aligned} \tag{3.8}$$

Da die Summe im Nenner identisch für alle Klassen c ist und $P(\boldsymbol{X} = \boldsymbol{x} | Y = y_c) \propto f^{(c)}(\boldsymbol{x})$ ist, ergibt sich eine äquivalente Regel durch die von Rao (1973, S. 574) bezeichneten *Diskriminanzscores* (*discriminant scores*) $p^{(c)} f^{(c)}(\boldsymbol{x})$ (Huberty, 1994, S. 48):

$$\begin{aligned} & \frac{p^{(k)} f_{\boldsymbol{X}|Y=y_k}(\boldsymbol{x}|y_k)}{\sum_{c=1}^{M} p^{(c)} f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c)} > \frac{p^{(j)} f_{\boldsymbol{X}|Y=y_j}(\boldsymbol{x}|y_j)}{\sum_{c=1}^{M} p^{(c)} f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c)} \\ \Leftrightarrow\ & p^{(k)} f^{(k)}(\boldsymbol{x}) > p^{(j)} f^{(j)}(\boldsymbol{x}). \end{aligned} \tag{3.9}$$

Dies ist äquivalent zu

$$k = \underset{c=1,\ldots,M}{\arg\max}\, p^{(c)} f^{(c)}(\boldsymbol{x}) =: G_{\text{Bayes}}(\boldsymbol{x}). \tag{3.10}$$

Hier ist zu sehen, dass diese Regel eine Verallgemeinerung der Maximum-Likelihood Regel (3.6) ist, bei welcher identische a-priori Wahrscheinlichkeiten $p^{(c)} = \frac{1}{M}$ für alle M Klassen unterstellt werden.

Optimalitätseigenschaften Die Bayes-Regel $G_{\text{Bayes}}(\boldsymbol{x})$, welche einer Beobachtung $\boldsymbol{x}$ nach (3.10) die Klasse k zuweist, weist im Vergleich zu allen Klassifikationsregeln die geringste Fehlerrate, bedingt darauf, dass $\boldsymbol{X} = \boldsymbol{x}$ bekannt ist (bedingte Fehlerrate), sowie auch die geringste Gesamtfehlerrate auf (Fahrmeir et al., 1996a, S. 361 f.).

Sowohl die Bayes- als auch die Maximum-Likelihood Regel besitzen zudem Optimalitätseigenschaften bezüglich auftretender Kosten durch eine Klassifikation. Beide sind kostenoptimale Regeln, d. h. sie minimieren die erwarteten Fehlklassifikationskosten jeweils unter der Betrachtung einer bestimmten Kostenfunktion C.

Solch eine Kostenfunktion

$$C : Y \times Y \to \mathbb{R}$$

quantifiziert dabei die Kosten $C(y_k, y_j)$, welche durch die Klassifizierung einer Beobachtung aus Klasse k in Klasse j resultieren.

In vielen Fällen werden identische positive Kosten für alle Klassen betrachtet. Alle Fehlklassifikationen werden demnach gleich behandelt. Diese Verlust- bzw. Kostenfunktion wird als *einfache symmetrische Kostenfunktion* bezeichnet und ist folgendermaßen definiert (Fahrmeir et al., 1996a, S. 362 f.):

$$C_e(y_k, y_j) = \begin{cases} 0, & k = j, \\ K > 0, & k \neq j. \end{cases} \tag{3.11}$$

Als Spezialfall der *einfachen symmetrischen Kostenfunktion* ergibt sich mit $K = 1$ der *0-1-Verlust*:

$$C_0(y_k, y_j) = \begin{cases} 0, & k = j, \\ 1, & k \neq j, \end{cases} \quad \text{bzw.} \quad C_0(y_k, y_j) = \mathbb{1}_{\{k \neq j\}}. \tag{3.12}$$

Die gesamten Fehlklassifikationskosten entsprechen in diesem Fall der Summe der fehlklassifizierten Beobachtungen.

Eine weitere spezielle Kostenfunktion ist die sogenannte *umgekehrt proportionale Kostenfunktion* (Fahrmeir et al., 1996a, S. 363), welche eine proportionale Gewichtung der Kosten einer Fehlklassifikation mit der Größe der wahren Klasse betrachtet:

$$C_p(y_k, y_j) = \begin{cases} 0, & k = j, \\ \frac{K}{p^{(k)}}, & k \neq j, \end{cases} \quad \text{mit } K > 0. \tag{3.13}$$

Wird eine Beobachtung aus Klasse k fälschlicherweise Klasse j zugeordnet, so sind die Kosten umso größer, je kleiner die Klasse k ist.

Sei G eine Klassifikationsregel, welche eine Beobachtung $\boldsymbol{x} \in \boldsymbol{X}$ durch $G(\boldsymbol{x})$ einer Klasse zuordnet. Eine *kostenoptimale Regel* minimiert die *erwarteten Kosten* (Fahrmeir et al., 1996a, S. 362):

$$\begin{aligned} &\arg\min_G \int \underbrace{\sum_{y_c \in Y} P(Y = y_c | \boldsymbol{X} = \boldsymbol{x})\, C(y_c, G(\boldsymbol{x}))}_{\text{bedingte erwartete Kosten}}\, \mathrm{d}\boldsymbol{x} \\ &\overset{(3.7)}{=} \arg\min_G \int \sum_{y_c \in Y} p^{(c)} P(\boldsymbol{X} = \boldsymbol{x} | Y = y_c)\, C(y_c, G(\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}. \end{aligned}$$

Für die einfache symmetrische Kostenfunktion (3.11) lässt sich die Minimierung der erwarteten Kosten vereinfachen zu:

$$\arg\min_G \int \sum_{y_c \neq G(\boldsymbol{x})} p^{(c)} P(\boldsymbol{X} = \boldsymbol{x} | Y = y_c)\, K\, \mathrm{d}\boldsymbol{x} \;\Leftrightarrow\; \arg\min_G \int K \sum_{y_c \neq G(\boldsymbol{x})} p^{(c)} f^{(c)}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}. \tag{3.14}$$

Für jedes $\boldsymbol{x} \in \boldsymbol{X}$ ist $K \sum_{y_c \neq G(\boldsymbol{x})} p^{(c)} f^{(c)}(\boldsymbol{x})$ minimal für $G(\boldsymbol{x}) = \arg\max_{c=1,\ldots,M} p^{(c)} f^{(c)}(\boldsymbol{x})$. Damit ergibt sich die Bayes-Regel (3.10) bei Betrachtung der einfachen symmetrischen Kostenfunktion und auch im Speziellen des 0-1-Verlustes durch $K = 1$, sodass in diesen Fällen die Bayes-Regel kostenoptimal ist.

Wird hingegen die umgekehrt proportionale Kostenfunktion (3.13) betrachtet, so ergibt sich aus der kostenoptimalen Regel die Maximum-Likelihood Regel (3.6), da

$$\begin{aligned} &\arg\min_G \int \sum_{y_c \neq G(\boldsymbol{x})} p^{(c)} P(\boldsymbol{X} = \boldsymbol{x} | Y = y_c) \cdot \frac{K}{p^{(c)}}\, \mathrm{d}\boldsymbol{x} \Leftrightarrow \arg\min_G \int K \sum_{y_c \neq G(\boldsymbol{x})} f^{(c)}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \\ &= \arg\max_{c=1,\ldots,M} f^{(c)}(\boldsymbol{x}) = G_{\mathrm{ML}}(\boldsymbol{x}). \end{aligned}$$

Diese ist somit kostenoptimal bezüglich der umgekehrt proportionalen Kostenfunktion.

3.1.1 Spezialfall: Bayesfehler bei 0-1-Verlust

Der mit der Bayes-Regel (3.10) zusammenhängende Fehler wird als *Bayesfehler* bezeichnet. Bei Betrachtung des 0-1-Verlustes (3.12) lässt sich der Bayesfehler im allgemeinen Fall von M Klassen definieren durch (Toussaint, 1974, S. 472):

$$err_{\text{Bayes}} = 1 - \int \max_{c=1,\ldots,M} \left(p^{(c)} f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c) \right) \mathrm{d}\boldsymbol{x} = 1 - \int \max_{c=1,\ldots,M} \left(p^{(c)} f^{(c)}(\boldsymbol{x}) \right) \mathrm{d}\boldsymbol{x}. \tag{3.15}$$

Differenzierter gilt analog ohne Betrachtung der Maximumsfunktion (Garber und Djouadi, 1988, S. 281):

$$err_{\text{Bayes}} = 1 - \sum_{c=1}^{M} \int_{R_c} p^{(c)} f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c) \,\mathrm{d}\boldsymbol{x} = 1 - \sum_{c=1}^{M} \int_{R_c} p^{(c)} f^{(c)}(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}. \tag{3.16}$$

Die Maximumsfunktion wird dabei in die Integrationsgrenzen übertragen. R_c beschreibt die Region der Einflussvariablen $\boldsymbol{x} \in \mathbb{R}^p$, in welcher die a-posteriori Wahrscheinlichkeit für Klasse c maximal ist (Garber und Djouadi, 1988, S. 282; Tumer und Ghosh, 1996, S. 696) und der Bayes-Klassifikator (3.10) demnach Klasse c prognostiziert. Formal lässt sich diese Region folgendermaßen darstellen (Garber und Djouadi, 1988, S. 282):

$$R_c = \left\{ \boldsymbol{x} \in \mathbb{R}^p \,\middle|\, P(Y = y_c|\boldsymbol{X} = \boldsymbol{x}) > \max_{\substack{j=1,\ldots,M,\\ j\neq c}} P(Y = y_j|\boldsymbol{X} = \boldsymbol{x}) \right\}.$$

Diese Menge lässt sich analog umformulieren zu

$$R_c = \left\{ \boldsymbol{x} \in \mathbb{R}^p \,\middle|\, \arg\max_{j=1,\ldots,M} P(Y = y_j|\boldsymbol{X} = \boldsymbol{x}) = c \right\}.$$

Wegen des Satzes von Bayes (vgl. (3.7) auf Seite 33) ist die a-posteriori Wahrscheinlichkeit für Klasse c maximal, wenn (der Zähler) $p^{(c)} f^{(c)}(\boldsymbol{x})$ maximal ist, da die Dichte im Nenner für alle Klassen $c = 1, \ldots, M$ identisch ist. Daher kann bei der Menge R_c auch das Produkt $p^{(c)} f^{(c)}(\boldsymbol{x})$ anstelle der a-posteriori Wahrscheinlichkeit betrachtet werden.

Eine alternative Herleitung des Bayesfehlers im Falle von zwei Klassen liefert Fukunaga (1990, S. 52 f.). Diese Herleitung basiert auf der Betrachtung des bedingten Fehlers bzw. der erwarteten Kosten (vgl. Seite 34 f.), falls die Verteilung der Einflussvariablen bekannt ist und die Bayes-Regel betrachtet wird.

Im Falle von zwei Klassen ist dieser bedingte Fehler bzw. sind die minimalen Kosten bei Betrachtung des 0-1-Verlustes in jedem Punkt $\boldsymbol{x}$

$$r(\boldsymbol{X}) = \min \left(P(Y = y_1 \,|\, \boldsymbol{X} = \boldsymbol{x}),\ P(Y = y_2 \,|\, \boldsymbol{X} = \boldsymbol{x}) \right).$$

Der Bayesfehler ist der erwartete Fehler, also (mit $f_{\boldsymbol{X}}(\boldsymbol{x}) = \sum_{c=1}^{M} p^{(c)} f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c)$)

$$\begin{aligned} err_{\text{Bayes}} &= \mathrm{E}\left(r(\boldsymbol{X})\right) = \int \min\left(P(Y = y_1 \mid \boldsymbol{X} = \boldsymbol{x}),\ P(Y = y_2 \mid \boldsymbol{X} = \boldsymbol{x})\right) f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\ &= \int \min\left(P(Y = y_1 \mid \boldsymbol{X} = \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}),\ P(Y = y_2 \mid \boldsymbol{X} = \boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x})\right)\,\mathrm{d}\boldsymbol{x} \\ &\overset{(3.8)/(3.9)}{=} \int \min\left(p^{(1)} f_{\boldsymbol{X}|Y=y_1}(\boldsymbol{x}|y_1),\ p^{(2)} f_{\boldsymbol{X}|Y=y_2}(\boldsymbol{x}|y_2)\right)\,\mathrm{d}\boldsymbol{x} \\ &= \int \min\left(p^{(1)} f^{(1)}(\boldsymbol{x}),\ p^{(2)} f^{(2)}(\boldsymbol{x})\right)\,\mathrm{d}\boldsymbol{x} \end{aligned} \tag{3.17}$$

(vgl. dazu (3.14)).

Mithilfe der Regionen R_c kann diese Darstellung noch weiter aufgespalten werden. Dadurch lässt sich (3.17) im Zwei-Klassen-Fall auch folgendermaßen definieren:

$$err_{\text{Bayes}} = p^{(1)} \int_{R_2} f^{(1)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + p^{(2)} \int_{R_1} f^{(2)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}. \tag{3.18}$$

Mit wachsender Anzahl an Klassen M wird die Herleitung dieser zweiten Variante des Bayesfehlers jedoch immer komplexer, da $\sum_{i=2}^{M-1} \binom{M}{i}$ Schnitte von 2 bis $M-1$ normierten Dichten der M Klassen betrachtet werden müssen. Bei der Minimierung der erwarteten Kosten durch den Bayes-Klassifikator für $M > 2$ Klassen wird daher besser

$$err_{\text{Bayes}} = \min \int \sum_{y_c \neq G_{\text{Bayes}}(\boldsymbol{x})} p^{(c)} f^{(c)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

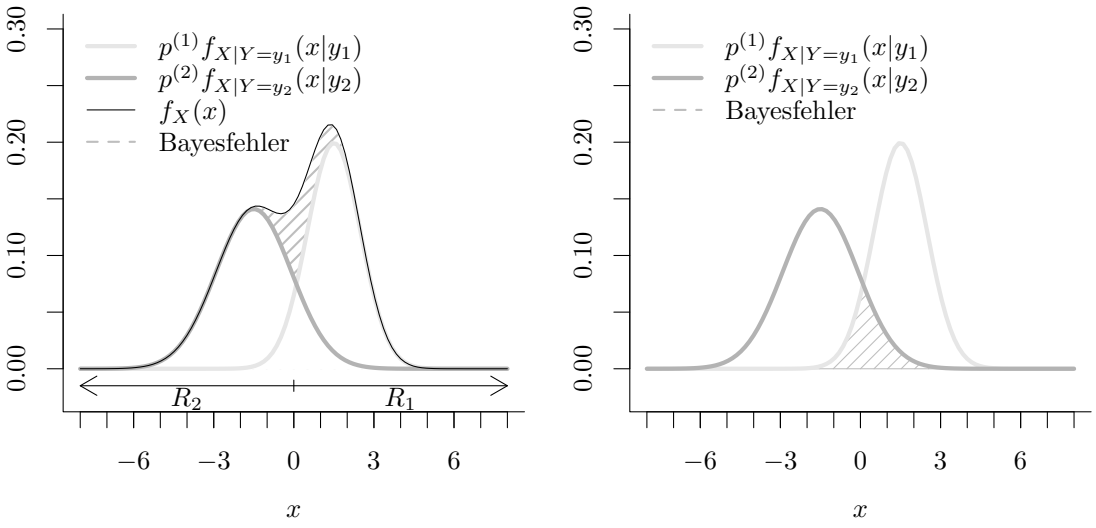
betrachtet (vgl. (3.14)).

Der Bayesfehler und die unterschiedliche Idee der zwei verschiedenen Herleitungen ist in Abbildung 3.1 für ein eindimensionales Beispiel und zwei Klassen veranschaulicht. In diesem Beispiel werden zwei Normalverteilungen $X|(Y = y_1) \sim \mathcal{N}(1.5, 1)$ und $X|(Y = y_2) \sim \mathcal{N}(-1.5, 2)$ betrachtet. Die a-priori Wahrscheinlichkeiten sind für beide Klassen identisch $p^{(1)} = p^{(2)} = 0.5$.

In der Abbildung 3.1 (a) ist die erste Variante des Bayesfehlers nach (3.15) bzw. (3.16) veranschaulicht. Dieser ergibt sich aus der Differenz zwischen der gesamten Wahrscheinlichkeit 1 (Integral von $f_{\boldsymbol{X}}(\boldsymbol{x})$ über den gesamten Wertebereich) und dem Maximum der normierten Dichtefunktionen beider Klassen.

Im Zwei-Klassen-Fall lässt sich dieser Bayesfehler auch ohne die gemeinsame Dichte $f_{\boldsymbol{X}}(\boldsymbol{x})$ vereinfacht veranschaulichen (s. Abbildung 3.1 (b)) durch das Integral über den gesamten Wertebereich der Minimumsfunktion beider normierten Dichten nach (3.17) bzw. (3.18). Der Flächeninhalt beider Varianten ist identisch.

Der Bayesfehler kann als untere Grenze beim Vergleich der Fehlklassifikationsraten verschiedener Klassifikationsmethoden betrachtet werden, da er aufgrund der Optimalitäts-



(a) Bestimmung nach (3.15)/(3.16)

(b) Bestimmung nach (3.17)/(3.18)

Abbildung 3.1: Veranschaulichung des Bayesfehlers für zwei Klassen mit $X|(Y = y_1) \sim \mathcal{N}(1.5, 1)$ und $X|(Y = y_2) \sim \mathcal{N}(-1.5, 2)$ und identischen a-priori Wahrscheinlichkeiten $p^{(1)} = p^{(2)} = 0.5$.

eigenschaft der Bayes-Regel der minimale Fehler ist (vgl. Seite 34). Wird die Fehlerrate eines Klassifikationsverfahrens kleiner als der Bayesfehler geschätzt, so ist von Overfitting auf den vorliegenden Daten auszugehen. Hier ist jedoch zu beachten, dass der Bayesfehler im Allgemeinen nicht exakt bestimmt werden kann, da die wahre Verteilung der Daten bekannt sein muss, und zudem – selbst wenn die wahre Verteilung bekannt ist – die analytische Bestimmung in den meisten Fällen sehr aufwändig ist (Garber und Djouadi, 1988, S. 282). In Simulationsstudien ist der Vergleich mit dem Bayesfehler jedoch sehr hilfreich, um die Güte von betrachteten Klassifikationsverfahren zu beurteilen.

3.2 Kanonische Diskriminanzanalyse

Die *Kanonische Diskriminanzanalyse* leitet sich aus der allgemeinen Bayes-Regel (3.10) ab, indem für die Verteilungen innerhalb der Klassen multivariate Normalverteilungen unterstellt werden (Huberty, 1994, S. 53 ff.):

$$\boldsymbol{X}^{(c)} \sim \mathcal{N}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}).$$

3.2.1 Kanonische Lineare Diskriminanzanalyse

Im Falle der *Kanonischen Linearen Diskriminanzanalyse* wird zudem angenommen, dass alle Klassen dieselbe Kovarianzmatrix aufweisen (Hastie et al., 2009, S. 108): $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}^{(1)} = \ldots = \boldsymbol{\Sigma}^{(M)}$:

Voraussetzung 1 (Lineare Diskriminanzanalyse)**.** Für die Parameter in den Klassen $c = 1, \ldots, M$ gilt:

(LDA1) Unterschiedliche Erwartungswertvektoren: $\boldsymbol{\mu}^{(c)}$ können sich unterscheiden.
(LDA2) Identische Kovarianzmatrizen: $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}^{(c)}$.
(LDA3) Multivariate Normalverteilungen: $\boldsymbol{X}^{(c)} \sim \mathcal{N}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma})$.
(LDA4) Identische Fehlklassifikationskosten (3.11).

Mit diesen Annahmen besitzt die Dichtefunktion für Klasse c demnach die folgende Form (Hastie et al., 2009, S. 108) (vgl. (3.4)):

$$f^{(c)}(\boldsymbol{x}; \boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p}} \cdot |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(c)}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(c)}\right)\right). \tag{3.19}$$

Bei zwei Klassen $k \neq j$ besagt die Bayes-Regel (3.8)/(3.9) nun, dass eine neue Beobachtung $\boldsymbol{x}$ Klasse k zugeordnet bekommt, falls die a-posteriori Wahrscheinlichkeit für Klasse k größer ist:

$$P(k|\boldsymbol{x}) > P(j|\boldsymbol{x}) \Leftrightarrow \frac{p^{(k)} f^{(k)}(\boldsymbol{x})}{\sum_{c\in\{j,k\}} p^{(c)} f^{(c)}(\boldsymbol{x})} > \frac{p^{(j)} f^{(j)}(\boldsymbol{x})}{\sum_{c\in\{j,k\}} p^{(c)} f^{(c)}(\boldsymbol{x})}.$$

Da die Summe über die beiden Klassen im Nenner bzw. die gemeinsame Dichtefunktion $f_{\boldsymbol{X}}(\boldsymbol{x})$ auf beiden Seiten identisch ist, ebenso wie der Vorfaktor in der Dichtefunktion (3.19) aufgrund der Annahme (LDA2), vereinfacht sich die Ungleichung zu

$$\begin{aligned} & p^{(k)} \exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(k)}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(k)}\right)\right) \\ & > p^{(j)} \exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(j)}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(j)}\right)\right). \end{aligned} \tag{3.20}$$

Logarithmierung und Ausmultiplizieren führt zu folgender Darstellung der Regel:

$$\begin{aligned} h_L^{(k)}(\boldsymbol{x}) &:= \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^{(k)} - \frac{1}{2}\left(\boldsymbol{\mu}^{(k)}\right)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^{(k)} + \log p^{(k)} \\ &> \\ h_L^{(j)}(\boldsymbol{x}) &:= \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^{(j)} - \frac{1}{2}\left(\boldsymbol{\mu}^{(j)}\right)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^{(j)} + \log p^{(j)}. \end{aligned}$$

Aufgrund der Annahme (LDA2) wird der quadratische Term $\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}$ bei der Umformung eliminiert. Die *Diskriminanzfunktionen* $h_L^{(c)}(\boldsymbol{x})$ sind demnach linear in $\boldsymbol{x}$, wodurch sich eine lineare Trennung durch $h_L^{(k)}(\boldsymbol{x}) = h_L^{(j)}(\boldsymbol{x})$ (Hyperebene für $p \geq 2$) ergibt (Hastie et al., 2009, S. 108). Dies erklärt die Bezeichnung *Lineare Diskriminanzanalyse*.

Für $M = 2$ Klassen und $p = 2$ Merkmale sieht die lineare Trenngerade folgendermaßen aus (van Meegen, 2015, S. 17; van Meegen et al., 2019, S. 17):

$$
\begin{aligned}
& h_L^{(1)}(\boldsymbol{x}) = h_L^{(2)}(\boldsymbol{x}) \\
\Leftrightarrow\ & \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}\right) = \frac{1}{2}\left(\boldsymbol{\mu}^{(1)}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(1)} - \frac{1}{2}\left(\boldsymbol{\mu}^{(2)}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(2)} \\
& \qquad - \log p^{(1)} + \log p^{(2)} \\
\Leftrightarrow\ & \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}\right) = \frac{1}{2}\Bigg(\left(\boldsymbol{\mu}^{(1)}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(1)} - \left(\boldsymbol{\mu}^{(2)}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(2)} \\
& \qquad + \left(\boldsymbol{\mu}^{(1)}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(2)} - \left(\boldsymbol{\mu}^{(2)}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(1)}\Bigg) \\
& \qquad - \log p^{(1)} + \log p^{(2)} \\
\Leftrightarrow\ & \boldsymbol{x}^T\underbrace{\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}\right)}_{=:\boldsymbol{a}=(a_1,a_2)^T} = -\frac{1}{2}\Bigg(\left(\boldsymbol{\mu}^{(1)}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}\right) \\
& \qquad - \left(\boldsymbol{\mu}^{(2)}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}\right)\Bigg) + \log p^{(1)} - \log p^{(2)} \\
\Leftrightarrow\ & \boldsymbol{x}^T\boldsymbol{a} = -\frac{1}{2}\left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}\right) + \log p^{(1)} - \log p^{(2)} \\
\Leftrightarrow\ & x_1a_1 + x_2a_2 = \frac{1}{2}\cdot\boldsymbol{a}^T\left(\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}\right) + \log p^{(1)} - \log p^{(2)} \\
\Leftrightarrow\ & x_2 = -\frac{a_1}{a_2}\cdot x_1 + \frac{\boldsymbol{a}^T}{a_2}\cdot\frac{\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}}{2} + \frac{\log p^{(1)} - \log p^{(2)}}{a_2}.
\end{aligned}
\tag{3.21}
$$

Im Falle identischer a-priori Wahrscheinlichkeiten $p^{(1)} = p^{(2)} = 0.5$ fällt der hintere Teil des Achsenabschnittes weg. Die Steigung ist unabhängig von den a-priori Wahrscheinlichkeiten. Eine Veränderung jener zieht also lediglich eine Verschiebung der Trenngerade entlang der Achsen nach sich.

Im verallgemeinerten Fall von $M \geq 2$ Klassen werden implizit die $\binom{M}{2}$ Ungleichungen aller Kombinationen aus zwei Klassen betrachtet. Folglich wird dann für eine neue Beobachtung $\boldsymbol{x}^*$ die Klasse k prognostiziert, deren *lineare Diskriminanzfunktion* (Hastie et al., 2009, S. 109; Johnson und Wichern, 2007, S. 611)

$$
h_L^{(c)}(\boldsymbol{x}) := \boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(c)} - \frac{1}{2}\left(\boldsymbol{\mu}^{(c)}\right)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^{(c)} + \log p^{(c)} \tag{3.22}
$$

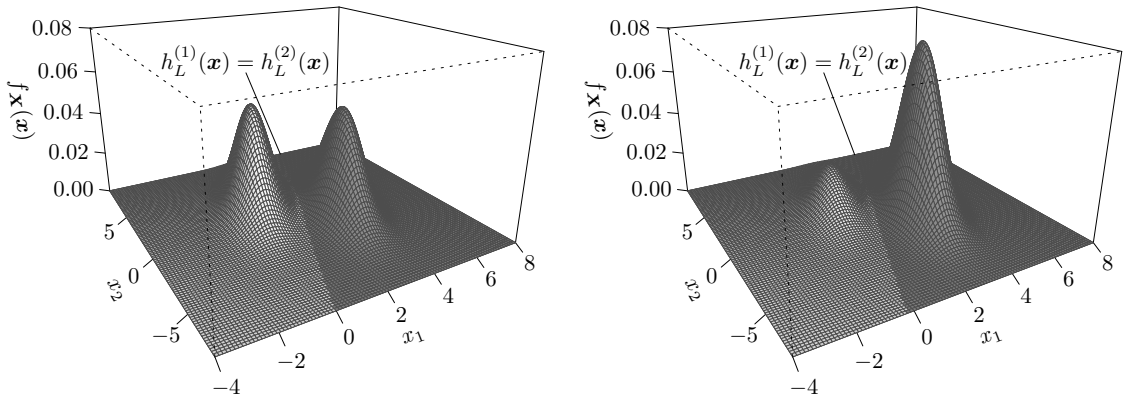
maximal ist:

$$
k = \underset{c=1,\ldots,M}{\arg\max}\ h_L^{(c)}(\boldsymbol{x}^*).
$$

Die Klassifikationsgrenzen setzen sich aus einer Menge von $\binom{M}{2}$ Hyperebenen zusammen.

Alternativ kann auch ein Minimierungsproblem betrachtet werden. Ausgehend von (3.20) ergibt sich durch Logarithmierung und Multiplikation mit dem Faktor -2 zunächst

$$
\left(\boldsymbol{x} - \boldsymbol{\mu}^{(k)}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(k)}\right) - 2\log p^{(k)} < \left(\boldsymbol{x} - \boldsymbol{\mu}^{(j)}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(j)}\right) - 2\log p^{(j)}.
$$



(a) Identische a-priori Wahrscheinlichkeiten: $p^{(1)} = p^{(2)} = 0.5$.

(b) Unterschiedliche a-priori Wahrscheinlichkeiten: $p^{(1)} = 0.2$ und $p^{(2)} = 0.8$.

Abbildung 3.2: Beispiel für Klassifikationsgrenze bei der Linearen Diskriminanzanalyse bei zwei Klassen und zweidimensionalen Verteilungen.

Bei identischen a-priori Wahrscheinlichkeiten hängt die Zuordnung einer Beobachtung somit lediglich von der quadrierten Mahalanobis-Distanz (Mahalanobis, 1936) ab (van Meegen, 2015, S. 7). Laut Mitchell und Krzanowski (1985) ist die Mahalanobis-Distanz ein geeignetes Abstandsmaß im p-dimensionalen Raum für zwei Verteilungen mit verschiedenen Erwartungswertvektoren und gemeinsamer elliptischer Kovarianzmatrix. Eine Beobachtung $\boldsymbol{x}$ wird somit jener Klasse zugeordnet, derer sie am nächsten liegt. Bei unterschiedlichen a-priori Wahrscheinlichkeiten werden Klassen mit einer geringen a-priori Wahrscheinlichkeit durch den negativen Summanden stärker bestraft, da $p^{(c)} \in (0, 1) \Rightarrow \lim_{p^{(c)} \to 0} 2 \log p^{(c)} = -\infty$ und $\lim_{p^{(c)} \to 1} 2 \log p^{(c)} = 0$. Die Klassifikationsgrenzen verschieben sich dadurch ausgehend von der quadrierten Mahalanobis-Distanz zugunsten der Klasse mit verhältnismäßig größerer a-priori Wahrscheinlichkeit.

Auch hier kann beim Ausmultiplizieren der quadratische Term vernachlässigt werden, da dieser für alle Klassen identisch ist, sodass die folgende umformulierte Diskriminanzfunktion

$$h_{L,2}^{(c)}(\boldsymbol{x}) := -2\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^{(c)} + \left(\boldsymbol{\mu}^{(c)}\right)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}^{(c)} - 2 \log p^{(c)} \tag{3.23}$$

die Regel (Huberty, 1994, S. 60 f.)

$$k = \underset{c=1,\ldots,M}{\arg\min}\, h_{L,2}^{(c)}(\boldsymbol{x}^*) \tag{3.24}$$

für eine Beobachtung $\boldsymbol{x}^*$ liefert.

Die Klassifikationsregel der Linearen Diskriminanzanalyse ist in Abbildung (3.2) anhand von zweidimensionalen Normalverteilungen veranschaulicht. Im Falle von zwei Klassen ($M = 2$) und $p = 2$ Merkmalen ergibt sich die Trennung durch die Gerade $h_L^{(1)}(\boldsymbol{x}) = h_L^{(2)}(\boldsymbol{x})$

bzw. $h_{L,2}^{(1)}(\boldsymbol{x}) = h_{L,2}^{(2)}(\boldsymbol{x})$. Es wird deutlich, dass sich durch den Übergang von identischen a-priori Wahrscheinlichkeiten (a) zu unterschiedlichen a-priori Wahrscheinlichkeiten (b) die Klassifikationsgrenze zugunsten der Klasse mit größerem $p^{(c)}$ leicht verschiebt.

Schätzung der Parameter Im allgemeinen Fall sind die zugrunde liegenden Verteilungen nicht bekannt. Die theoretischen Erwartungswertvektoren $\boldsymbol{\mu}^{(c)}$, die Kovarianzmatrix $\boldsymbol{\Sigma}$ sowie die a-priori Wahrscheinlichkeiten $p^{(c)}$, $c = 1, \ldots, M$, müssen demnach geschätzt werden, um die Klassifikationsregel aufstellen und anwenden zu können.

Für die Schätzung des Erwartungswertvektors in Klasse c wird im Allgemeinen das arithmetische Mittel (Hastie et al., 2009, S. 109; Huberty, 1994, S. 54)

$$\hat{\boldsymbol{\mu}}^{(c)} = \frac{1}{n^{(c)}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq n}} \boldsymbol{x}_i = \frac{1}{\sum_{i=1}^{n} \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}} \sum_{i=1}^{n} \boldsymbol{x}_i \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} = \frac{1}{n^{(c)}} \sum_{i=1}^{n^{(c)}} \boldsymbol{x}_{(i)}, \quad c = 1, \ldots, M, \tag{3.25}$$

als Maximum-Likelihood-Schätzer (Duda et al., 2001, S. 89) herangezogen.

Hastie et al. (2009, S. 109) schlagen die gepoolte Kovarianzmatrix als Schätzer für $\boldsymbol{\Sigma}$ vor. Diese setzt sich aus den gewichteten Kovarianzmatrizen der M Klassen,

$$\hat{\boldsymbol{\Sigma}}^{(c)} = \frac{1}{n^{(c)} - 1} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq n}} \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}^{(c)}\right) \left(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}^{(c)}\right)^T, \tag{3.26}$$

zusammen, welche unverzerrte Schätzer für $\boldsymbol{\Sigma}^{(c)}$ sind (Duda et al., 2001, S. 90):

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n - M} \sum_{c=1}^{M} (n^{(c)} - 1) \hat{\boldsymbol{\Sigma}}^{(c)}. \tag{3.27}$$

Filzmoser et al. (2006, S. 522) schlagen eine alternative Variante der Bestimmung der gepoolten Kovarianzmatrix $\boldsymbol{\Sigma}_{\text{gew}}$ vor, welche eine Gewichtung durch die a-priori Wahrscheinlichkeiten beinhaltet. Auf Basis dieser lässt sich eine empirische Variante als Schätzer ableiten, indem die theoretischen Größen durch die Schätzer $\hat{p}^{(c)}$ und $\hat{\boldsymbol{\Sigma}}^{(c)}$ ersetzt werden:

$$\hat{\boldsymbol{\Sigma}}_{\text{gew}} = \sum_{c=1}^{M} \hat{p}^{(c)} \hat{\boldsymbol{\Sigma}}^{(c)}. \tag{3.28}$$

Auch wenn theoretisch $\boldsymbol{\Sigma}_{\text{gew}} = \boldsymbol{\Sigma}$ (unter der Annahme (LDA2)), so unterscheiden sich die Schätzer $\hat{\boldsymbol{\Sigma}}$ und $\hat{\boldsymbol{\Sigma}}_{\text{gew}}$ im Allgemeinen (van Meegen, 2015, S. 12).

Als Schätzer für die a-priori Wahrscheinlichkeiten werden meistens relative Häufigkeiten (Hastie et al., 2009, S. 109)

$$\hat{p}^{(c)} = \frac{n^{(c)}}{n}, \quad c = 1, \ldots, M, \tag{3.29}$$

oder die Betrachtung einer diskreten Gleichverteilung herangezogen (McLachlan, 1992, S. 288):

$$\hat{p}^{(c)} = \frac{1}{M}, \quad c = 1, \ldots, M. \tag{3.30}$$

Van Meegen et al. (2019, S. 5) zeigen, dass die Schätzer (3.27) und (3.28) identisch sind im Falle identischer Klassengrößen $n^{(c)} = \frac{n}{M}$, $c = 1, \ldots, M$, und identischer a-priori Wahrscheinlichkeiten $p^{(c)} = \frac{1}{M}$, $c = 1, \ldots, M$, (falls diese bekannt sind) bzw. bei Heranziehen von (3.30) als Schätzer oder (3.29) als Folge aus $n^{(c)} = \frac{n}{M}$.

3.2.2 Quadratische Diskriminanzanalyse

Bei der *Quadratischen Diskriminanzanalyse* wird die Annahme identischer Kovarianzmatrizen in den Klassen aufgehoben (Hastie et al., 2009, S. 110). Für die Verteilungen innerhalb der Klassen werden multivariate Normalverteilungen $\boldsymbol{X}^{(c)} \sim \mathcal{N}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)})$ mit Dichtefunktionen

$$f^{(c)}(\boldsymbol{x}; \boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}) = \frac{1}{\sqrt{(2\pi)^p}} \cdot \left|\boldsymbol{\Sigma}^{(c)}\right|^{-1/2} \exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \boldsymbol{\mu}^{(c)}\right)^T \left(\boldsymbol{\Sigma}^{(c)}\right)^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}^{(c)}\right)\right) \tag{3.31}$$

unterstellt:

> **Voraussetzung 2** (Quadratische Diskriminanzanalyse)**.** Für die Parameter in den Klassen $c = 1, \ldots, M$ gilt:
>
> (QDA1) Unterschiedliche Erwartungswertvektoren: $\boldsymbol{\mu}^{(c)}$ können sich unterscheiden.
> (QDA2) Unterschiedliche Kovarianzmatrizen: $\boldsymbol{\Sigma}^{(c)}$ können sich unterscheiden.
> (QDA3) Multivariate Normalverteilungen: $\boldsymbol{X}^{(c)} \sim \mathcal{N}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)})$.
> (QDA4) Identische Fehlklassifikationskosten (3.11).

Auch hier wird in Bezug auf die Bayes-Regel (3.8) jene Klasse mit größter a-posteriori Wahrscheinlichkeit prognostiziert.

Aufgrund nicht identischer Kovarianzmatrizen in den Klassen kürzt sich dabei anders als bei der Linearen Diskriminanzanalyse nicht der gesamte Vorfaktor von (3.31) bei der Aufstellung der Diskriminanzfunktionen weg, sondern nur $\frac{1}{\sqrt{(2\pi)^p}}$. Ebenso bleiben die quadratischen Terme in $\boldsymbol{x}$ erhalten, was die Bezeichnung *Quadratische Diskriminanzanalyse* erklärt. Lediglich die Nenner aus (3.8) bzw. (3.9) können auch hier außer Acht gelassen werden, da sie für alle Klassen identisch sind.

Durch Logarithmierung von (3.9) unter Beachtung von (3.31) sehen die *quadratischen Diskriminanzfunktionen* somit folgendermaßen aus (Hastie et al., 2009, S. 110; Johnson und Wichern, 2007, S. 610):

$$h_Q^{(c)}(\boldsymbol{x}) := -\frac{1}{2}\log\left|\boldsymbol{\Sigma}^{(c)}\right| - \frac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)^T\left(\boldsymbol{\Sigma}^{(c)}\right)^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right) + \log p^{(c)}. \tag{3.32}$$

Die Klassifikationsregel in der Quadratischen Diskriminanzanalyse basiert auf diesen quadratischen Diskriminanzfunktionen. Eine Beobachtung $\boldsymbol{x}^*$ wird Klasse (Huberty, 1994, S. 58)

$$k = \underset{c=1,\ldots,M}{\arg\max}\; h_Q^{(c)}(\boldsymbol{x}^*)$$

zugeordnet.

Auch hier schlägt Huberty (1994, S. 60) die äquivalente Regel auf Basis des Minimierungsproblems

$$k = \underset{c=1,\ldots,M}{\arg\min}\; h_{Q,2}^{(c)}(\boldsymbol{x}^*)$$

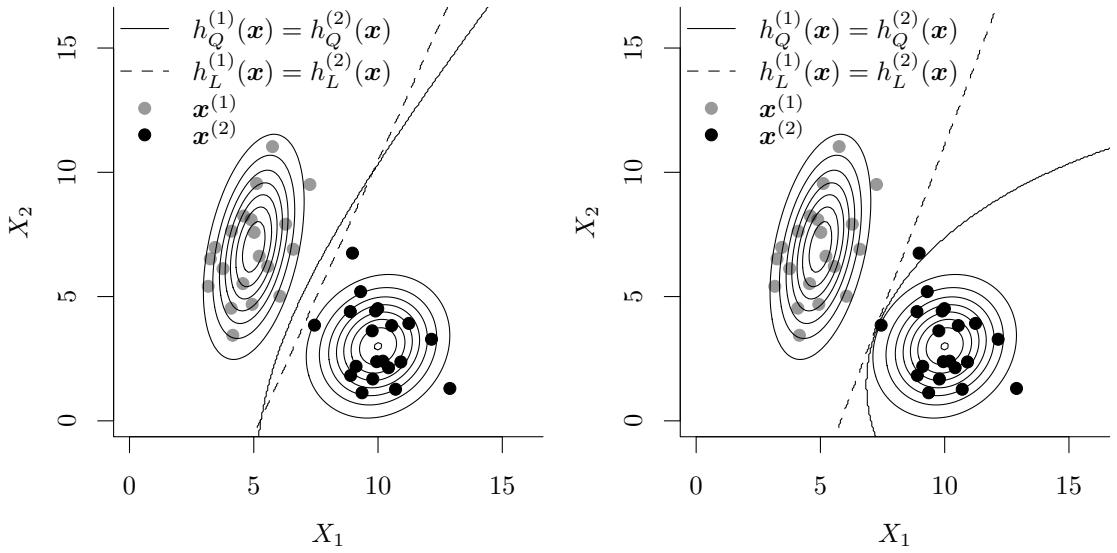
vor, wobei

$$h_{Q,2}^{(c)}(\boldsymbol{x}) := \log\left|\boldsymbol{\Sigma}^{(c)}\right| + \left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)^T\left(\boldsymbol{\Sigma}^{(c)}\right)^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right) - 2\log p^{(c)}. \tag{3.33}$$

Die Klassifikationsgrenzen ergeben sich durch die Menge der $\binom{M}{2}$ quadratischen Gleichungen für $j, k \in \{1,\ldots,M\}$, $j \neq k$, (Hastie et al., 2009, S. 110): $\left\{\boldsymbol{x} : h_Q^{(j)}(\boldsymbol{x}) = h_Q^{(k)}(\boldsymbol{x})\right\}$.

Durch die Betrachtung differenzierter Kovarianzmatrizen $\boldsymbol{\Sigma}^{(c)}$ für die einzelnen Klassen sind flexiblere (quadratische) Klassifikationsgrenzen möglich als bei der Linearen Diskriminanzanalyse. Dies wird in Abbildung 3.3 deutlich. Als Beispiel ist die Klassifikationsgrenze für zweidimensionale Normalverteilungen $\boldsymbol{X}^{(1)} \sim \mathcal{N}\left(\begin{pmatrix}5\\7\end{pmatrix}, \begin{pmatrix}1 & 0.9\\0.9 & 5\end{pmatrix}\right)$ und $\boldsymbol{X}^{(2)} \sim \mathcal{N}\left(\begin{pmatrix}10\\3\end{pmatrix}, \begin{pmatrix}2 & 0.3\\0.3 & 2\end{pmatrix}\right)$ für (a) identische a-priori Wahrscheinlichkeiten $p^{(1)} = p^{(2)} = 0.5$ und (b) unterschiedliche a-priori Wahrscheinlichkeiten $p^{(1)} = 0.7$ und $p^{(2)} = 0.3$ in den Klassen veranschaulicht. Zum Vergleich ist ebenfalls die Trennung durch die LDA eingezeichnet, für welche die gemeinsame gepoolte Kovarianzmatrix $\boldsymbol{\Sigma} = \begin{pmatrix}1.5 & 0.6\\0.6 & 3.5\end{pmatrix}$ in (a) bzw. $\boldsymbol{\Sigma} = \begin{pmatrix}1.30 & 0.72\\0.72 & 4.10\end{pmatrix}$ in (b) nach (3.28) betrachtet wird. Während sich die Klassifikationsgrenze der LDA beim Übergang von identischen zu unterschiedlichen a-priori Wahrscheinlichkeiten nur entlang der Achsen verschiebt (vgl. Seite 41 f. und Abbildung 3.2), ändert die Klassifikationsgrenze der QDA auch ihre Form und nicht nur ihre Lage im Raum.

Bei der Anwendung der QDA können die im vorherigen Abschnitt eingeführten Schätzer als empirische Varianten von $\boldsymbol{\mu}^{(c)}$, $\boldsymbol{\Sigma}^{(c)}$ und $p^{(c)}$ herangezogen werden.



(a) Identische a-priori Wahrscheinlichkeiten: $p^{(1)} = p^{(2)} = 0.5$.

(b) Unterschiedliche a-priori Wahrscheinlichkeiten: $p^{(1)} = 0.7$ und $p^{(2)} = 0.3$.

Abbildung 3.3: Beispiel für Klassifikationsgrenze bei der QDA bei zwei Klassen und zweidimensionalen Verteilungen. Zum Vergleich ist die Klassifikationsgrenze der LDA bei Betrachtung der gepoolten Kovarianzmatrix (3.28) mit eingezeichnet.

3.3 Fisher Diskriminanzanalyse

Fisher (1936) entwickelt eine verteilungsfreie Form der (linearen) Diskriminanzanalyse (kurz: *Fisher LDA*), welche ohne die Annahme einer Normalverteilung auskommt. Aufgrund dieser fehlenden Verteilungsannahme können keine a-posteriori Wahrscheinlichkeiten bestimmt werden, um die Klassifikationsregel aufzustellen. Stattdessen wird eine geometrische Idee zur Trennung von Datenpunkten betrachtet. Weiterhin werden jedoch wie bei der Kanonischen LDA identische Kovarianzmatrizen innerhalb der Klassen $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}^{(c)}$, $c = 1, \ldots, M$, unterstellt, die Erwartungswertvektoren dürfen sich unterscheiden. Also:

> **Voraussetzung 3** (Fisher Diskriminanzanalyse)**.** Für die Parameter in den Klassen $c = 1, \ldots, M$ gilt:
>
> (FDA1) Unterschiedliche Erwartungswertvektoren: $\boldsymbol{\mu}^{(c)}$ können sich unterscheiden.
> (FDA2) Identische Kovarianzmatrizen: $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}^{(c)}$.
> (FDA3) Identische Fehlklassifikationskosten (3.11).

Fisher (1936) fokussiert sich größtenteils auf den Zwei-Klassen-Fall. Rao (1948) und Bryan (1951) verallgemeinern den Ansatz auf mehr als zwei Klassen (Krzanowski und Marriott, 1995, S. 7). Mukhopadhyay (2009) beschreibt die Idee der Fisher Diskriminanzanalyse vor dem Hintergrund der Dimensionsreduktion für $M \geq 2$ Klassen. Mithilfe einer geringeren Anzahl an Variablen ($r < p$) soll eine bestmögliche Trennung der verschiedenen $M \geq 2$

Klassen erfolgen. Dazu werden die Zufallsvektoren $\boldsymbol{X}^{(c)}$ mithilfe von Vektoren $\boldsymbol{\alpha} \in \mathbb{R}^p$, den sogenannten *Diskriminanzkomponenten*, in einen niedriger-dimensionalen Raum transformiert und die transformierten Zufallsvektoren $\boldsymbol{\mathcal{Y}}^{(c)}$ betrachtet. Van Meegen et al. (2019, S. 3) beschreiben diesen Zusammenhang in mehreren Schritten:

$$\boldsymbol{\mathcal{Y}}^{(c)} = \begin{pmatrix} \mathcal{Y}_1^{(c)} \\ \vdots \\ \mathcal{Y}_r^{(c)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_1^T \boldsymbol{X}^{(c)} \\ \vdots \\ \boldsymbol{\alpha}_r^T \boldsymbol{X}^{(c)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\alpha}_1^T \\ \vdots \\ \boldsymbol{\alpha}_r^T \end{pmatrix} \boldsymbol{X}^{(c)} =: \boldsymbol{A}^T \boldsymbol{X}^{(c)}. \tag{3.34}$$

Betrachte zunächst eine beliebige, aber feste Diskriminanzkomponente $\boldsymbol{\alpha}$ und demnach die einzelne Transformation, welche auch als *Diskriminanzfunktion* bezeichnet wird:

$$\mathcal{Y}^{(c)} = \boldsymbol{\alpha}^T \boldsymbol{X}^{(c)} \in \mathbb{R}.$$

Aufgrund der Linearität lassen sich Erwartungswert und Kovarianzmatrix (bzw. Varianz) der Transformation aus der Verteilung des Zufallsvektors $\boldsymbol{X}^{(c)}$ herleiten:

$$\begin{aligned} \boldsymbol{\mu}_{\mathcal{Y}}^{(c)} &= \mathrm{E}\left(\mathcal{Y}^{(c)}\right) = \mathrm{E}\left(\boldsymbol{\alpha}^T \boldsymbol{X}^{(c)}\right) = \boldsymbol{\alpha}^T \mathrm{E}\left(\boldsymbol{X}^{(c)}\right) = \boldsymbol{\alpha}^T \boldsymbol{\mu}^{(c)}, \\ \mathrm{Var}\left(\mathcal{Y}^{(c)}\right) &= \mathrm{Cov}\left(\mathcal{Y}^{(c)}\right) = \mathrm{Cov}\left(\boldsymbol{\alpha}^T \boldsymbol{X}^{(c)}\right) = \boldsymbol{\alpha}^T \mathrm{Cov}\left(\boldsymbol{X}^{(c)}\right) \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}. \end{aligned}$$

Die geometrische Idee der Fisher Diskriminanzanalyse besteht darin, dass eine gute Trennung der Gruppen erfolgen kann, falls die jeweiligen Erwartungswertvektoren der einzelnen Klassen weit voneinander entfernt liegen, während die Streuung innerhalb der Klassen gering ist. Das erste Kriterium wird dabei durch die Summe der quadrierten Abstände der Erwartungswertvektoren der einzelnen Klassen und dem Mittel der Erwartungswertvektoren über alle Klassen abgebildet. Die Streuung innerhalb der Klassen wird durch die Kovarianzmatrix charakterisiert, welche für alle Klassen als identisch vorausgesetzt wird (vgl. (FDA2)).

Die Transformation durch $\boldsymbol{\alpha}$ soll demnach so erfolgen, dass das folgende Maximierungsproblem, welches beide Problemstellungen kombiniert, gelöst wird (van Meegen, 2015, S. 9; Filzmoser et al., 2006, S. 522):

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{\sum_{c=1}^{M} \left(\boldsymbol{\mu}_{\mathcal{Y}}^{(c)} - \bar{\boldsymbol{\mu}}_{\mathcal{Y}}\right)^2}{\mathrm{Cov}\left(\mathcal{Y}^{(c)}\right)} = \max_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{\sum_{c=1}^{M} \left(\boldsymbol{\mu}_{\mathcal{Y}}^{(c)} - \bar{\boldsymbol{\mu}}_{\mathcal{Y}}\right)^2}{\mathrm{Var}\left(\mathcal{Y}^{(c)}\right)} = \max_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{\sum_{c=1}^{M} \left(\boldsymbol{\mu}_{\mathcal{Y}}^{(c)} - \bar{\boldsymbol{\mu}}_{\mathcal{Y}}\right)^2}{\boldsymbol{\alpha}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}}, \tag{3.35}$$

wobei

$$\bar{\boldsymbol{\mu}}_{\mathcal{Y}} = \frac{1}{M} \sum_{c=1}^{M} \boldsymbol{\mu}_{\mathcal{Y}}^{(c)} = \frac{1}{M} \sum_{c=1}^{M} \boldsymbol{\alpha}^T \boldsymbol{\mu}^{(c)} = \boldsymbol{\alpha}^T \cdot \frac{1}{M} \sum_{c=1}^{M} \boldsymbol{\mu}^{(c)} = \boldsymbol{\alpha}^T \bar{\boldsymbol{\mu}}.$$

Eine gewichtete Variante des Erwartungswertvektors ist (Filzmoser et al., 2006, S. 521)

$$\bar{\boldsymbol{\mu}}_{\text{gew}} = \sum_{c=1}^{M} p^{(c)} \boldsymbol{\mu}^{(c)}$$

und resultierend daraus $\bar{\boldsymbol{\mu}}_{\mathcal{Y}_{\text{gew}}} = \boldsymbol{\alpha}^T \bar{\boldsymbol{\mu}}_{\text{gew}}$.

Der Zähler des Maximierungsproblems (3.35) lässt sich analog durch die transformierte gewichtete Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ darstellen, da (van Meegen, 2015, S. 9)

$$\sum_{c=1}^{M} \left(\boldsymbol{\mu}_{\mathcal{Y}}^{(c)} - \bar{\boldsymbol{\mu}}_{\mathcal{Y}}\right)^2 = \sum_{c=1}^{M} \left(\boldsymbol{\alpha}^T \boldsymbol{\mu}^{(c)} - \boldsymbol{\alpha}^T \bar{\boldsymbol{\mu}}\right)^2 = \boldsymbol{\alpha}^T \underbrace{\left(\sum_{c=1}^{M} \left(\boldsymbol{\mu}^{(c)} - \bar{\boldsymbol{\mu}}\right)\left(\boldsymbol{\mu}^{(c)} - \bar{\boldsymbol{\mu}}\right)^T\right)}_{=:\boldsymbol{B}} \boldsymbol{\alpha}. \tag{3.36}$$

Es resultiert das analoge Maximierungsproblem:

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{\boldsymbol{\alpha}^T \boldsymbol{B} \boldsymbol{\alpha}}{\boldsymbol{\alpha}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}}. \tag{3.37}$$

Es wird also eine Transformation gesucht, sodass die Kovarianzmatrix innerhalb der Gruppen minimiert und jene Zwischen-den-Gruppen maximiert wird. Die Lösung ist dabei bis auf einen Skalierungsfaktor eindeutig.

Wird eine gewichtete Variante betrachtet (Filzmoser et al., 2006, S. 522), dann gilt für den Zähler des Maximierungsproblems (van Meegen et al., 2019, S. 3):

$$\sum_{c=1}^{M} p^{(c)} \left(\boldsymbol{\mu}_{\mathcal{Y}}^{(c)} - \bar{\boldsymbol{\mu}}_{\mathcal{Y}_{\text{gew}}}\right)^2 = \sum_{c=1}^{M} p^{(c)} \left(\boldsymbol{\alpha}^T \boldsymbol{\mu}^{(c)} - \boldsymbol{\alpha}^T \bar{\boldsymbol{\mu}}_{\text{gew}}\right)^2 = \boldsymbol{\alpha}^T \boldsymbol{B}_{\text{gew}} \boldsymbol{\alpha}$$

mit

$$\boldsymbol{B}_{\text{gew}} = \sum_{c=1}^{M} p^{(c)} \left(\boldsymbol{\mu}^{(c)} - \bar{\boldsymbol{\mu}}_{\text{gew}}\right)\left(\boldsymbol{\mu}^{(c)} - \bar{\boldsymbol{\mu}}_{\text{gew}}\right)^T. \tag{3.38}$$

Im Nenner wird bei der Transformation die gewichtete Kovarianzmatrix innerhalb der Gruppen anstelle von $\boldsymbol{\Sigma}$ betrachtet (Filzmoser et al., 2006, S. 522):

$$\boldsymbol{\Sigma}_{\text{gew}} = \sum_{c=1}^{M} p^{(c)} \boldsymbol{\Sigma}^{(c)}.$$

Van Meegen (2015, S. 10) zeigt jedoch, dass $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{\text{gew}}$ im Falle der Annahme (FDA2) (vgl. auch Seite 42).

Mardia et al. (1979, S. 319) nennen als Lösung $\boldsymbol{\alpha}^*$ des Maximierungsproblems (3.35) bzw. (3.37) den entsprechenden Eigenvektor zum größten Eigenwert des Eigenwertproblems von $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$. Unter der Nebenbedingung $\boldsymbol{\alpha}_j^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_j = 1,\ j = 1, \ldots, r$, leitet Mukhopadhyay (2009, S. 429 ff.) die Lösung $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}_1$ ausführlich her.

Er betrachtet dazu die Spektralzerlegung der Kovarianzmatrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_p \end{pmatrix} \begin{pmatrix} \gamma_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \gamma_p \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_p^T \end{pmatrix},$$

wobei γ_j, $j = 1, \ldots, p$, die Eigenwerte und $\boldsymbol{v}_j$ die entsprechenden normalisierten Eigenvektoren von $\boldsymbol{\Sigma}$ sind mit $\boldsymbol{v}_j^T\boldsymbol{v}_j = 1$ und $\boldsymbol{v}_j^T\boldsymbol{v}_k = 0$, $k \neq j$.

Es lässt sich zeigen, dass sich die Inverse der Quadratwurzel der Kovarianzmatrix durch den Übergang zur Spektraldarstellung folgendermaßen bestimmen lässt (Harville, 2008, S. 545/550; Mukhopadhyay, 2009, S. 476) (für positive Eigenwerte):

$$\boldsymbol{\Sigma}^{-1/2} = \begin{pmatrix} \boldsymbol{v}_1 & \ldots & \boldsymbol{v}_p \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{\gamma_1}} & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \frac{1}{\sqrt{\gamma_p}} \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_p^T \end{pmatrix}. \tag{3.39}$$

Mithilfe dieser Umformungen kann das Maximierungsproblem (3.37) umdefiniert werden (Mukhopadhyay, 2009, S. 430):

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^p} \frac{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}\boldsymbol{\alpha}} =: \max_{\boldsymbol{u} \in \mathbb{R}^p} \frac{\boldsymbol{u}^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{u}}{\boldsymbol{u}^T\boldsymbol{u}}.$$

Daher lässt sich analog das Produkt $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ betrachten, dessen Eigenwertzerlegung numerisch stabiler als jene von $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ ist. Im Folgenden seien dazu mit $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r > 0$ die $r \leq \min(p, M-1)$ positiven Eigenwerte von $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ bzw. $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ bezeichnet. Eine Erklärung, wieso sie für beide Eigenwertprobleme identisch sind, folgt in (3.40).

Die Anzahl $r \leq \min(p, M-1)$ ergibt sich durch die Betrachtung der Ränge der Kovarianzmatrizen. Es gilt $\operatorname{rg}(\boldsymbol{\Sigma}^{-1}) \leq p$, da $\boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{p \times p}$ und $\operatorname{rg}(\boldsymbol{B}) \leq M-1$, da $\sum_{c=1}^{M} (\boldsymbol{\mu}^{(c)} - \bar{\boldsymbol{\mu}}) = \boldsymbol{0}$. Die M Vektoren $(\boldsymbol{\mu}^{(c)} - \bar{\boldsymbol{\mu}})$ aus der Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ aus (3.36) sind somit linear abhängig, mindestens ein Vektor lässt sich in Abhängigkeit der anderen darstellen, weshalb der durch die M Vektoren aufgespannte Raum $\{(\boldsymbol{\mu}^{(1)} - \bar{\boldsymbol{\mu}}), \ldots, (\boldsymbol{\mu}^{(M)} - \bar{\boldsymbol{\mu}})\}$ höchstens $(M-1)$-dimensional ist (van Meegen, 2015, S. 10; Johnson und Wichern, 2007, S. 629; Mukhopadhyay, 2009, S. 531). Mithilfe von Rechenregeln bezüglich des Rangs

eines Produktes von Matrizen gilt (Fischer, 2014, S. 149): $\text{rg}(\boldsymbol{\Sigma}^{-1}\boldsymbol{B}) \leq \min(\text{rg}(\boldsymbol{\Sigma}^{-1}), \text{rg}(\boldsymbol{B}))$ $\leq \min(p, M-1)$. Da der Rang einer symmetrischen Matrix identisch zur Anzahl an positiven Eigenwerten ist (Schmidt und Trenkler, 2006, S. 91), gibt es also insgesamt $r \leq \min(p, M-1)$ positive Eigenwerte.

Bezeichne mit $\boldsymbol{\nu}_j$ die normalisierten Eigenvektoren von $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$, also $\boldsymbol{\nu}_j^T\boldsymbol{\nu}_j = 1$ und $\boldsymbol{\nu}_j^T\boldsymbol{\nu}_k = 0,\ k \neq j$. Die Matrix $\boldsymbol{N}$ umfasse alle r Eigenvektoren: $\boldsymbol{N} = \begin{pmatrix}\boldsymbol{\nu}_1 & \dots & \boldsymbol{\nu}_r\end{pmatrix}$. Aufgrund der Orthogonalität der Eigenvektoren (Mukhopadhyay, 2009, S. 430 f.) gilt $\boldsymbol{N}\boldsymbol{N}^T = \boldsymbol{N}^T\boldsymbol{N} = \boldsymbol{I}_{r\times r}$. Die Eigenwerte beider Matrizenprodukte $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ und $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ sind identisch, da (Mukhopadhyay, 2009, S. 431)

$$\begin{aligned}
&\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_j &&= \lambda_j\boldsymbol{\nu}_j \\
\Leftrightarrow &\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_j &&= \boldsymbol{\Sigma}^{-1/2}\lambda_j\boldsymbol{\nu}_j \\
\Leftrightarrow &\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\underbrace{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_j}_{=:\boldsymbol{\alpha}_j} &&= \lambda_j\underbrace{\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_j}_{=:\boldsymbol{\alpha}_j}.
\end{aligned} \tag{3.40}$$

Die zugehörigen Eigenvektoren unterscheiden sich. Die Eigenvektoren $\boldsymbol{\alpha}_j$ von $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ sind jedoch proportional zu $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_j$ und lassen sich somit durch Transformation aus den Eigenvektoren $\boldsymbol{\nu}_j$ des Eigenwertproblems von $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ bestimmen durch

$$\boldsymbol{\alpha}_j := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_j. \tag{3.41}$$

In Matrixschreibweise gilt für alle r Eigenvektoren gemeinsam:

$$\boldsymbol{A} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{N} \Leftrightarrow \begin{pmatrix}\boldsymbol{\alpha}_1 & \dots & \boldsymbol{\alpha}_r\end{pmatrix} = \boldsymbol{\Sigma}^{-1/2}\begin{pmatrix}\boldsymbol{\nu}_1 & \dots & \boldsymbol{\nu}_r\end{pmatrix}. \tag{3.42}$$

Der Eigenvektor $\boldsymbol{\alpha}_1 := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_1$ zum größten Eigenwert λ_1 zur Lösung des Maximierungsproblems (3.35) bzw. (3.37) wird *1. Diskriminanzkomponente* genannt. Die *2. Diskriminanzkomponente* $\boldsymbol{\alpha}_2 := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_2$ zum zweitgrößten Eigenwert λ_2 hat zusätzlich die Eigenschaft $\text{Cov}(\boldsymbol{\alpha}_1^T\boldsymbol{X}, \boldsymbol{\alpha}_2^T\boldsymbol{X}) = 0$ und maximiert (3.35) bzw. (3.37) unter eben dieser Nebenbedingung. Die *k-te Diskriminanzkomponente* $\boldsymbol{\alpha}_k := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_k$ zum k-ten größten Eigenwert λ_k löst das Problem unter der Nebenbedingung $\text{Cov}(\boldsymbol{\alpha}_l^T\boldsymbol{X}, \boldsymbol{\alpha}_k^T\boldsymbol{X}) = 0$, $l \leq k$, $k = 3, \dots, r$. Als zusätzliche Eigenschaft besitzen alle Diskriminanzfunktionen Varianz 1, da (Mukhopadhyay, 2009, S. 430 f.)

$$\text{Var}\left(\boldsymbol{\alpha}_j^T\boldsymbol{X}\right) = \boldsymbol{\alpha}_j^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_j = \boldsymbol{\nu}_j^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_j = \boldsymbol{\nu}_j^T\boldsymbol{\nu}_j = 1. \tag{3.43}$$

Damit gilt auch die Eigenschaft der Diskriminanzkomponenten: $\boldsymbol{\alpha}_j^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_j = 1$. Für die Kovarianzen gilt (siehe oben):

$$\text{Cov}\left(\boldsymbol{\alpha}_l^T\boldsymbol{X}, \boldsymbol{\alpha}_k^T\boldsymbol{X}\right) = \boldsymbol{\alpha}_l^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_k = \boldsymbol{\nu}_l^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_k = \boldsymbol{\nu}_l^T\boldsymbol{\nu}_k = 0, \tag{3.44}$$

da $\boldsymbol{\nu}_l^T$ und $\boldsymbol{\nu}_k$ orthogonal zueinander sind (Mukhopadhyay, 2009, S. 430 f.).

Die Diskriminanzkomponenten bzw. Diskriminanzfunktionen werden einerseits zur Dimensionsreduktion verwendet. Die p-dimensionalen Zufallsvektoren $\boldsymbol{X}^{(c)}$ werden durch (3.34) in r-dimensionale Zufallsvektoren $\boldsymbol{\mathcal{Y}}^{(c)}$ transformiert. Im Zusammenhang mit Klassifikationsverfahren kann diese geometrische Idee jedoch andererseits zur Trennung von Beobachtungen bzw. der Klassifikation neuer Beobachtungen in eine von M vorliegenden Klassen herangezogen werden.

Eine einfache Form der Klassifikationsregel bei der Fisher LDA formulieren Mardia et al. (1979, S. 319) für den Fall, dass nur eine Diskriminanzkomponente $\boldsymbol{\alpha}$ betrachtet wird: Eine Beobachtung $\boldsymbol{x}^*$ wird Klasse k zugeordnet, falls

$$\left|\boldsymbol{\alpha}^T\boldsymbol{x}^* - \boldsymbol{\alpha}^T\boldsymbol{\mu}^{(k)}\right| < \left|\boldsymbol{\alpha}^T\boldsymbol{x}^* - \boldsymbol{\alpha}^T\boldsymbol{\mu}^{(j)}\right| \quad \text{für alle } j \neq k.$$

Eine Beobachtung wird demnach der Klasse k zugeordnet, zu deren projizierten Erwartungswertvektor die projizierte Beobachtung den geringsten absoluten Abstand hat.

Eine weitere klassische Form der Klassifikationsregel bei der Fisher LDA, welche mehrere oder sogar alle r Diskriminanzkomponenten beachtet, formulieren Johnson und Wichern (2007, S. 629). Eine Beobachtung $\boldsymbol{x}^*$ wird der Klasse

$$k = \underset{c=1,\ldots,M}{\arg\min}\; h_F^{(c)}(\boldsymbol{x}^*) \tag{3.45}$$

zugeordnet, wobei

$$h_F^{(c)}(\boldsymbol{x}) := \sum_{j=1}^{r}\left(\boldsymbol{\alpha}_j^T\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)\right)^2 = \left(\boldsymbol{A}^T\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)\right)^T\left(\boldsymbol{A}^T\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)\right). \tag{3.46}$$

Die Idee besteht darin, dass eine neue Beobachtung $\boldsymbol{x}^*$ jener Klasse k zugeordnet wird, zu deren projizierten Erwartungswertvektor $\boldsymbol{\alpha}_j^T\boldsymbol{\mu}^{(k)}$ ihr projizierter Wert $\boldsymbol{\alpha}_j^T\boldsymbol{x}^*$ den geringsten quadrierten Abstand hat summiert über alle Dimensionen $j = 1,\ldots,r < p$, d. h. zu welchem Erwartungswertvektor die Beobachtung im niedriger-dimensionalen Raum am nächsten liegt. Eine Einbeziehung der Kovarianzmatrix ist hier unerheblich, da nach (3.43) und (3.44) die Varianzen aller Diskriminanzfunktionen Eins sind und ihre Kovarianzen Null. Die geometrische Idee ist in Abbildung 3.4 für eine zweidimensionale Verteilung und zwei Klassen veranschaulicht.

Eine alternative Diskriminanzregel schlagen Filzmoser et al. (2006, S. 522) durch die Einführung eines Strafterms basierend auf den a-priori Wahrscheinlichkeiten der Klassen vor:

$$k = \underset{c=1,\ldots,M}{\arg\min}\; h_{F,2}^{(c)}(\boldsymbol{x}^*), \tag{3.47}$$

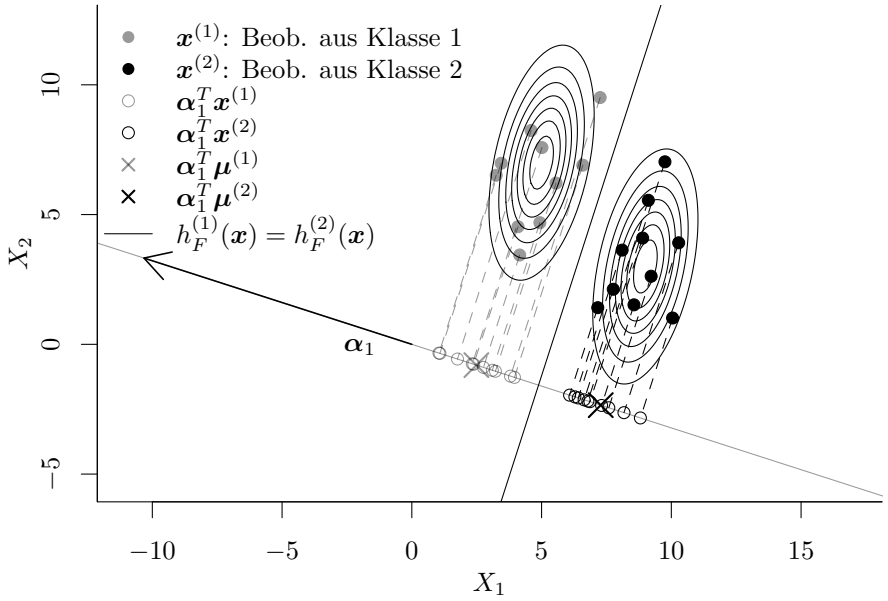


Abbildung 3.4: Veranschaulichung der geometrischen Idee der Fisher LDA für eine zweidimensionale Verteilung und zwei Klassen. Es erfolgt eine Projektion in den eindimensionalen Raum, sodass die Varianz innerhalb der Klassen möglichst klein und jene Zwischen-den-Klassen möglichst groß ist. Für die 1. Diskriminanzkomponente (bestimmt durch (3.41)) gilt: $\boldsymbol{\alpha}_1^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = 1$.

wobei

$$\begin{aligned} h_{F,2}^{(c)}(\boldsymbol{x}) :&= \sum_{j=1}^{r}\left(\boldsymbol{\alpha}_j^T\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)\right)^2 - 2\log p^{(c)} \\ &= \left(\boldsymbol{A}^T\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)\right)^T\left(\boldsymbol{A}^T\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)\right) - 2\log p^{(c)}. \end{aligned} \tag{3.48}$$

Wie bei der Kanonischen Linearen Diskriminanzanalyse werden Klassen mit einer geringen a-priori Wahrscheinlichkeit durch den negativen Summanden stärker bestraft, da $p^{(c)} \in (0,1) \Rightarrow \lim_{p^{(c)}\to 0} 2\log p^{(c)} = -\infty$ und $\lim_{p^{(c)}\to 1} 2\log p^{(c)} = 0$. Die Klassifikationsgrenzen verschieben sich dadurch zugunsten der Klasse mit verhältnismäßig größerer a-priori Wahrscheinlichkeit.

Die Klassifikationsgrenzen setzen sich auch bei der Fisher LDA aus einer Menge von $\binom{M}{2}$ Hyperebenen zusammen: $\left\{\boldsymbol{x} : h_F^{(j)}(\boldsymbol{x}) = h_F^{(k)}(\boldsymbol{x})\right\}$ bzw. $\left\{\boldsymbol{x} : h_{F,2}^{(j)}(\boldsymbol{x}) = h_{F,2}^{(k)}(\boldsymbol{x})\right\}$ für $j,k \in \{1,\ldots,M\}$, $j \neq k$.

Schätzung der Parameter Auch bei der Fisher LDA sind die betrachteten Parameter im Allgemeinen in der Anwendung nicht bekannt und müssen geschätzt werden. Zusätzlich zu den ebenfalls für die Kanonische LDA benötigten Schätzern für die Mittelwertvektoren der Klassen und die Kovarianzmatrix innerhalb der Klassen (vgl. Seite 42) wird hier ein Schätzer für die Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ benötigt.

In der Literatur werden dazu verschiedene Schätzer vorgeschlagen. Diese verschiedenen gewichteten und ungewichteten Schätzer werden von van Meegen (2015, S. 11 f.) zusammengefasst und gegenübergestellt.

Die Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ basiert auf dem Mittelwert $\bar{\boldsymbol{\mu}}$ der Erwartungswertvektoren über alle Klassen. Auch für diesen wird bei der Schätzung ein empirisches Äquivalent benötigt. Johnson und Wichern (2007, S. 623) betrachten einen einfachen ungewichteten Mittelwertschätzer:

$$\hat{\bar{\boldsymbol{\mu}}} = \frac{1}{M} \sum_{c=1}^{M} \hat{\boldsymbol{\mu}}^{(c)}. \tag{3.49}$$

Van Meegen (2015, S. 12), van Meegen et al. (2019, S. 5) und Filzmoser et al. (2006, S. 521) schlagen eine durch die a-priori Wahrscheinlichkeiten gewichtete Variante vor:

$$\hat{\bar{\boldsymbol{\mu}}}_{\text{gew}} = \sum_{c=1}^{M} \hat{p}^{(c)} \hat{\boldsymbol{\mu}}^{(c)}, \tag{3.50}$$

wobei die a-priori Wahrscheinlichkeiten als Gewichte auch geschätzt werden müssen, falls sie nicht bekannt sind. Duda et al. (2001, S. 121) konkretisieren diese Gewichtung durch

$$\hat{\bar{\boldsymbol{\mu}}}_{\text{gew},2} = \frac{1}{n} \sum_{c=1}^{M} n^{(c)} \hat{\boldsymbol{\mu}}^{(c)}. \tag{3.51}$$

Falls die a-priori Wahrscheinlichkeiten durch die relativen Häufigkeiten geschätzt werden, sind (3.50) und (3.51) identisch. Falls eine Gleichverteilung betrachtet wird, sind hingegen (3.49) und (3.50) identisch. Zu beachten ist dabei, dass die beiden letzteren Schätzer eher „ungewichtete“ Schätzer bezüglich des gesamten Erwartungswertvektors sind. Werden nämlich die einzelnen Mittelwertvektoren der Klassen nicht gewichtet und sind die Klassen unterschiedlich groß, so ist (3.49) kein unverzerrter Schätzer für den gesamten Erwartungswertvektor $\boldsymbol{\mu}$.

In direkter Anlehnung an die theoretische Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ nach (3.36) definieren Johnson und Wichern (2007, S. 623) den intuitiven Schätzer:

$$\hat{\boldsymbol{B}} = \sum_{c=1}^{M} \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\bar{\boldsymbol{\mu}}} \right) \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\bar{\boldsymbol{\mu}}} \right)^T. \tag{3.52}$$

Krzanowski und Marriott (1995, S. 7) fassen zwei Varianten mit Vorfaktor zusammen. Zum einen eine mit Verweis auf Rao (1948, S. 188) (erweitert um den Vorfaktor) ungewichtete Variante

$$\hat{\boldsymbol{B}}^* = \frac{1}{n - M} \sum_{c=1}^{M} \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\bar{\boldsymbol{\mu}}} \right) \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\bar{\boldsymbol{\mu}}} \right)^T. \tag{3.53}$$

Zum anderen einen entsprechenden gewichteten Schätzer, welcher von Bryan (1951, S. 90 f.) betrachtet und um einen Vorfaktor ergänzt wird:

$$\hat{\boldsymbol{B}}^{*}_{\text{gew}} = \frac{1}{n-M} \sum_{c=1}^{M} n^{(c)} \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\tilde{\boldsymbol{\mu}}}_{\text{gew},2} \right) \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\tilde{\boldsymbol{\mu}}}_{\text{gew},2} \right)^{T}. \tag{3.54}$$

Auch Fahrmeir et al. (1996a, S. 52/380), Mukhopadhyay (2009, S. 432) und Duda et al. (2001, S. 122) beziehen eine Gewichtung der einzelnen Klassen mit ihren Klassengrößen ein, betrachten dabei den gewichteten Mittelwertschätzer (3.51) und verzichten wie Bryan auf den Vorfaktor:

$$\hat{\boldsymbol{B}}_{\text{gew},2} = \sum_{c=1}^{M} n^{(c)} \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\tilde{\boldsymbol{\mu}}}_{\text{gew},2} \right) \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\tilde{\boldsymbol{\mu}}}_{\text{gew},2} \right)^{T}. \tag{3.55}$$

Im Kontext der Fisher LDA spielt der Vorfaktor jedoch keine Rolle. Die Eigenvektoren im Eigenwertproblem bleiben von dieser Konstante unberührt, lediglich die Eigenwerte sind um den Faktor $(n-M)$ größer (van Meegen, 2015, S. 12). Daher ist es irrelevant, ob (3.52) oder (3.53) bzw. (3.54) oder (3.55) betrachtet werden.

Anlehnend an die theoretische gewichtete Zwischen-den-Klassen Kovarianzmatrix (3.38) nach Filzmoser et al. (2006, S. 522) schlagen van Meegen (2015, S. 12) und van Meegen et al. (2019, S. 6) der Vollständigkeit halber auch den folgenden Schätzer vor:

$$\hat{\boldsymbol{B}}_{\text{gew}} = \sum_{c=1}^{M} \hat{p}^{(c)} \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}} \right) \left(\hat{\boldsymbol{\mu}}^{(c)} - \hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}} \right)^{T}. \tag{3.56}$$

3.4 Vergleich von Kanonischer und Fisher LDA

In der Literatur herrscht Uneinigkeit darüber, in welchen Situationen die Kanonische LDA und die Fisher LDA identische Klassifikationsergebnisse liefern. Beispielsweise stellen Fahrmeir et al. (1996a, S. 381) heraus, dass beide Verfahren unter der Annahme identischer a-priori Wahrscheinlichkeiten identisch sind und verweisen auf ein Manuskript von Knüsel (1993). Rencher und Christensen (2012, S. 313) hingegen erwähnen die Übereinstimmung der Klassifikationsregel nur für den Zwei-Klassen-Fall und $p^{(1)} = p^{(2)}$. Über den allgemeinen Fall wird jedoch keine eindeutige Aussage getroffen, sodass unklar bleibt, ob diese Übereinstimmung sich auf den betrachteten Spezialfall beschränkt. Insbesondere die Vergleichbarkeit beider Verfahren für den Fall, dass ungleiche a-priori Wahrscheinlichkeiten der Klassen unterstellt werden, ist somit unklar.

Van Meegen (2015) hat sich in ihrer Bachelorarbeit „Ungleiche a priori Wahrscheinlichkeiten in linearen Diskriminanzanalyseverfahren“ daher ausführlich mit einem theoretischen Vergleich der Kanonischen Linearen Diskriminanzanalyse und der Fisher Diskriminanzanalyse auseinandergesetzt. Zudem wird anhand einer Simulationsstudie untersucht, welche

Auswirkung die Verwendung der unterschiedlichen Schätzer für Erwartungswertvektoren und Kovarianzmatrizen in der Anwendung hat. Es erfolgt somit ein Vergleich der Methoden sowohl in der Theorie als auch in der Anwendung.

Van Meegen (2015) zeigt dabei, dass die Kanonische LDA und die Fisher LDA in der Theorie dieselben Ergebnisse liefern – unabhängig von der Anzahl an Klassen M, Anzahl an Variablen p und der Form der a-priori Wahrscheinlichkeiten. Die wichtigsten Ergebnisse wurden von van Meegen et al. (2019) zusammengefasst.

Zwei Klassen und zwei Merkmale ($M = 2$, $p = 2$) Für den Zwei-Klassen-Fall und zwei Dimensionen erfolgt der Beweis über die Übereinstimmung beider Methoden durch den Vergleich beider Trennhyperebenen.

Die Diskriminanzkomponenten der Fisher LDA wurden in Abschnitt 3.3 hergeleitet. Nach (3.41) gilt

$$\boldsymbol{\alpha}_j := \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\nu}_j$$

mit der Eigenschaft $\boldsymbol{\alpha}_j^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_j = 1$. Im Falle von zwei Klassen und zwei Dimensionen kann eine Diskriminanzkomponente hergeleitet werden ($r \leq \min(p, M-1) = \min(2, 2-1) = 1$). Laut Pires und Branco (1996, S. 415) ist diese Diskriminanzkomponente proportional zu $\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}\right)$: $\boldsymbol{\alpha}_1 \propto \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}\right)$.

Van Meegen (2015, S. 14 ff.) und van Meegen et al. (2019, S. 14 f.) leiten ausführlich her, dass die exakte Form folgendermaßen aussieht:

$$\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12})^T := \frac{\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}\right)}{\left(\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)}\right)\right)^{1/2}}. \tag{3.57}$$

Die Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ hat somit keinen Einfluss auf die Diskriminanzkomponente, ebenso wenig wie der gemittelte Erwartungswertvektor $\bar{\boldsymbol{\mu}}$. In der Praxis hat demnach die Wahl des Schätzers für $\boldsymbol{B}$ bzw. $\bar{\boldsymbol{\mu}}$ keinen direkten Einfluss auf die Form der Diskriminanzkomponente.

Die lineare Trenngerade bei der Fisher LDA ergibt sich durch Gleichsetzen von (3.46) bzw. (3.48) für beide Klassen (van Meegen, 2015, S. 17; van Meegen et al., 2019, S. 16):

$$\begin{aligned}
& h_{F,2}^{(1)}(\boldsymbol{x}) = h_{F,2}^{(2)}(\boldsymbol{x}) \\
\Leftrightarrow \quad & \left(\boldsymbol{\alpha}_1^T\left(\boldsymbol{x} - \boldsymbol{\mu}^{(1)}\right)\right)^2 - 2\log p^{(1)} = \left(\boldsymbol{\alpha}_1^T\left(\boldsymbol{x} - \boldsymbol{\mu}^{(2)}\right)\right)^2 - 2\log p^{(2)} \\
\Leftrightarrow \quad & -2\log p^{(1)} + 2\log p^{(2)} = \left(\boldsymbol{\alpha}_1^T\left(\boldsymbol{x} - \boldsymbol{\mu}^{(1)}\right) + \boldsymbol{\alpha}_1^T\left(\boldsymbol{x} - \boldsymbol{\mu}^{(2)}\right)\right) \\
& \qquad \cdot \left(\boldsymbol{\alpha}_1^T\left(\boldsymbol{x} - \boldsymbol{\mu}^{(2)}\right) - \boldsymbol{\alpha}_1^T\left(\boldsymbol{x} - \boldsymbol{\mu}^{(1)}\right)\right) \\
\Leftrightarrow \quad & -\boldsymbol{\alpha}_1^T\left(2\boldsymbol{x} - \boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}\right)\boldsymbol{\alpha}_1^T\left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)}\right) = 2\log p^{(1)} - 2\log p^{(2)}
\end{aligned}$$

$$\Leftrightarrow \quad -2\boldsymbol{\alpha}_1^T \left(\boldsymbol{x} - \frac{\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}}{2} \right) = \frac{2\log p^{(1)} - 2\log p^{(2)}}{\boldsymbol{\alpha}_1^T \left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)} \right)}$$

$$\Leftrightarrow \quad \boldsymbol{\alpha}_1^T \left(\boldsymbol{x} - \bar{\boldsymbol{\mu}} \right) = -\frac{\log p^{(1)} - \log p^{(2)}}{\boldsymbol{\alpha}_1^T (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)})}$$

$$\Leftrightarrow \quad \alpha_{11} x_1 + \alpha_{12} x_2 = \boldsymbol{\alpha}_1^T \bar{\boldsymbol{\mu}} - \frac{\log p^{(1)} - \log p^{(2)}}{\boldsymbol{\alpha}_1^T \left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(2)} \right)}$$

$$\Leftrightarrow \quad x_2 = -\frac{\alpha_{11}}{\alpha_{12}} \cdot x_1 + \frac{\boldsymbol{\alpha}_1^T \bar{\boldsymbol{\mu}}}{\alpha_{12}} + \frac{\log p^{(1)} - \log p^{(2)}}{\alpha_{12} \boldsymbol{\alpha}_1^T \left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right)}.$$

Mit (3.57) gilt:

$$x_2 = -\frac{\alpha_{11}}{\alpha_{12}} \cdot x_1 + \frac{\boldsymbol{\alpha}_1^T \bar{\boldsymbol{\mu}}}{\alpha_{12}} + \frac{\log p^{(1)} - \log p^{(2)}}{\alpha_{12} \left(\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right) \right)^{1/2}}. \tag{3.58}$$

Die lineare Trenngerade bei der Kanonischen LDA wurde bereits in (3.21) hergeleitet:

$$x_2 = -\frac{a_1}{a_2} \cdot x_1 + \frac{\boldsymbol{a}^T}{a_2} \cdot \frac{\boldsymbol{\mu}^{(1)} + \boldsymbol{\mu}^{(2)}}{2} + \frac{\log p^{(1)} - \log p^{(2)}}{a_2}$$

mit $\boldsymbol{a} := (a_1, a_2)^T = \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right)$.

Da nun mit (3.57)

$$\begin{aligned} \boldsymbol{a} = (a_1, a_2)^T &= \boldsymbol{\alpha}_1 \left(\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right) \right)^{1/2} \\ &= \begin{pmatrix} \alpha_{11} \left(\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right) \right)^{1/2} \\ \alpha_{12} \left(\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right) \right)^{1/2} \end{pmatrix} \end{aligned}$$

lässt sich die lineare Trenngerade der Kanonischen LDA somit auch in Abhängigkeit der Diskriminanzkomponente schreiben:

$$x_2 = -\frac{\alpha_{11}}{\alpha_{12}} \cdot x_1 + \frac{\boldsymbol{\alpha}_1^T}{\alpha_{12}} \cdot \bar{\boldsymbol{\mu}} + \frac{\log p^{(1)} - \log p^{(2)}}{\alpha_{12} \left(\left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(1)} \right) \right)^{1/2}}.$$

Für $M = 2$ und $p = 2$ sind somit die linearen Trenngeraden der Kanonischen LDA $h_L^{(1)}(\boldsymbol{x}) = h_L^{(2)}(\boldsymbol{x})$ und der Fisher LDA $h_{F,2}^{(1)}(\boldsymbol{x}) = h_{F,2}^{(2)}(\boldsymbol{x})$ identisch.

Mehr als zwei Klassen und zwei Merkmale ($M > 2$, $p > 2$) Bei der Betrachtung von $M > 2$ Klassen und mehr als zwei Merkmalen ist die Herleitung der Trennhyperebenen komplexer, weshalb van Meegen (2015, S. 19 ff.) (bzw. van Meegen et al. (2019, S. 6 ff.)) den Fokus auf den Vergleich der Klassifikationsregeln legt.

Bezeichne $\boldsymbol{N}^* := \begin{pmatrix} \boldsymbol{N} & \boldsymbol{\nu}_{r+1} & \dots & \boldsymbol{\nu}_p \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nu}_1 & \dots & \boldsymbol{\nu}_r & \dots & \boldsymbol{\nu}_p \end{pmatrix}$ die Matrix, welche alle p normalisierten Eigenvektoren zu den Eigenwerten $\lambda_1 \geq \dots \geq \lambda_r > 0$, $\lambda_{r+1} = \dots =$

$\lambda_p = 0$ der Matrix $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ umfasst (vgl. Seite 49). Die Betrachtung der zusätzlichen $(p-r)$ Eigenvektoren mit Eigenwert gleich Null ändert nichts an der Orthogonalität, weshalb $\boldsymbol{N}^*\boldsymbol{N}^{*^T} = \boldsymbol{N}^{*^T}\boldsymbol{N}^* = \boldsymbol{I}_{p\times p}$.

Sei zudem $\boldsymbol{A}^* := \begin{pmatrix}\boldsymbol{A} & \boldsymbol{\alpha}_{r+1} & \dots & \boldsymbol{\alpha}_p\end{pmatrix} = \begin{pmatrix}\boldsymbol{\alpha}_1 & \dots & \boldsymbol{\alpha}_r & \dots & \boldsymbol{\alpha}_p\end{pmatrix}$ die Matrix, welche alle p Diskriminanzkomponenten enthält. Nach (3.40) und (3.42) lassen sich die Eigenvektoren $\boldsymbol{\alpha}_j$ der Matrix $\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ aus den Eigenvektoren $\boldsymbol{\nu}_j$ des Matrizenproduktes $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ bestimmen: $\boldsymbol{A}^* = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{N}^*$. Die Diskriminanzfunktion der Fisher LDA mit Strafterm (3.48) lässt sich bei Betrachtung aller p Diskriminanzkomponenten unter diesen Vorüberlegungen folgendermaßen umformen (van Meegen, 2015, S. 20; van Meegen et al., 2019, S. 7):

$$\begin{aligned}
h_{F,2}^{*(c)}(\boldsymbol{x}) &= \sum_{j=1}^{p}\left(\boldsymbol{\alpha}_j^T\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)\right)^2 - 2\log p^{(c)} \\
&= \left(\boldsymbol{A}^{*^T}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)\right)^T\left(\boldsymbol{A}^{*^T}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)\right) - 2\log p^{(c)} \\
&= \left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{N}^*\boldsymbol{N}^{*^T}\boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right) - 2\log p^{(c)} \\
&= \left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)^T\boldsymbol{\Sigma}^{-1/2}\boldsymbol{I}_{p\times p}\boldsymbol{\Sigma}^{-1/2}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right) - 2\log p^{(c)} \\
&= \left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{x}-\boldsymbol{\mu}^{(c)}\right) - 2\log p^{(c)} \\
&= h_{L,2}^{(c)}(\boldsymbol{x}) \quad \text{(mit quadratischem Term, vgl. Seite 41).}
\end{aligned}$$

Bei Betrachtung aller p Diskriminanzkomponenten sind somit die Diskriminanzfunktionen mit Strafterm $h_{F,2}^{*(c)}(\boldsymbol{x})$ der Fisher LDA und $h_{L,2}^{(c)}(\boldsymbol{x})$ (mit quadratischem Term) der Kanonischen LDA identisch. In (3.48) fließen jedoch nur die ersten $r < p$ Diskriminanzkomponenten ein. Dies ändert jedoch nichts an der Zuordnung der Diskriminanzregel in der Fisher LDA, da nach Johnson und Wichern (2007, S. 630) die letzten $(p-r)$ Summanden $\boldsymbol{\alpha}_j^T(\boldsymbol{x}^* - \boldsymbol{\mu}^{(c)})$, $j = r+1, \dots, p$, in $h_{F,2}^{*(c)}(\boldsymbol{x})$ für alle Klassen $c = 1, \dots, M$ identisch sind (van Meegen, 2015, S. 20 f.; van Meegen et al., 2019, S. 7). Diese tragen somit nicht zur Trennung der Klassen bei, da sie gleichermaßen in das Minimierungsproblem von (3.48) einfließen. Zudem ist es irrelevant, ob der quadratische Term in $h_{L,2}^{(c)}(\boldsymbol{x})$ betrachtet oder wie in (3.23) eliminiert wird, da dieser für alle Klassen identisch ist. Die Klassifikationsregeln (3.24) und (3.47) sind somit – auch für unterschiedliche a-priori Wahrscheinlichkeiten der Klassen – identisch:

$$\underset{c=1,\dots,M}{\arg\min}\; h_{L,2}^{(c)}(\boldsymbol{x}) = \underset{c=1,\dots,M}{\arg\min}\; h_{F,2}^{(c)}(\boldsymbol{x}).$$

Die Kanonische LDA und die Fisher LDA liefern somit in der Theorie identische Resultate für eine beliebige Anzahl an Klassen und Merkmalen, sofern jeweils die Diskriminanzfunktionen mit Strafterm betrachtet werden. Wird jedoch die erste Variante (3.46) der Diskriminanzfunktion der Fisher LDA ohne Strafterm betrachtet, so sind beide Methoden nur im Falle identischer a-priori Wahrscheinlichkeiten in den Klassen gleich.

In der Praxis müssen die Parameter jedoch geschätzt werden. Entsprechende gewichtete und ungewichtete Schätzer wurden in Abschnitt 3.2.1 und 3.3 vorgestellt. Wie bereits auf Seite 54 erwähnt, hat die Schätzung der Zwischen-den-Klassen-Kovarianzmatrix $\boldsymbol{B}$ keinen Einfluss auf die Diskriminanzkomponenten und demnach die Trennhyperebenen in der Fisher LDA. Dies gilt ebenso für den Mehrklassenfall wie van Meegen (2015, S. 31 ff.) anhand einer Simulationsstudie verdeutlicht (s. auch van Meegen et al. (2019, S. 10 ff.)). Es ist somit irrelevant, ob (3.52) oder (3.56) als Schätzer herangezogen wird.

Van Meegen (2015, S. 22 ff.) vergleicht die Klassifikationsergebnisse der Kanonischen LDA und der Fisher LDA unter Verwendung verschiedener gewichteter und ungewichteter Schätzer für die Erwartungswertvektoren und Kovarianzmatrizen für zwei ($M = 2$) und drei Klassen ($M = 3$) sowie zwei Merkmale ($p = 2$) anhand einer Simulationsstudie. Als vergleichendes Zielkriterium jeder Kombination wird dabei die relative Anzahl an unterschiedlich klassifizierten Beobachtungen eines zweidimensionalen Gitters (wegen $p = 2$) betrachtet. Dabei werden feste Erwartungswertvektoren der einzelnen Klassen gewählt. Die Kovarianzmatrix ist für alle Klassen identisch, jedoch werden verschiedene Korrelationsstrukturen zwischen den Merkmalen betrachtet. Zudem werden verschiedene Aufteilungen der Anzahl an Beobachtungen in den einzelnen Klassen betrachtet, um implizit verschiedene Konstellationen von a-priori Wahrscheinlichkeiten abzubilden. Die Schätzer für die a-priori Wahrscheinlichkeiten werden auch variiert, sodass diese in manchen Fällen mit den relativen Häufigkeiten der Klassen übereinstimmen und in anderen Fällen komplett abweichen. Die Schlussfolgerungen aus dieser Simulationsstudie sind in Tabelle 3.1 zusammengefasst.

Hauptresultate: Die Kanonische LDA und Fisher LDA liefern in der praktischen Anwendung dieselben Ergebnisse, sofern dieselben Schätzer für die Kovarianzmatrix $\boldsymbol{\Sigma}$ verwendet werden. Die Wahl des Schätzers für die Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ hat laut Simulationsergebnissen auch im Mehrklassenfall keine Auswirkung auf das Klassifikationsergebnis. Die Wahl des Schätzers für die Kovarianzmatrix $\boldsymbol{\Sigma}$ kann die Trennhyperebenen verändern. Dies wird zumindest für $p = 2$ Merkmale gezeigt und wird auch anhand des theoretischen Ergebnisses der linearen Trenngerade (3.58) deutlich. Im Allgemeinen sind bei unterschiedlichen Klassengrößen $n^{(c)}$ die Trennhyperebenen bei Betrachtung von $\hat{\boldsymbol{\Sigma}}$ und $\hat{\boldsymbol{\Sigma}}_{\text{gew}}$ und verschiedenen Schätzern für die a-priori Wahrscheinlichkeiten $p^{(c)}$ verschieden. Lediglich bei gleich großen Klassen $n^{(c)}$ und identischen a-priori Wahrscheinlichkeiten $p^{(c)}$ bzw. bei Verwendung der relativen Häufigkeiten als Schätzer oder Betrachtung einer Gleichverteilung $\hat{p}^{(c)} = \frac{n^{(c)}}{n} = \frac{1}{n/n^{(c)}} = \frac{1}{M}$ sind die Trennhyperebenen für $\hat{\boldsymbol{\Sigma}}$ und $\hat{\boldsymbol{\Sigma}}_{\text{gew}}$ identisch (vgl. auch Fußnote zu Tabelle 3.1). Werden nicht die relativen Häufigkeiten zur Schätzung der a-priori Wahrscheinlichkeiten herangezogen, verschiebt sich die Trennhyperebene bei Verwendung von $\hat{\boldsymbol{\Sigma}}$ entlang der Achsen gegenüber jener, welche unter Verwendung von (3.29) und $\hat{\boldsymbol{\Sigma}}$ resultiert. Der „Intercept“ ändert sich, die „Steigung“ jedoch nicht (vgl. (3.58)). Werden die Trennhyperebenen unter Verwendung des Schätzers $\hat{\boldsymbol{\Sigma}}_{\text{gew}}$ verglichen,

Tabelle 3.1: Vergleich der Resultate aller Kombinationen von Kanonischer LDA (**G**(aussian)**DA**) und Fisher LDA (**F**(isher)**DA**) unter Verwendung verschiedener gewichteter und ungewichteter Schätzer für die Parameter: $\hat{\tilde{\boldsymbol{\mu}}}$ nach (3.49), $\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$ nach (3.50), $\hat{\boldsymbol{\mu}}^{(c)}$ nach (3.25), $\hat{\boldsymbol{B}}$ nach (3.52), $\hat{\boldsymbol{B}}_{\text{gew}}$ nach (3.56), $\hat{\boldsymbol{\Sigma}}$ nach (3.27), $\hat{\boldsymbol{\Sigma}}_{\text{gew}}$ nach (3.28).

Kombinationen für paarweise Vergleiche									identische Resultate?*
	Schätzer für					Schätzer für			
	$\boldsymbol{\mu}$	$\boldsymbol{\mu}^{(c)}$	$\boldsymbol{B}$	$\boldsymbol{\Sigma}$		$\boldsymbol{\mu}$	$\boldsymbol{\mu}^{(c)}$	$\boldsymbol{B}$	$\boldsymbol{\Sigma}$	
GDA	–	$\hat{\boldsymbol{\mu}}^{(c)}$	–	$\hat{\boldsymbol{\Sigma}}$	GDA	–	$\hat{\boldsymbol{\mu}}^{(c)}$	–	$\hat{\boldsymbol{\Sigma}}_{\text{gew}}$	nein
FDA	$\hat{\tilde{\boldsymbol{\mu}}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}$	$\hat{\boldsymbol{\Sigma}}$	FDA	$\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}_{\text{gew}}$	$\hat{\boldsymbol{\Sigma}}$	ja
FDA	$\hat{\tilde{\boldsymbol{\mu}}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}$	$\hat{\boldsymbol{\Sigma}}$	FDA	$\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}_{\text{gew}}$	$\hat{\boldsymbol{\Sigma}}_{\text{gew}}$	nein
FDA	$\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}_{\text{gew}}$	$\hat{\boldsymbol{\Sigma}}$	FDA	$\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}_{\text{gew}}$	$\hat{\boldsymbol{\Sigma}}_{\text{gew}}$	nein
GDA	–	$\hat{\boldsymbol{\mu}}^{(c)}$	–	$\hat{\boldsymbol{\Sigma}}$	FDA	$\hat{\tilde{\boldsymbol{\mu}}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}$	$\hat{\boldsymbol{\Sigma}}$	ja
GDA	–	$\hat{\boldsymbol{\mu}}^{(c)}$	–	$\hat{\boldsymbol{\Sigma}}$	FDA	$\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}_{\text{gew}}$	$\hat{\boldsymbol{\Sigma}}$	ja
GDA	–	$\hat{\boldsymbol{\mu}}^{(c)}$	–	$\hat{\boldsymbol{\Sigma}}$	FDA	$\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}_{\text{gew}}$	$\hat{\boldsymbol{\Sigma}}_{\text{gew}}$	nein
GDA	–	$\hat{\boldsymbol{\mu}}^{(c)}$	–	$\hat{\boldsymbol{\Sigma}}_{\text{gew}}$	FDA	$\hat{\tilde{\boldsymbol{\mu}}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}$	$\hat{\boldsymbol{\Sigma}}$	nein
GDA	–	$\hat{\boldsymbol{\mu}}^{(c)}$	–	$\hat{\boldsymbol{\Sigma}}_{\text{gew}}$	FDA	$\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}_{\text{gew}}$	$\hat{\boldsymbol{\Sigma}}$	nein
GDA	–	$\hat{\boldsymbol{\mu}}^{(c)}$	–	$\hat{\boldsymbol{\Sigma}}_{\text{gew}}$	FDA	$\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$	$\hat{\boldsymbol{\mu}}^{(c)}$	$\hat{\boldsymbol{B}}_{\text{gew}}$	$\hat{\boldsymbol{\Sigma}}_{\text{gew}}$	ja

* Alle Methoden liefern identische Klassifikationsergebnisse für die betrachteten Parameterkonstellationen, wenn gleich große Klassen $n^{(c)}$ und identische (bekannte) a-priori Wahrscheinlichkeiten $p^{(c)}$ betrachtet werden oder bei gleich großen Klassen $n^{(c)}$ und Verwendung der relativen Häufigkeiten als Schätzer oder Betrachtung einer Gleichverteilung $\hat{p}^{(c)} = \frac{n^{(c)}}{n} = \frac{1}{n/n^{(c)}} = \frac{1}{M}$.

so ändert sich ebenfalls die „Steigung“, da $\hat{\boldsymbol{\Sigma}}_{\text{gew}}$ von $\hat{p}^{(c)}$ beeinflusst wird und somit in die Steigung von (3.58) einfließt.

Für weitere Ergebnisse im Detail sei auf die Simulationsstudie der Bachelorarbeit von van Meegen (2015) bzw. für eine kürzere Variante auf die Veröffentlichung von van Meegen et al. (2019) verwiesen.

4 Methoden für Online LDA und QDA

In diesem Kapitel werden die verschiedenen Methoden für Online Diskriminanzanalyse bzw. Adaptionen der Linearen und Quadratischen Diskriminanzanalyse für Datenströme vorgestellt und erläutert. Diese dienen als Ausgangsmethoden, welche in Kapitel 7 in Hinblick auf eine Verbesserung der Prognosegüte bei Vorliegen von concept drift erweitert werden. Im folgenden Abschnitt wird zunächst die allgemeine Ausgangssituation beschrieben, die allen folgenden Methoden zugrunde liegt.

4.1 Ausgangssituation

Es wird ein Datenstrom $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ von Beobachtungen betrachtet, wobei $\boldsymbol{x}_i \in \mathbb{R}^p$ für $i = 1, 2, \ldots$. Für die sequentiellen Verfahren wird dabei zunächst angenommen, dass zu jedem Zeitpunkt i genau eine einzelne neue Beobachtung hinzukommt. Die jeweiligen Klassenlabels c_i der Beobachtungen liegen immer jeweils einen Zeitpunkt zeitversetzt vor, für eine Beobachtung $\boldsymbol{x}_t$ zu einem festen Zeitpunkt t also zum Zeitpunkt $t+1$. Die Zeitpunkte $i = 1, 2, \ldots$ sind dabei äquidistant.

Da bei Datenströmen die Annahme (3.1) einer unveränderten Verteilung $\boldsymbol{X} := \boldsymbol{X}_i$ über die Zeit häufig nicht gerechtfertigt ist, werden nun anders als in der Batch Variante der LDA (vgl. Kapitel 3) die Zufallsvektoren $\boldsymbol{X}_i$ mit Zeitindex i betrachtet. Die Dichtefunktion der klassenbedingten Verteilung (vgl. (3.3)) $\boldsymbol{X}_i^{(c)} := \boldsymbol{X}_i | (W^{(c)} = n_i^{(c)},\ W^{(j)} = 0,\ j \neq c) \sim \mathcal{N}\left(\boldsymbol{\mu}_i^{(c)}, \boldsymbol{\Sigma}_i^{(c)}\right)$ der Zufallsvektoren sei daher mit $f_i^{(c)}(\boldsymbol{x}; \boldsymbol{\mu}_i^{(c)}, \boldsymbol{\Sigma}_i^{(c)})$ bezeichnet.

In der LDA wird angenommen, dass die Kovarianzmatrix in allen Klassen identisch ist. Zudem wird eine zeitinvariante Kovarianzmatrix vorausgesetzt. Es gilt demnach zusätzlich:

$$\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_i^{(1)} = \ldots = \boldsymbol{\Sigma}_i^{(M)} \quad \forall i.$$

Zu einem festen Zeitpunkt t folgt die Verteilung der Klassen einer Multinomialverteilung mit Parametern t und den a-priori Wahrscheinlichkeiten $p^{(c)} \in [0, 1]$ für die einzelnen Klassen $c = 1, \ldots, M$ (vgl. (3.2) in Kapitel 3):

$$\boldsymbol{W} = (W^{(1)}, \ldots, W^{(M)}) \sim Mult\left(t, p^{(1)}, \ldots, p^{(M)}\right),$$

wodurch

$$P(W^{(1)} = n^{(1)}, \ldots, W^{(M)} = n^{(M)}) = \frac{t!}{n^{(1)}! \cdots n^{(M)}!} \cdot \left(p^{(1)}\right)^{n^{(1)}} \cdots \left(p^{(M)}\right)^{n^{(M)}}$$

die Wahrscheinlichkeit beschreibt, dass zum Zeitpunkt t genau $n^{(1)}$ Beobachtungen aus Klasse 1, $n^{(2)}$ Beobachtungen aus Klasse 2, usw. vorliegen.

Betrachtet wird nun der feste Zeitpunkt t. Bis zu diesem Zeitpunkt treten Beobachtungen in M verschiedenen Klassen auf, d. h. $\boldsymbol{x}_i$, $i = 1, \ldots, t$, hat Klassenlabel $c_i \in \{1, \ldots, M\}$. Insgesamt liegen $n_t^{(c)}$ *Beobachtungen mit Klassenlabel* c vor:

$$n_t^{(c)} = \sum_{i=1}^{t} \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}. \tag{4.1}$$

Die Funktion g beschreibt dabei, welche Klasse auftritt (vgl. (3.5) in Kapitel 3).

Die *Gesamtanzahl der Beobachtungen zum Zeitpunkt* t sei mit n_t bezeichnet, wobei

$$n_t = \sum_{c=1}^{M} n_t^{(c)}.$$

Diese Gesamtanzahl ist $n_t = t$ in der Situation, dass genau eine neue Beobachtung zu jedem Zeitpunkt im Datenstrom auftritt.

$\boldsymbol{m}_{n_t}$ bezeichnet den *Mittelwertvektor zum Zeitpunkt* t basierend auf n_t Beobachtungen:

$$\boldsymbol{m}_{n_t} = \frac{1}{n_t} \sum_{i=1}^{t} \boldsymbol{x}_i. \tag{4.2}$$

$\boldsymbol{m}_{n_t^{(c)}}^{(c)}$ bezeichnet den Maximum-Likelihood-Schätzer (kurz: ML-Schätzer) für den Erwartungswertvektor $\boldsymbol{\mu}_t^{(c)}$ von Klasse c zum Zeitpunkt t basierend auf $n_t^{(c)}$ Beobachtungen und demnach den *Mittelwertvektor von Klasse* c (vgl. (3.25) auf Seite 42):

$$\boldsymbol{m}_{n_t^{(c)}}^{(c)} = \frac{1}{n_t^{(c)}} \underbrace{\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t}} \boldsymbol{x}_i}_{n_t^{(c)} \text{ Summanden}} = \frac{1}{\sum_{i=1}^{t} \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}} \sum_{i=1}^{t} \left(\boldsymbol{x}_i \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right). \tag{4.3}$$

Die *empirische Kovarianzmatrix* als ML-Schätzer für die Kovarianzmatrix in Klasse c zum Zeitpunkt t hat die folgende Form (Hartung und Elpelt, 1999, S. 224/246):

$$\boldsymbol{S}_t^{(c)} = \frac{1}{n_t^{(c)} - 1} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_t^{(c)}}^{(c)}\right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_t^{(c)}}^{(c)}\right)^T =: \frac{1}{n_t^{(c)} - 1} \cdot \boldsymbol{Q}_t^{(c)}. \tag{4.4}$$

Die *gepoolte Kovarianzmatrix innerhalb der Klassen zum Zeitpunkt t* berechnet sich durch (Hastie et al., 2009, S. 109)

$$\boldsymbol{S}_t = \frac{1}{n_t - M} \sum_{c=1}^{M} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}} \right)^T =: \frac{1}{n_t - M} \sum_{c=1}^{M} \boldsymbol{Q}_t^{(c)}, \tag{4.5}$$

was analog ist zu $\boldsymbol{S}_t = \frac{1}{n_t - M} \sum_{c=1}^{M} \left(\left(n_t^{(c)} - 1 \right) \boldsymbol{S}_t^{(c)} \right)$ (vgl. (3.26)/(3.27) auf Seite 42).

Für die Fisher Diskriminanzanalyse (Fisher LDA) wird zusätzlich die *Zwischen-den-Klassen Kovarianzmatrix* $\boldsymbol{B}_t$ (Filzmoser et al., 2006, S. 522) bzw. ein Schätzer benötigt (vgl. gewichtete Variante (3.38)). Dabei wird zunächst auf einen Vorfaktor verzichtet, da sich die Aktualisierungsschritte so deutlich vereinfachen (s. Abschnitt 4.2). Ein Vorfaktor wird erst bei Bildung der Klassifikationsregel wieder betrachtet. Wichtig ist demnach zunächst nur $\tilde{\boldsymbol{B}}_t$ (ohne Vorfaktor):

$$\tilde{\boldsymbol{B}}_t = \sum_{c=1}^{M} n_t^{(c)} \left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_t} \right) \left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_t} \right)^T. \tag{4.6}$$

Es ist zu erwähnen, dass $\tilde{\boldsymbol{B}}_t$ auch bereits ohne einen Vorfaktor der Variante $\hat{\boldsymbol{B}}_{\text{gew},2}$ aus (3.55) in Abschnitt 3.3 (vgl. Seite 53) entspricht, da $\boldsymbol{m}_{n_t} = \hat{\tilde{\boldsymbol{\mu}}}_{\text{gew},2}$ aus (3.51). Die zusätzliche Betrachtung des Vorfaktors $\frac{1}{n_t}$ liefert die Variante $\hat{\boldsymbol{B}}_{\text{gew}}$ aus (3.56). Eine Erklärung folgt auf Seite 65. Auf Seite 53 wurde erläutert, dass die Wahl des Vorfaktors generell bei der Fisher LDA keine Rolle spielt.

Ebenso wird der Vorfaktor der gepoolten Kovarianzmatrix innerhalb der Klassen zum Zeitpunkt t bei der Aktualisierung zunächst nicht benötigt. Daher sei mit $\tilde{\boldsymbol{S}}_t$ nur der Teil der Kovarianzmatrix (4.5) ohne den Vorfaktor definiert:

$$\tilde{\boldsymbol{S}}_t = \sum_{c=1}^{M} \boldsymbol{Q}_t^{(c)} = \sum_{c=1}^{M} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}} \right)^T, \tag{4.7}$$

wobei $\boldsymbol{S}_t = \frac{1}{n_t - M} \cdot \tilde{\boldsymbol{S}}_t$.

Bei den Chunk Varianten der Methoden werden bei der Aktualisierung mehr als eine einzelne neue Beobachtung herangezogen. Es wird angenommen, dass die Beobachtungen weiterhin sequentiell, d. h. eine Beobachtung zu jedem Zeitpunkt, auftreten, ein Update des Klassifikationsmodells aber erst nach einigen Zeitpunkten erfolgt.

Für das Update des Klassifikationsmodells werden daher nicht alle Zeitpunkte i, $i = 1, 2, \ldots$, betrachtet, sondern allgemein die Zeitpunkte $t_0 := 1, \ldots, t_1, \ldots, t_2, \ldots$, wobei $t_{j+1} > t_j$ für alle $j = 0, 1, \ldots$, und beliebig viele Zeitpunkte zwischen t_j und t_{j+1} liegen können.

Die Gesamtanzahl der Beobachtungen zum Zeitpunkt t_1 ist

$$n_{t_1} = \sum_{c=1}^{M} n_{t_1}^{(c)},$$

wobei

$$n_{t_1}^{(c)} = \sum_{i=1}^{t_1} \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}$$

die Anzahl der Beobachtungen in Klasse c ist. n_{t_2} bezeichnet entsprechend die Anzahl an Beobachtungen zum Zeitpunkt t_2 nach Update durch einen Chunk von neuen Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$. Mit $n_{t_1:t_2}$ wird die Anzahl der Beobachtungen aus diesem neuen Chunk benannt, also

$$n_{t_1:t_2} = \sum_{c=1}^{M} n_{t_1:t_2}^{(c)} \quad \text{mit} \quad n_{t_1:t_2}^{(c)} = \sum_{i=t_1+1}^{t_2} \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}.$$

Es gilt dabei: $n_{t_1} + n_{t_1:t_2} = n_{t_2}$ bzw. $n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)} = n_{t_2}^{(c)}$. Mit $T := \{t_1 + 1, ..., t_2\}$ sei im Folgenden die Menge der Zeitpunkte des aktuellen Chunks bezeichnet.

Eine alternative Interpretation der Chunk Variante ist jene, dass zu jedem Zeitpunkt mehrere neue Beobachtungen (ein Chunk) auftreten und weiterhin zu jedem Zeitpunkt ein Update des Klassifikationsmodells erfolgt.

4.2 Sequential Incremental LDA und Chunk Incremental LDA (Pang et al., 2005a,b)

Pang et al. (2005a,b) beschäftigen sich mit der Adaption der **linearen Diskriminanzanalyse nach Fisher** (Fisher LDA, Abschnitt 3.3) (Fisher, 1936) für Datenströme. Sie ziehen dazu Standard Update-Formeln für Mittelwertvektoren und Kovarianzmatrizen (Zwischenden-Klassen und Innerhalb-der-Klassen) heran, um diese für die Fisher LDA schrittweise zu aktualisieren. Unterschieden wird dabei zwischen der sogenannten *Sequential Incremental LDA* (Pang et al., 2005b) und der *Chunk Incremental LDA* (Pang et al., 2005a,b).

Im Folgenden werden Korrekturen von Notationsfehlern in Pang et al. (2005b) und Erweiterungen mit (∗∗) gekennzeichnet. Größere Abweichungen werden im Abschnitt „Anmerkungen“ (s. Seite 68 f.) erläutert.

Sequential Incremental LDA (kurz: Sequential ILDA)

Die Autoren entwickeln zunächst eine Update-Methode durch Aktualisierung der benötigten Mittelwertvektoren und Kovarianzmatrizen durch eine einzelne neue Beobachtung.

Je nachdem, welche Klassenausprägung die neue Beobachtung $\boldsymbol{x}_{t+1}$ im Datenstrom hat, werden die Mittelwertvektoren aller Klassen folgendermaßen aktualisiert:

$$\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}} = \begin{cases} \boldsymbol{m}^{(c)}_{n^{(c)}_t}, & \text{falls } g(\boldsymbol{x}_{t+1}) \neq c, \\ \dfrac{n^{(c)}_t \boldsymbol{m}^{(c)}_{n^{(c)}_t} + \boldsymbol{x}_{t+1}}{n^{(c)}_t + 1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c \in \{1, \ldots, M\}, \\ \boldsymbol{x}_{t+1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c = M+1. \end{cases} \tag{4.8}$$

Für den neuen Gesamtmittelwertvektor gilt:

$$\boldsymbol{m}_{n_{t+1}} = \frac{n_t \boldsymbol{m}_{n_t} + \boldsymbol{x}_{t+1}}{n_t + 1}. \tag{4.9}$$

Für die Aktualisierung der Kovarianzmatrizen werden zusätzlich zu den einzelnen Mittelwertvektoren auch die neuen Anzahlen der Beobachtungen in jeder Klasse c benötigt. Diese ergeben sich schrittweise folgendermaßen:

$$n^{(c)}_{t+1} = \begin{cases} n^{(c)}_t, & \text{falls } g(\boldsymbol{x}_{t+1}) \neq c, \\ n^{(c)}_t + 1, & \text{falls } g(\boldsymbol{x}_{t+1}) = c \in \{1, \ldots, M\}, \\ 1, & \text{falls } g(\boldsymbol{x}_{t+1}) = c = M+1. \end{cases} \tag{4.10}$$

Die aktuelle Gesamtanzahl an Beobachtungen n_{t+1} wird nicht explizit als Aktualisierungsschritt aufgeführt, ist jedoch $n_{t+1} = \sum_{c=1}^{M} n^{(c)}_{t+1}$ oder einfacher $n_{t+1} = n_t + 1$ bzw. $n_{t+1} = t + 1$ bei schrittweiser Aktualisierung durch eine einzelne Beobachtung zu jedem Zeitpunkt.

Aus diesen Anzahlen an Beobachtungen lassen sich auch Schätzer für die a-priori Wahrscheinlichkeiten in Form von aktualisierten relativen Häufigkeiten herleiten. Die a-priori Wahrscheinlichkeiten werden zwar von Pang et al. (2005b) nicht erwähnt, Schätzer für diese Wahrscheinlichkeiten werden jedoch für eine spezielle Klassifikationsregel (3.47) bei der Fisher LDA benötigt. Daher wird diese Formel hier ergänzend hergeleitet (∗∗):

$$P^{(c)}_{t+1} = \frac{n^{(c)}_{t+1}}{n_{t+1}} = \begin{cases} \dfrac{n^{(c)}_t}{n_t + 1}, & \text{falls } g(\boldsymbol{x}_{t+1}) \neq c, \\ \left(1 - \dfrac{1}{n_t + 1}\right) P^{(c)}_t + \dfrac{1}{n_t + 1} = \dfrac{n^{(c)}_t + 1}{n_t + 1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c \in \{1, \ldots, M\}, \\ \dfrac{1}{n_t + 1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c = M+1. \end{cases} \tag{4.11}$$

Eine Initialisierung kann dabei auf Basis der relativen Häufigkeiten an Beobachtungen in den Klassen $c = 1, \ldots, M$ auf den ersten Beobachtungen im Datenstrom erfolgen.

Die Betrachtung einer zusätzlichen Beobachtung $\boldsymbol{x}_{t+1}$ bei der Bildung einer Klassifikationsregel durch die Fisher LDA zieht zudem veränderte Kovarianzmatrizen (Zwischen-den-Klassen und Innerhalb-der-Klassen) nach sich. Diese neuen $\tilde{\boldsymbol{B}}_{t+1}$ und $\tilde{\boldsymbol{S}}_{t+1}$ lassen mithilfe

der aktualisierten Mittelwertvektoren (4.8) und (4.9) und Anzahlen an Beobachtungen (4.10) folgendermaßen herleiten (∗∗):

$$\tilde{\boldsymbol{B}}_{t+1} = \begin{cases} \sum_{c=1}^{M} n_{t+1}^{(c)} \left(\boldsymbol{m}_{n_{t+1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t+1}} \right) \left(\boldsymbol{m}_{n_{t+1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t+1}} \right)^T, & \text{falls} \\ & g(\boldsymbol{x}_{t+1}) = k \in \{1, \ldots, M\}, \\ \sum_{c=1}^{M} n_t^{(c)} \left(\boldsymbol{m}_{n_t^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t+1}} \right) \left(\boldsymbol{m}_{n_t^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t+1}} \right)^T & \\ \quad + \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_{t+1}} \right) \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_{t+1}} \right)^T & \\ = \sum_{c=1}^{M+1} n_{t+1}^{(c)} \left(\boldsymbol{m}_{n_{t+1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t+1}} \right) \left(\boldsymbol{m}_{n_{t+1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t+1}} \right)^T, & \text{falls } g(\boldsymbol{x}_{t+1}) = M+1, \end{cases} \tag{4.12}$$

$$\tilde{\boldsymbol{S}}_{t+1} = \begin{cases} \sum_{c=1, c\neq k}^{M} \boldsymbol{Q}_t^{(c)} + \boldsymbol{Q}_{t+1}^{(k)}, & \text{falls } g(\boldsymbol{x}_{t+1}) = k \in \{1, \ldots, M\}, \\ \sum_{c=1}^{M} \boldsymbol{Q}_t^{(c)} + \underbrace{\boldsymbol{Q}_{t+1}^{(M+1)}}_{=\boldsymbol{0}} = \sum_{c=1}^{M} \boldsymbol{Q}_t^{(c)} = \tilde{\boldsymbol{S}}_t, & \text{falls } g(\boldsymbol{x}_{t+1}) = M+1, \end{cases}$$

wobei

$$\boldsymbol{Q}_t^{(c)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_t^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_t^{(c)}}^{(c)} \right)^T \text{ wie in (4.4)}$$

und

$$\boldsymbol{Q}_{t+1}^{(k)} = \boldsymbol{Q}_t^{(k)} + \frac{n_t^{(k)}}{n_t^{(k)} + 1} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_t^{(k)}}^{(k)} \right) \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_t^{(k)}}^{(k)} \right)^T.$$

Damit gilt für die aktualisierte gepoolte Kovarianzmatrix innerhalb der Klassen vereinfacht lediglich auf Basis der aktuellen Schätzer und der neuen Beobachtung $\boldsymbol{x}_{t+1}$:

$$\tilde{\boldsymbol{S}}_{t+1} = \begin{cases} \tilde{\boldsymbol{S}}_t + \frac{n_t^{(k)}}{n_t^{(k)}+1} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_t^{(k)}}^{(k)} \right) \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_t^{(k)}}^{(k)} \right)^T, & \text{falls} \\ & g(\boldsymbol{x}_{t+1}) = k \in \{1, \ldots, M\}, \\ \tilde{\boldsymbol{S}}_t, & \text{falls } g(\boldsymbol{x}_{t+1}) = M+1, \end{cases} \tag{4.13}$$

da $\sum_{c=1, c\neq k}^{M} \boldsymbol{Q}_t^{(c)} + \boldsymbol{Q}_{t+1}^{(k)} = \underbrace{\sum_{c=1, c\neq k}^{M} \boldsymbol{Q}_t^{(c)} + \boldsymbol{Q}_t^{(k)}}_{=\tilde{\boldsymbol{S}}_t \text{ nach (4.7)}} + \frac{n_t^{(k)}}{n_t^{(k)}+1} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_t^{(k)}}^{(k)} \right) \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_t^{(k)}}^{(k)} \right)^T.$

Streng genommen handelt es sich bei der Formel (4.12) der Zwischen-den-Klassen Kovarianzmatrix um keine Aktualisierung, da die Berechnung nicht auf dem bisherigen Schätzer $\tilde{\boldsymbol{B}}_t$ basiert. Vielmehr wird die Kovarianzmatrix in jedem Schritt neu berechnet. Dabei hängt sie jedoch nicht von allen alten Beobachtungen des Datenstroms ab, sondern nur von den aktuellen Mittelwertvektoren und der neuen Beobachtung $\boldsymbol{x}_{t+1}$.

Mit den Mittelwertvektoren als Schätzer für die Erwartungswertvektoren sowie Kovarianzmatrizen und Schätzer für die a-priori Wahrscheinlichkeiten liegen alle benötigten Größen für die Klassifikationsregel (3.45) bzw. (3.47) der Fisher LDA vor.

Bei der Bestimmung der Diskriminanzkomponenten durch das Eigenwertproblem in der Fisher LDA werden die Schätzer $\tilde{\boldsymbol{S}}_t$ und $\tilde{\boldsymbol{B}}_t$ mit aktuellen Vorfaktoren betrachtet. In der Implementierung dieser Arbeit wird der Vorfaktor $\frac{1}{n_t}$ bei der Zwischen-den-Klassen Kovarianzmatrix $\tilde{\boldsymbol{B}}_t$ betrachtet. Da die einzelnen Summanden in (4.12) jeweils mit der Anzahl der Beobachtungen in der Klasse $n_t^{(c)}$ gewichtet werden, führt die Multiplikation mit dem Vorfaktor $\frac{1}{n_t}$ zu der Batch Variante

$$\frac{1}{n_t}\sum_{c=1}^{M} n_t^{(c)} \left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_t}\right)\left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_t}\right)^T = \sum_{c=1}^{M} \frac{n_t^{(c)}}{n_t} \left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_t}\right)\left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_t}\right)^T. \tag{4.14}$$

Dieser Schätzer ist identisch mit der von van Meegen et al. (2019, S. 6) vorgeschlagenen Version (3.56), wenn die a-priori Wahrscheinlichkeiten der Klassen durch die relativen Häufigkeiten $\hat{p}_t^{(c)} = \frac{n_t^{(c)}}{n_t}$ geschätzt werden, da dann auch $\boldsymbol{m}_{n_t}$ aus (4.2) identisch ist zu $\hat{\tilde{\boldsymbol{\mu}}}_{\text{gew}}$ aus (3.50).

Ein alternativer Vergleich ist jener mit dem Schätzer (3.53). Da die Eigenvektoren im Eigenwertproblem der Fisher LDA von dem Vorfaktor unberührt bleiben und sich lediglich die Eigenwerte um den Faktor entsprechend verändern (vgl. Seite 53), ist es unerheblich, ob $\frac{1}{n_t}$ oder wie in (3.53) $\frac{1}{n_t - M}$ betrachtet wird. Bei einem großen Datenstrom und demnach großen n_t spielt zudem die Anzahl der Klassen M in der Differenz $n_t - M$ für $M << n_t$ kaum noch eine Rolle.

Als Schätzer für die gepoolte Kovarianzmatrix innerhalb der Klassen wird die von Hastie et al. (2009, S. 109) vorgeschlagene ungewichtete gepoolte Kovarianzmatrix (3.27) herangezogen, sodass $\tilde{\boldsymbol{S}}_t$ mit dem Vorfaktor $\frac{1}{n_t - M}$ betrachtet wird, was zu folgender Batch Variante des Schätzers führt:

$$\frac{1}{n_t - M}\sum_{c=1}^{M} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}}\right)\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}}\right)^T.$$

Chunk Incremental LDA

In einem weiteren Schritt führen Pang et al. (2005a,b) eine Möglichkeit zur Aktualisierung der Fisher LDA auf Basis einer ganzen Reihe von neuen Beobachtungen (Chunk) ein. Die Anzahl der neuen Beobachtungen, die für eine Aktualisierung herangezogen werden, können dabei für jeden Chunk zufällig sein. Im Folgenden sei mit $T := \{t_1 + 1, \ldots, t_2\}$ die Menge der Zeitpunkte bzw. Beobachtungen des aktuellen Chunks bezeichnet (vgl. Seite 62).

Für den gesamten Mittelwertvektor gilt zum Zeitpunkt t_2 durch Hinzunahme von $n_{t_1:t_2}$ neuen Beobachtungen:

$$\boldsymbol{m}_{n_{t_2}} = \frac{n_{t_1}\boldsymbol{m}_{n_{t_1}} + n_{t_1:t_2}\boldsymbol{m}_{n_{t_1:t_2}}}{n_{t_1} + n_{t_1:t_2}} = \frac{n_{t_1}\boldsymbol{m}_{n_{t_1}} + n_{t_1:t_2}\boldsymbol{m}_{n_{t_1:t_2}}}{n_{t_2}},$$

wobei

$$\boldsymbol{m}_{n_{t_1:t_2}} = \frac{1}{n_{t_1:t_2}} \sum_{i=t_1+1}^{t_2} \boldsymbol{x}_i.$$

Die Mittelwertvektoren der einzelnen Klassen c können analog auf Basis der alten Mittelwertvektoren durch Hinzunahme der neuen Mittelwertvektoren $\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}$ aktualisiert werden, wenn Beobachtungen in bereits bekannten Klassen auftreten:

$$\boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} = \begin{cases} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}}, & \text{falls } \forall\, i \in T : g(\boldsymbol{x}_i) \neq c, \\ \dfrac{n^{(c)}_{t_1}\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2}\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}}{n^{(c)}_{t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \in \{1,\ldots,M\}, \\ \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \notin \{1,\ldots,M\}, \end{cases} \tag{4.15}$$

wobei

$$\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} = \frac{1}{n^{(c)}_{t_1:t_2}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ t_1<i\leq t_2}} \boldsymbol{x}_i = \frac{1}{\sum_{i=t_1+1}^{t_2} \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}} \sum_{i=t_1+1}^{t_2} \left(\boldsymbol{x}_i \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right). \tag{4.16}$$

Die neuen Anzahlen an Beobachtungen in jeder Klasse c berechnen sich durch:

$$n^{(c)}_{t_2} = \begin{cases} n^{(c)}_{t_1}, & \text{falls } \forall\, i \in T : g(\boldsymbol{x}_i) \neq c, \\ n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \in \{1,\ldots,M\}, \\ n^{(c)}_{t_1:t_2}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \notin \{1,\ldots,M\}. \end{cases} \tag{4.17}$$

Auch hier werden ergänzend zu Pang et al. (2005a,b) Update-Formeln für die relativen Häufigkeiten als Schätzer für die a-priori Wahrscheinlichkeiten $p^{(c)}$ der einzelnen Klassen hergeleitet, da diese für die Klassifikationsregel (3.47) bei der Fisher LDA benötigt werden (**):

$$P^{(c)}_{t_2} = \frac{n^{(c)}_{t_2}}{n_{t_2}} = \begin{cases} \dfrac{n^{(c)}_{t_1}}{n_{t_1} + n_{t_1:t_2}}, & \text{falls } \forall\, i \in T : g(\boldsymbol{x}_i) \neq c, \\ \dfrac{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}}{n_{t_1} + n_{t_1:t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \in \{1,\ldots,M\}, \\ \dfrac{n^{(c)}_{t_1:t_2}}{n_{t_1} + n_{t_1:t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \notin \{1,\ldots,M\}. \end{cases} \tag{4.18}$$

Zusätzlich zur Anzahl der Beobachtungen in jeder Klasse wird hier die Größe $n_{t_1:t_2}$ des aktuellen Chunks benötigt.

Mithilfe von (4.15)–(4.17) verändern sich die Kovarianzmatrizen zum Zeitpunkt t_2 folgendermaßen (∗∗):

$$\tilde{\boldsymbol{B}}_{t_2} = \begin{cases} \sum_{c=1}^{M} n_{t_2}^{(c)} \left(\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)\left(\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)^T, & \text{falls } \forall\, i \in T: \\ & g(\boldsymbol{x}_i) = c \in \{1,\dots,M\}, \\ \sum_{c=1}^{M} n_{t_2}^{(c)} \left(\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)\left(\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)^T & \\ \quad + n_{t_1:t_2}^{(M+1)} \left(\boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)} - \boldsymbol{m}_{n_{t_2}}\right)\left(\boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)} - \boldsymbol{m}_{n_{t_2}}\right)^T & \\ = \sum_{c=1}^{M+1} n_{t_2}^{(c)} \left(\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)\left(\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)^T, & \text{falls } n_{t_1:t_2}^{(M+1)} \text{ Elemente aus} \\ & T \text{ existieren mit} \\ & g(\boldsymbol{x}_i) = M+1, \end{cases} \tag{4.19}$$

$$\tilde{\boldsymbol{S}}_{t_2} = \begin{cases} \sum_{c=1}^{M} \boldsymbol{Q}_{t_2}^{(c)}, & \text{falls } \forall\, i \in T: \\ & g(\boldsymbol{x}_i) = c \in \{1,\dots,M\}, \\ \sum_{c=1}^{M} \boldsymbol{Q}_{t_2}^{(c)} + \boldsymbol{Q}_{t_1:t_2}^{(M+1)} = \sum_{c=1}^{M} \boldsymbol{Q}_{t_2}^{(c)} + \boldsymbol{Q}_{t_2}^{(M+1)} = \sum_{c=1}^{M+1} \boldsymbol{Q}_{t_2}^{(c)}, & \text{falls } n_{t_1:t_2}^{(M+1)} \text{ Elemente aus} \\ & T \text{ existieren mit} \\ & g(\boldsymbol{x}_i) = M+1, \end{cases} \tag{4.20}$$

wobei

$$\boldsymbol{Q}_{t_1:t_2}^{(M+1)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=M+1 \\ t_1 < i \leq t_2}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)}\right)\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)}\right)^T \tag{4.21}$$

und

$$\begin{aligned} \boldsymbol{Q}_{t_2}^{(c)} &= \boldsymbol{Q}_{t_1}^{(c)} + \frac{n_{t_1}^{(c)} \left(n_{t_1:t_2}^{(c)}\right)^2}{\left(n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)}\right)^2} \cdot \boldsymbol{D}^{(c)} + \frac{\left(n_{t_1}^{(c)}\right)^2}{\left(n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)}\right)^2} \cdot \boldsymbol{E}^{(c)} \\ &\quad + \frac{n_{t_1:t_2}^{(c)} \left(n_{t_1:t_2}^{(c)} + 2n_{t_1}^{(c)}\right)}{\left(n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)}\right)^2} \cdot \boldsymbol{F}^{(c)} \\ &= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t_1}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)}\right)^T + \frac{n_{t_1}^{(c)} \left(n_{t_1:t_2}^{(c)}\right)^2}{\left(n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)}\right)^2} \cdot \boldsymbol{D}^{(c)} \\ &\quad + \frac{\left(n_{t_1}^{(c)}\right)^2}{\left(n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)}\right)^2} \cdot \boldsymbol{E}^{(c)} + \frac{n_{t_1:t_2}^{(c)} \left(n_{t_1:t_2}^{(c)} + 2n_{t_1}^{(c)}\right)}{\left(n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)}\right)^2} \cdot \boldsymbol{F}^{(c)} \end{aligned}$$

mit

$$\boldsymbol{D}^{(c)} = \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}}\right)\left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}}\right)^T,$$

$$\boldsymbol{E}^{(c)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ t_1<i\leq t_2}} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}}\right)\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}}\right)^T,$$

$$\boldsymbol{F}^{(c)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ t_1<i\leq t_2}} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}\right)\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}\right)^T \quad (**).$$

Eine alternative Sichtweise dieses Modells ist der Zusammenschluss von zwei bestehenden Modellen auf verschiedenen Beobachtungen anstelle des Updates eines alten Modells durch einen Chunk von neuen Beobachtungen. In diesem Fall sind die interessierenden Größen des ersten Modells n_{t_1}, $n^{(c)}_{t_1}$, $\boldsymbol{m}_{n_{t_1}}$, $\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}}$, $c = 1, \ldots, M$, $\tilde{\boldsymbol{B}}_{t_1}$ und $\tilde{\boldsymbol{S}}_{t_1}$ basierend auf den Beobachtungen der Zeitpunkte 1 bis t_1. Das zweite Fisher Diskriminanzanalysemodell besteht aus den Größen $n_{t_1:t_2}$, $n^{(c)}_{t_1:t_2}$, $\boldsymbol{m}_{n_{t_1:t_2}}$, $\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}$, $c = 1, \ldots, M(+1)$, $\tilde{\boldsymbol{B}}_{t_1:t_2}$ und $\tilde{\boldsymbol{S}}_{t_1:t_2}$ basierend auf den Beobachtungen der Zeitpunkte $t_1 + 1$ bis t_2. Bei den beiden entwickelten Update-Methoden ist jeweils das Auftreten einer neuen Klasse $M + 1$ im Datenstrom möglich.

Letztendlich ist das resultierende Modell aus den inkrementellen Anpassungen der Sequential Incremental LDA oder Chunk Incremental LDA dasselbe, welches bei Anpassung an alle Beobachtungen gemeinsam resultieren würde (also in der Batch-Methode der linearen Diskriminanzanalyse nach Fisher aus Abschnitt 3.3). Folglich erfolgt keine Anpassung an einen eventuell vorliegenden concept drift im Datenstrom. Es werden keine Gewichte betrachtet, sondern alle Beobachtungen fließen gleichermaßen in das Modell ein.

Anmerkungen Die aktualisierte Zwischen-den-Klassen Kovarianzmatrix zum Zeitpunkt $t + 1$ wurde für den Fall $g(\boldsymbol{x}_{t+1}) = M + 1$ in (4.12) korrigiert. Pang et al. (2005b, S. 906) definieren diese als

$$\begin{aligned}\tilde{\boldsymbol{B}}_{t+1} &= \sum_{c=1}^{M} n_t^{(c)} \left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_{t+1}}\right)\left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_{t+1}}\right)^T + \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_{t+1}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_{t+1}}\right)^T \\ &= \sum_{c=1}^{M+1} n_{t+1}^{(c)} \left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_{t+1}}\right)\left(\boldsymbol{m}^{(c)}_{n_t^{(c)}} - \boldsymbol{m}_{n_{t+1}}\right)^T.\end{aligned}$$

Die Autoren sagen zwar, dass $\boldsymbol{m}^{(M+1)}_{n_t^{(M+1)}} = \boldsymbol{x}_{t+1}$, aber streng genommen existiert $\boldsymbol{m}^{(M+1)}_{n_t^{(M+1)}}$ zum Zeitpunkt t noch nicht. Da nach (4.8) $\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}} = \boldsymbol{m}^{(c)}_{n_t^{(c)}}$ für $c = 1, \ldots, M$ in diesem Fall, können stattdessen in der Endformel die aktuellen Mittelwertvektoren $\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}$ betrachtet werden, was zur korrigierten Variante (4.12) führt.

Bei der Chunk Methode wurde die aktualisierte Zwischen-den-Klassen Kovarianzmatrix zum Zeitpunkt t_2 durch (4.19) korrigiert. In Pang et al. (2005b, S. 907) wird der Teil, falls $n_{t_1:t_2}^{(M+1)}$ Elemente aus T existieren mit $g(\boldsymbol{x}_i) = M+1$ definiert als:

$$\begin{aligned}\tilde{\boldsymbol{B}}_{t_2} &= \sum_{c=1}^{M} n_{t_2}^{(c)} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)\left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)^T \\ &\quad + n_{t_1:t_2}^{(M+1)} \left(\boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)} - \boldsymbol{m}_{n_{t_2}}\right)\left(\boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)} - \boldsymbol{m}_{n_{t_2}}\right)^T \\ &= \sum_{c=1}^{M+1} n_{t_2}^{(c)} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)\left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_2}}\right)^T.\end{aligned}$$

Es müssen jedoch die aktuellen Mittelwertvektoren $\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)}$ der Klassen in der Kovarianzmatrix betrachtet werden, da auch Beobachtungen der übrigen Klassen $c = 1, \ldots, M$ im neuen Chunk auftreten können.

Ebenso wird in Pang et al. (2005b, S. 907) der falsche Index t_1 statt t_2 im entsprechenden Fall bei der aktualisierten gepoolten Kovarianzmatrix innerhalb der Klassen (4.20) verwendet:

$$\tilde{\boldsymbol{S}}_{t_2} = \sum_{c=1}^{M} \boldsymbol{Q}_{t_1}^{(c)} + \boldsymbol{Q}_{t_1:t_2}^{(M+1)} = \sum_{c=1}^{M+1} \boldsymbol{Q}_{t_2}^{(c)}$$

und

$$\boldsymbol{Q}_{t_1:t_2}^{(M+1)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ t_1 < i \leq t_2}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}\right)^T$$

anstelle von (4.21).

Die korrigierte Version stellen Formeln (4.20) und (4.21) dar.

Zur Veranschaulichung werden alle Methoden im Anhang auf einen einfachen eindimensionalen Beispiel-Datenstrom (vgl. Tabelle A.1 in Anhang A) angewandt. Das Rechenbeispiel für die Methode *Sequential Incremental LDA* ist in Anhang A.1 zu finden.

4.3 Online Linear Discriminant Classifier (kurz: OLDC): Feste und adaptive Lernrate (Kuncheva und Plumpton, 2008)

Kuncheva und Plumpton (2008) betrachten die **Kanonische Lineare Diskriminanzanalyse**. Es werden ebenfalls Update-Formeln herangezogen, um die benötigten Schätzer der a-priori Wahrscheinlichkeiten, die Mittelwertvektoren der Klassen und die inverse Kovarianzmatrix bei Auftreten neuer Beobachtungen inkrementell zu aktualisieren und somit eine adaptive Online Variante der Methode zu entwickeln. Dabei erfolgen die Aktualisierungen immer auf Basis einer einzelnen neuen Beobachtung. Die folgenden Formeln und

Herleitungen sind Kuncheva und Plumpton (2008) entnommen und auf die Notation dieser Dissertation angepasst bzw. teilweise ergänzt und ausführlicher hergeleitet. An einigen Stellen wurden Korrekturen und Erweiterungen der Formeln von Kuncheva und Plumpton (2008) vorgenommen. Die entsprechenden Stellen sind mit (∗∗) markiert und die wichtigsten Korrekturen sind im Abschnitt „Anmerkungen“ (s. Seite 75 f.) erläutert.

Update ohne Lernrate

Der aktualisierte Mittelwertvektor $\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}$ der einzelnen Klassen c ergibt sich analog zu (4.8) der Sequential Incremental LDA.

Die a-priori Wahrscheinlichkeiten bzw. entsprechend relativen Häufigkeiten als Schätzer ergeben sich zum Zeitpunkt $t+1$ bei Auftreten einer neuen Beobachtung $\boldsymbol{x}_{t+1}$ durch:

$$P^{(c)}_{t+1} = \begin{cases} \dfrac{n^{(c)}_t}{n_t+1}, & \text{falls } g(\boldsymbol{x}_{t+1}) \neq c, \\ \left(1-\dfrac{1}{n_t+1}\right) P^{(c)}_t + \dfrac{1}{n_t+1} = \dfrac{n^{(c)}_t+1}{n_t+1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c \in \{1,\dots,M\}, \\ \dfrac{1}{n_t+1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c = M+1. \end{cases} \tag{4.22}$$

Die Kovarianzmatrix $\boldsymbol{S}_{t+1}$ zum Zeitpunkt $t+1$ bei Update durch eine einzelne neue Beobachtung lässt sich folgendermaßen berechnen, falls $g(\boldsymbol{x}_{t+1}) = k \in \{1,\dots,M\}$ (∗∗):

$$\begin{aligned} \boldsymbol{S}_{t+1} &= \frac{1}{n_t+1} \sum_{c=1}^{M} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t+1}} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}\right)\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}\right)^T \\ &= \frac{n_t}{n_t+1}\left(\boldsymbol{S}_t + \frac{n^{(k)}_t}{n_t\left(n^{(k)}_t+1\right)}\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n^{(k)}_t}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n^{(k)}_t}\right)^T\right). \end{aligned}$$

Die Autorinnen betrachten nur den Vorfaktor $\frac{1}{n_t}$ bei der Bestimmung der gepoolten Kovarianzmatrix anstelle des von u. a. Hastie et al. (2009, S. 109) herangezogenen Vorfaktors $n_t - M$, welcher die Anzahl der Klassen berücksichtigt (vgl. (3.27)). Es ist jedoch zu betonen, dass im Datenstrom der Unterschied zwischen n_t und $n_t - M$ immer unwichtiger wird, da im Laufe der Zeit $n_t >> M$.

Insgesamt gilt (ausführliche Herleitung (B.1) in Anhang B.1):

$$\boldsymbol{S}_{t+1} = \begin{cases} \dfrac{n_t}{n_t+1}\left(\boldsymbol{S}_t + \dfrac{n^{(k)}_t\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n^{(k)}_t}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n^{(k)}_t}\right)^T}{n_t\left(n^{(k)}_t+1\right)}\right), & \text{falls } g(\boldsymbol{x}_{t+1}) = k \in \{1,\dots,M\}, \\ \dfrac{n_t}{n_t+1} \cdot \boldsymbol{S}_t, & \text{falls } g(\boldsymbol{x}_{t+1}) = M+1. \end{cases} \tag{4.23}$$

Die neue Kovarianzmatrix zum Zeitpunkt $t+1$ lässt sich demnach aus der alten Kovarianzmatrix $\boldsymbol{S}_t$ herleiten. Sei $\boldsymbol{z} := \boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}$ $(**)$, dann:

$$\boldsymbol{S}_{t+1} = \frac{n_t}{n_t+1}\left(\boldsymbol{S}_t + \sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}^T\right). \tag{4.24}$$

Bei der Vorhersage durch die Kanonische Lineare Diskriminanzanalyse wird die inverse Kovarianzmatrix und nicht die Kovarianzmatrix selbst benötigt (vgl. (3.22) und (3.23) in Abschnitt 3.2). Kuncheva und Plumpton (2008) nutzen die Sherman-Morrison-Woodbury Formel für ein Rang-1-Update, um direkt die inverse Kovarianzmatrix in jedem Schritt zu aktualisieren, wenn eine einzelne neue Beobachtung betrachtet wird. Eine Invertierung in jedem Schritt wird somit vermieden. Die Sherman-Morrison-Woodbury Formel für ein Rang-d-Update sieht folgendermaßen aus (Golub und Van Loan, 1996, S. 50):

Satz 2. *Sherman-Morrison-Woodbury Formel für ein Rang-d-Update*
Für eine reguläre $(p \times p)$-Matrix $\boldsymbol{S}$ und Matrizen $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{p\times d}$ mit $\boldsymbol{I}_{d\times d} + \boldsymbol{V}^T\boldsymbol{S}^{-1}\boldsymbol{U}$ nicht-singulär gilt:

$$\left(\boldsymbol{S}+\boldsymbol{U}\boldsymbol{V}^T\right)^{-1} = \boldsymbol{S}^{-1} - \boldsymbol{S}^{-1}\boldsymbol{U}\left(\boldsymbol{I}_{d\times d} + \boldsymbol{V}^T\boldsymbol{S}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}^T\boldsymbol{S}^{-1}.$$

Satz 2 lässt sich für ein Rang-1-Update, d. h. bei Betrachtung von Vektoren vereinfachen:

Satz 3. *Sherman-Morrison-Woodbury Formel für ein Rang-1-Update*
Für eine reguläre $(p\times p)$-Matrix $\boldsymbol{S}$ und zwei Vektoren $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$, wobei $\boldsymbol{v}^T\boldsymbol{S}^{-1}\boldsymbol{u} \neq -1$, gilt:

$$\left(\boldsymbol{S}+\boldsymbol{u}\boldsymbol{v}^T\right)^{-1} = \boldsymbol{S}^{-1} - \frac{\boldsymbol{S}^{-1}\boldsymbol{u}\boldsymbol{v}^T\boldsymbol{S}^{-1}}{1+\boldsymbol{v}^T\boldsymbol{S}^{-1}\boldsymbol{u}}. \tag{4.25}$$

Die Verwendung dieser Formel auf (4.24) ermöglicht die Aktualisierung der inversen Kovarianzmatrix zum Zeitpunkt $t+1$ auf Basis der vorherigen inversen Kovarianzmatrix. Eine Invertierung muss demnach nur einmal zu Beginn als Initialisierung erfolgen $(**)$:

$$\begin{aligned}\boldsymbol{S}_{t+1}^{-1} &= \frac{n_t+1}{n_t}\left(\boldsymbol{S}_t^{-1} - \frac{\boldsymbol{S}_t^{-1}\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}^T\boldsymbol{S}_t^{-1}}{1+\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}^T\boldsymbol{S}_t^{-1}\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}}\right)\\ &= \frac{n_t+1}{n_t}\left(\boldsymbol{S}_t^{-1} - \frac{\boldsymbol{S}_t^{-1}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{S}_t^{-1}}{\frac{n_t\left(n_t^{(k)}+1\right)}{n_t^{(k)}}+\boldsymbol{z}^T\boldsymbol{S}_t^{-1}\boldsymbol{z}}\right).\end{aligned}$$

Insgesamt (∗∗):

$$\boldsymbol{S}_{t+1}^{-1} = \begin{cases} \dfrac{n_t+1}{n_t}\left(\boldsymbol{S}_t^{-1} - \dfrac{\boldsymbol{S}_t^{-1}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{S}_t^{-1}}{\frac{n_t\left(n_t^{(k)}+1\right)}{n_t^{(k)}} + \boldsymbol{z}^T\boldsymbol{S}_t^{-1}\boldsymbol{z}}\right), & \text{falls } g(\boldsymbol{x}_{t+1}) = k \in \{1,\dots,M\}, \\ \dfrac{n_t+1}{n_t}\cdot\boldsymbol{S}_t^{-1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = M+1. \end{cases} \tag{4.26}$$

Update mit Lernrate

Die Autorinnen entwickeln erweiternd eine Möglichkeit zur Anpassung an concept drift durch Einführung einer Lernrate $\lambda \in (0,1)$ bei den Update-Formeln der einzelnen Größen und somit indirekt einer Gewichtung der neuen Beobachtungen (Kuncheva und Plumpton, 2008, S. 514 ff.).

Die Idee besteht darin, dass die Lernrate folgende Eigenschaften aufweist:

- $\lambda \to 0$: Die neue Beobachtung wird nicht betrachtet (bei $\lambda = 0$), es erfolgt keine Aktualisierung des Modells durch die Beobachtung $\boldsymbol{x}_{t+1}$.
- $\lambda = 1/2$: Keine Gewichtung; alle Formeln sind identisch mit jenen ohne Lernrate: (4.27) = (4.8), (4.28) = (4.22), (4.29) = (4.26).
- $\lambda \to 1$: Die Vergangenheit wird nicht mehr betrachtet (bei $\lambda = 1$), der ganze Fokus liegt auf der neuen Beobachtung $\boldsymbol{x}_{t+1}$.

Bezüglich der Aktualisierung der Mittelwertvektoren der Klassen sowie der relativen Häufigkeiten ist die Erweiterung durch die Lernrate damit recht anschaulich. In der Update-Formel (4.8) wird die neue Beobachtung $\boldsymbol{x}_{t+1}$ mit λ gewichtet, während die (gewichtete) „Summe aller alten Beobachtungen“ $n_t^{(c)}\boldsymbol{m}_{n_t^{(c)}}^{(c)}$ und somit der „alte“ Anteil der Beobachtungen das Gewicht $1-\lambda$ erhält. Zur Normierung wird zusätzlich auch die bisherige Anzahl aller Beobachtungen in Klasse c zum Zeitpunkt $t+1$ durch die gewichtete Variante $(1-\lambda)n_t^{(c)} + \lambda$ ersetzt:

$$\boldsymbol{m}_{n_{t+1}^{(c)}}^{(c)} = \begin{cases} \boldsymbol{m}_{n_t^{(c)}}^{(c)}, & \text{falls } g(\boldsymbol{x}_{t+1}) \neq c \quad (**), \\ \dfrac{(1-\lambda)n_t^{(c)}\boldsymbol{m}_{n_t^{(c)}}^{(c)} + \lambda\boldsymbol{x}_{t+1}}{(1-\lambda)n_t^{(c)} + \lambda}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c \in \{1,\dots,M\}, \\ \boldsymbol{x}_{t+1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c = M+1. \end{cases} \tag{4.27}$$

Eine ähnliche Idee wird auf die Bestimmung der relativen Häufigkeiten zum Zeitpunkt $t+1$ übertragen. Die Anzahl der Beobachtungen in Klasse c zum Zeitpunkt t wird mit

$1-\lambda$ gewichtet, während die gesamte Anzahl zum neuen Zeitpunkt $t+1$ in $(1-\lambda)n_t+\lambda$ gespalten wird:

$$P_{t+1}^{(c)} = \begin{cases} \dfrac{(1-\lambda)n_t^{(c)}}{(1-\lambda)n_t+\lambda}, & \text{falls } g(\boldsymbol{x}_{t+1}) \neq c, \\ \dfrac{(1-\lambda)n_t^{(c)}+\lambda}{(1-\lambda)n_t+\lambda}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c \in \{1,\ldots,M\}, \\ \dfrac{\lambda}{(1-\lambda)n_t+\lambda}, & \text{falls } g(\boldsymbol{x}_{t+1}) = c = M+1 \quad (**). \end{cases} \tag{4.28}$$

Die Erweiterung der Update-Formel für die inverse Kovarianzmatrix ist etwas weniger anschaulich. Die Grundidee ist jedoch dieselbe. Die Herleitung erfolgt anhand der Update-Formel (4.24) für die Kovarianzmatrix $\boldsymbol{S}_{t+1}$ selbst. Kuncheva und Plumpton (2008, S. 514) gewichten die Vergangenheit, also $\boldsymbol{S}_t$, mit $1-\lambda$. Der Term $\boldsymbol{z}$, der die neue Beobachtung enthält, erhält das Gewicht λ. Zusätzlich wird erneut die gesamte Anzahl der Beobachtungen durch die gewichtete Variante $(1-\lambda)n_t+\lambda$ ersetzt:

$$\boldsymbol{S}_{t+1} = \frac{n_t}{(1-\lambda)n_t+\lambda}\left((1-\lambda)\boldsymbol{S}_t + \lambda\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}^T\right).$$

Daraus ergibt sich durch Anwendung der Sherman-Morrison-Woodbury Formel aus Satz 3 die inverse Kovarianzmatrix zum Zeitpunkt $t+1$ $(**)$:

$$\begin{aligned} \boldsymbol{S}_{t+1}^{-1} &= \frac{(1-\lambda)n_t+\lambda}{n_t}\cdot \\ &\quad\left(\frac{1}{(1-\lambda)}\cdot\boldsymbol{S}_t^{-1} - \frac{\frac{1}{(1-\lambda)}\cdot\boldsymbol{S}_t^{-1}\lambda\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}^T\cdot\frac{1}{(1-\lambda)}\cdot\boldsymbol{S}_t^{-1}}{1+\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}^T\cdot\frac{1}{(1-\lambda)}\cdot\boldsymbol{S}_t^{-1}\lambda\sqrt{\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}}\,\boldsymbol{z}}\right) \\ &= \frac{(1-\lambda)n_t+\lambda}{n_t}\left(\frac{1}{(1-\lambda)}\left(\boldsymbol{S}_t^{-1} - \frac{\frac{\lambda}{(1-\lambda)}\cdot\boldsymbol{S}_t^{-1}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{S}_t^{-1}}{\frac{n_t\left(n_t^{(k)}+1\right)}{n_t^{(k)}}+\frac{\lambda}{(1-\lambda)}\cdot\boldsymbol{z}^T\boldsymbol{S}_t^{-1}\boldsymbol{z}}\right)\right) \\ &= \frac{(1-\lambda)n_t+\lambda}{(1-\lambda)n_t}\left(\boldsymbol{S}_t^{-1} - \frac{\boldsymbol{S}_t^{-1}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{S}_t^{-1}}{\frac{(1-\lambda)n_t\left(n_t^{(k)}+1\right)}{\lambda n_t^{(k)}}+\boldsymbol{z}^T\boldsymbol{S}_t^{-1}\boldsymbol{z}}\right). \end{aligned}$$

Damit auch $\boldsymbol{S}_{t+1}^{-1}$ die entsprechenden Werte für $\lambda=1/2$ und die Grenzwerte $\lambda\to 0$ sowie $\lambda\to 1$ aufweist, wird in dieser Arbeit zusätzlich der Term $\boldsymbol{z} := \boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_t^{(k)}}^{(k)}$ durch $\boldsymbol{z}^*$ ersetzt mit:

$$\boldsymbol{z}^* := \boldsymbol{x}_{t+1} - \frac{\left((1-\lambda)n_t^{(k)}+\lambda\right)\boldsymbol{m}_{n_{t+1}^{(k)}}^{(k)} - \lambda\boldsymbol{x}_{t+1}}{(1-\lambda)n_t^{(k)}} \quad (**).$$

Dies ist möglich, da

$$\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}} \overset{(4.27)}{=} \frac{(1-\lambda)n_t^{(k)}\boldsymbol{m}^{(k)}_{n_t^{(k)}} + \lambda \boldsymbol{x}_{t+1}}{(1-\lambda)n_t^{(k)} + \lambda} \iff \boldsymbol{m}^{(k)}_{n_t^{(k)}} = \frac{\left((1-\lambda)n_t^{(k)} + \lambda\right)\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}} - \lambda \boldsymbol{x}_{t+1}}{(1-\lambda)n_t^{(k)}}.$$

Durch all diese Ideen ergibt sich die aktualisierte inverse Kovarianzmatrix mit Lernrate $\lambda \in (0,1)$ schlussendlich durch ($\lambda \neq 0,\ \lambda \neq 1$):

$$\boldsymbol{S}_{t+1}^{-1} = \begin{cases} \dfrac{(1-\lambda)n_t + \lambda}{(1-\lambda)n_t}\left(\boldsymbol{S}_t^{-1} - \dfrac{\boldsymbol{S}_t^{-1}\boldsymbol{z}^*\boldsymbol{z}^{*T}\boldsymbol{S}_t^{-1}}{\frac{(1-\lambda)n_t\left(n_t^{(k)}+1\right)}{\lambda n_t^{(k)}} + \boldsymbol{z}^{*T}\boldsymbol{S}_t^{-1}\boldsymbol{z}^*}\right), & \text{falls } g(\boldsymbol{x}_{t+1}) = k \in \{1,\dots,M\}, \\ \dfrac{(1-\lambda)n_t + \lambda}{(1-\lambda)n_t}\cdot \boldsymbol{S}_t^{-1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = M+1. \end{cases} \tag{4.29}$$

Die Formel (4.29) erfüllt die geforderten Eigenschaften (vgl. Seite 72):

- $$\underbrace{\frac{(1-\lambda)n_t + \lambda}{(1-\lambda)n_t}}_{=1 \text{ (für } \lambda=0)}\left(\boldsymbol{S}_t^{-1} - \underbrace{\frac{\boldsymbol{S}_t^{-1}\boldsymbol{z}^*\boldsymbol{z}^{*T}\boldsymbol{S}_t^{-1}}{\underbrace{\frac{(1-\lambda)n_t\left(n_t^{(k)}+1\right)}{\lambda n_t^{(k)}}}_{\underset{\lambda\to 0}{\longrightarrow}\infty} + \boldsymbol{z}^{*T}\boldsymbol{S}_t^{-1}\boldsymbol{z}^*}}_{\underset{\lambda\to 0}{\longrightarrow} 0}\right) \underset{\lambda\to 0}{\longrightarrow} \boldsymbol{S}_t^{-1},$$
- $$\underbrace{\frac{(1-\lambda)n_t + \lambda}{(1-\lambda)n_t}}_{=\frac{n_t+1}{n_t} \text{ (für } \lambda=1/2)}\left(\boldsymbol{S}_t^{-1} - \frac{\boldsymbol{S}_t^{-1}\boldsymbol{z}^*\boldsymbol{z}^{*T}\boldsymbol{S}_t^{-1}}{\underbrace{\frac{(1-\lambda)n_t\left(n_t^{(k)}+1\right)}{\lambda n_t^{(k)}}}_{=\frac{n_t\left(n_t^{(k)}+1\right)}{n_t^{(k)}} \text{ (für } \lambda=1/2)} + \boldsymbol{z}^{*T}\boldsymbol{S}_t^{-1}\boldsymbol{z}^*}\right) \underset{\lambda=1/2,\ \boldsymbol{z}=\boldsymbol{z}^*}{=} (4.26),$$
- $$\underbrace{\frac{(1-\lambda)n_t + \lambda}{(1-\lambda)n_t}}_{\underset{\lambda\to 1}{\longrightarrow}\infty}\left(\boldsymbol{S}_t^{-1} - \underbrace{\frac{\boldsymbol{S}_t^{-1}\boldsymbol{z}^*\boldsymbol{z}^{*T}\boldsymbol{S}_t^{-1}}{\frac{(1-\lambda)n_t\left(n_t^{(k)}+1\right)}{\lambda n_t^{(k)}} + \boldsymbol{z}^{*T}\boldsymbol{S}_t^{-1}\boldsymbol{z}^*}}_{\text{nicht definiert für } \lambda=1, \text{ da } \frac{\mathbf{0}}{0}\ (*)}\right) \text{ nicht definiert,}$$

$$\text{da } \lim_{\lambda\to 1}\boldsymbol{z}^* = \lim_{\lambda\to 1}\left(\boldsymbol{x}_{t+1} - \frac{\left((1-\lambda)n_t^{(k)} + \lambda\right)\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}} - \lambda\boldsymbol{x}_{t+1}}{(1-\lambda)n_t^{(k)}}\right)$$

$$\overset{\text{l'Hospital}}{=} \lim_{\lambda\to 1}\left(\boldsymbol{x}_{t+1} - \frac{\left(-n_t^{(k)} + 1\right)\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}} - \boldsymbol{x}_{t+1}}{-n_t^{(k)}}\right) = \mathbf{0},$$

da $\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}} = \boldsymbol{x}_{t+1}$ für $\lambda = 1$ (vgl. (4.27)) und demnach $\boldsymbol{z}^* \underset{\lambda=1}{=} \mathbf{0}$ $(*)$.

Anschaulich bedeutet dies, dass im Falle von $\lambda = 1/2$ die inverse Kovarianzmatrix ohne Gewichtung aus (4.26) betrachtet wird. Für $\lambda \to 0$ wird die neue Beobachtung nicht betrachtet und demnach wird die inverse Kovarianzmatrix aus dem vorherigen Schritt herangezogen. Der Fall $\lambda \to 1$ bedeutet, dass die gesamte Vergangenheit vergessen wird und der Fokus nur auf der neuen Beobachtung liegt. In diesem Fall ist die inverse Kovarianzmatrix jedoch nicht definiert, da nur eine einzelne Beobachtung betrachtet werden würde.

Die Lernrate λ lässt sich einerseits für den gesamten betrachteten Zeitraum, in dem das Modell immer durch neue Beobachtungen angepasst wird, festsetzen. Andererseits stellen Kuncheva und Plumpton (2008, S. 516 f.) eine adaptive Lernrate durch heuristisches Vorgehen vor. Die Lernrate wird dabei als Funktion des „running estimate“ des Klassifikationsfehlers aufgefasst. Dies bedeutet, dass zu jedem Zeitpunkt im Datenstrom bestimmt wird, ob die aktuelle Beobachtung mittels des aktuellen Modells richtig klassifiziert wird oder nicht. Mit diesen Klassifikationen wird laufend die Fehlerrate des Fensters der letzten L Beobachtungen $W_{t,L} = \{\boldsymbol{x}_{t-L}, \ldots, \boldsymbol{x}_t\}$ im Datenstrom kalkuliert. Zu jedem Zeitpunkt t wird gleitend die Fehlerdifferenz $\Delta_{t,L}$ der Fehlerraten zweier aufeinanderfolgender Fenster der Größe L berechnet und die Lernrate anhand dieser angepasst durch (Kuncheva und Plumpton, 2008, S. 516)

$$\lambda = \lambda^{1+\Delta_{t,L}}.$$

In der Kanonischen Linearen Diskriminanzanalyse basiert die Klassifikationsregel schlussendlich auf M verschiedenen Diskriminanzfunktionen (3.22) (bzw. analog (3.23)), wobei die theoretischen Größen hier durch die aktuellen Schätzer ersetzt werden:

$$\underset{c=1,\ldots,M}{\arg\max}\, \hat{h}_{L,t}^{(c)}(\boldsymbol{x}) = \underset{c=1,\ldots,M}{\arg\max} \left(\boldsymbol{x}^T \boldsymbol{S}_t^{-1} \boldsymbol{m}_{n_t^{(c)}}^{(c)} - \frac{1}{2} \left(\boldsymbol{m}_{n_t^{(c)}}^{(c)} \right)^T \boldsymbol{S}_t^{-1} \boldsymbol{m}_{n_t^{(c)}}^{(c)} + \log P_t^{(c)} \right). \quad (4.30)$$

Die Klasse der nächsten Beobachtung $\boldsymbol{x}_{t+1}$ im Datenstrom wird demnach mithilfe der aktuellen Klassifikationsregel prognostiziert durch

$$\hat{c}_{t+1} = \underset{c=1,\ldots,M}{\arg\max}\, \hat{h}_{L,t}^{(c)}(\boldsymbol{x}_{t+1}).$$

Anmerkungen In dieser Dissertation wurden einige Korrekturen der Formeln aus Kuncheva und Plumpton (2008) vorgenommen. Die Autorinnen definieren $\boldsymbol{z} := \boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_{t+1}^{(k)}}^{(k)}$ anstelle von $\boldsymbol{z} := \boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_t^{(k)}}^{(k)}$ und leiten die neue Kovarianzmatrix zum Zeitpunkt $t+1$ durch

$$\boldsymbol{S}_{t+1} = \frac{n_t}{n_t + 1} \left(\boldsymbol{S}_t + \frac{1}{n_t} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_{t+1}^{(k)}}^{(k)} \right) \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}_{n_{t+1}^{(k)}}^{(k)} \right)^T \right) \quad (4.31)$$

her (vgl. Kuncheva und Plumpton (2008, S. 513)).

Es lässt sich jedoch zeigen, dass diese Update-Formel nicht korrekt ist, da sie nicht zu demselben Ergebnis führt, welches man erhält, wenn die Kovarianzmatrix auf allen Daten der

Zeitpunkte 1 bis $t+1$ berechnet wird (ohne schrittweise Aktualisierungen). Die korrigierte Version (4.23) wird in (B.1) im Anhang B hergeleitet.

Im Falle von (4.31) ergibt sich die inverse Kovarianzmatrix mithilfe der Sherman-Morrison-Woodbury Formel aus Satz 3 durch (Kuncheva und Plumpton, 2008, S. 513)

$$\boldsymbol{S}_{t+1}^{-1} = \frac{n_t + 1}{n_t} \left(\boldsymbol{S}_t^{-1} - \frac{\boldsymbol{S}_t^{-1} \boldsymbol{z}\boldsymbol{z}^T \boldsymbol{S}_t^{-1}}{n_t + \boldsymbol{z}^T \boldsymbol{S}_t^{-1} \boldsymbol{z}} \right)$$

anstelle der korrigierten Version (4.26).

Die Einführung der Lernrate λ in der Kovarianzmatrix zieht weitere nötige Veränderungen der Formeln nach sich. Während die Gewichtung von $\boldsymbol{z}$ mit λ, der alten Kovarianzmatrix $\boldsymbol{S}_t$ mit $1-\lambda$ sowie die Ersetzung der gesamten Anzahl der Beobachtungen durch die gewichtete Variante $(1-\lambda)n_t + \lambda$ in der aktuellen Kovarianzmatrix $\boldsymbol{S}_{t+1}$ aus (4.31) und Anwendung der Sherman-Morrison-Woodbury Formel zur aktualisierten inversen Kovarianzmatrix

$$\boldsymbol{S}_{t+1}^{-1} = \frac{(1-\lambda)n_t + \lambda}{(1-\lambda)n_t} \left(\boldsymbol{S}_t^{-1} - \frac{\boldsymbol{S}_t^{-1} \boldsymbol{z}\boldsymbol{z}^T \boldsymbol{S}_t^{-1}}{\frac{(1-\lambda)n_t}{\lambda} + \boldsymbol{z}^T \boldsymbol{S}_t^{-1} \boldsymbol{z}} \right)$$

führt (vgl. Kuncheva und Plumpton (2008, S. 514)), die die nötigen Eigenschaften bezüglich der Grenzwerte von λ aufweist, ist dies bei den jeweiligen Ergänzungen und Ersetzungen in (4.23) nicht mehr der Fall, da auch der Term $\boldsymbol{z}$ korrigiert wurde (s. o.). Die weiteren nötigen Anpassungen für die aktualisierte inverse Kovarianzmatrix mit Lernrate (4.29) wurden in diesem Abschnitt hergeleitet.

Für ein Rechenbeispiel (mit fester Lernrate $\lambda = 1/2$) auf einem eindimensionalen Datenstrom sei auf Anhang A.2 verwiesen.

4.4 Online Diskriminanzanalyse mit adaptivem Vergessen (kurz: LDA-AF/QDA-AF) (Anagnostopoulos et al., 2012)

Anagnostopoulos et al. (2012) nutzen die Methode des adaptiven exponentiellen Vergessens, um eine Online Diskriminanzanalyse (Lineare sowie Quadratische Diskriminanzanalyse) zu entwickeln, welche sich adaptiv an eine Veränderung der zugrunde liegenden Verteilung der Beobachtungen im Datenstrom anpassen soll (engl. *Online Discriminant Analysis with Adaptive Forgetting for Streaming Classification*). Dazu ziehen sie eine Reihe von rekursiven Formeln für die nötigen Größen der Diskriminanzanalyse heran und führen Faktoren für exponentielles Vergessen ein, um die Verteilung vergangener Beobachtungen immer geringer zu gewichten und somit adaptives Anpassen der Schätzer der Diskriminanzanalyse an eine Veränderung der zugrunde liegenden Verteilung zu ermöglichen. Diese Faktoren

passen sich im Laufe des Datenstroms beim Auftreten neuer Beobachtungen selbst adaptiv an (*self-tuning*). Dazu wird die Idee des Gradientenabstiegs verwendet.

Im Folgenden wird diese Methode mit Anpassung an die Notation dieser Arbeit erläutert. An einigen Stellen wurden Korrekturen von Notationsfehlern oder Erweiterungen vorgenommen. Diese sind mit $(**)$ gekennzeichnet. Einige größere Abweichungen werden auch hier zusätzlich im Abschnitt „Anmerkungen“ (s. Seite 89 ff.) erläutert. Für die ursprüngliche Beschreibung inklusive weiterer Herleitungen und Beweise sei auf Anagnostopoulos et al. (2012) verwiesen.

Mit $\boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)})}^{(c)}$ seien die $n_t^{(c)}$ Beobachtungen aus $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ bezeichnet, die bis zum Zeitpunkt t jeweils in Klasse c realisiert wurden. Zudem wird angenommen, dass die Beobachtung zum Zeitpunkt t der Klasse c zugehörig ist, also $g(\boldsymbol{x}_t) = c$ gilt (vgl. Seite 32).

Die Idee basiert auf dem generativen Ansatz zur Bildung einer Klassifikationsregel, bei dem die gemeinsame Verteilung durch das Produkt aus klassenbedingter Verteilung und apriori Verteilung der Klassen $P(\boldsymbol{X}_t, \boldsymbol{W}) = P(\boldsymbol{X}_t | \boldsymbol{W} = \boldsymbol{w})\, P(\boldsymbol{W})$ (vgl. Satz 1 auf Seite 11) gebildet wird und demnach in zwei Schätzprobleme aufgeteilt wird.

Sei zunächst die klassenbedingte Verteilung $P(\boldsymbol{X}_t | \boldsymbol{W} = \boldsymbol{w})$ betrachtet. Bei der Diskriminanzanalyse wird diese für jede Klasse durch eine multivariate Normalverteilung (vgl. (3.3) und Seite 59) modelliert. Alle folgenden Herleitungen und Formeln werden beispielhaft für die Verteilung der Klasse $c \in \{1, \ldots, M\}$ beschrieben. Bei dieser Methode wird zunächst angenommen, dass die Anzahl der Klassen M für den gesamten Datenstrom, d. h. bereits bei der Initialisierung bekannt ist. Eine Erweiterung hierzu wird ab Seite 93 eingeführt.

Der Erwartungswertvektor sowie die Kovarianzmatrix der Klasse c lassen sich jeweils auf Basis der Beobachtungen $\boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)})}^{(c)}$ mithilfe der Maximum-Likelihood Methode schätzen, indem die negative log-Likelihood (NLL) der multivariaten Normalverteilung

$$
\begin{aligned}
&\mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)})}^{(c)}) \\
&= -\log f(\boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)})}^{(c)}; \boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}) \\
&= \frac{1}{2} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t}} \left(\log \left| \boldsymbol{\Sigma}^{(c)} \right| + \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right) \right) + const. \qquad (4.32)
\end{aligned}
$$

minimiert wird. Lösungen des Minimierungsproblems sind die ML-Schätzer ((4.3) für den Erwartungswertvektor) bzw. die folgenden rekursiven Varianten:

$$
\begin{aligned}
\boldsymbol{m}_{n_t^{(c)}}^{(c)} &= \left(1 - \frac{1}{n_t^{(c)}} \right) \boldsymbol{m}_{n_{t-1}^{(c)}}^{(c)} + \frac{1}{n_t^{(c)}} \cdot \boldsymbol{x}_t \\
&= \frac{1}{n_t^{(c)}} \left(\left(n_t^{(c)} - 1 \right) \boldsymbol{m}_{n_{t-1}^{(c)}}^{(c)} + \boldsymbol{x}_t \right) \quad \text{(vgl. dazu (4.8))}, \quad \boldsymbol{m}_{n_0^{(c)}}^{(c)} := \boldsymbol{0},
\end{aligned}
$$

$$\hat{\mathbf{\Pi}}_t^{(c)} = \left(1 - \frac{1}{n_t^{(c)}}\right) \hat{\mathbf{\Pi}}_{t-1}^{(c)} + \frac{1}{n_t^{(c)}} \cdot \boldsymbol{x}_t \boldsymbol{x}_t^T = \frac{1}{n_t^{(c)}} \left(\left(n_t^{(c)} - 1\right) \hat{\mathbf{\Pi}}_{t-1}^{(c)} + \boldsymbol{x}_t \boldsymbol{x}_t^T\right), \quad \hat{\mathbf{\Pi}}_0^{(c)} := \mathbf{0},$$

$$\hat{\mathbf{\Sigma}}_t^{(c)} = \hat{\mathbf{\Pi}}_t^{(c)} - \boldsymbol{m}_{n_t^{(c)}}^{(c)} \left(\boldsymbol{m}_{n_t^{(c)}}^{(c)}\right)^T.$$

Nun wird angenommen, dass sich Erwartungswertvektor und Kovarianzmatrix über die Zeit ändern können, die zugrunde liegende Verteilung also einem concept drift unterliegen kann. Anstelle von (4.32) wird eine Summe von Likelihood Termen betrachtet, bei der Summanden zeitlich vergangener Beobachtungen durch die Faktoren $\lambda_{(j)}^{(c)} \in [0,1]$, $j = 1, \ldots, n_t^{(c)} - 1$, ein zunehmend geringeres Gewicht (exponentielles Vergessen) bekommen. Es wird angenommen, dass zum Zeitpunkt t eine Beobachtung aus Klasse c realisiert wird:

$$\mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \mathbf{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)})}^{(c)}) = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} \left(\underbrace{\left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right)}_{=:v_i^{(c)}} \mathcal{L}(\boldsymbol{\mu}^{(c)}, \mathbf{\Sigma}^{(c)}; \boldsymbol{x}_i) \right)$$
$$+ \underbrace{v_t^{(c)}}_{:=1} \mathcal{L}(\boldsymbol{\mu}^{(c)}, \mathbf{\Sigma}^{(c)}; \boldsymbol{x}_t). \qquad (**) \qquad (4.33)$$

Ausformuliert ist das Produkt $v_i^{(c)}$ der Faktoren zum Zeitpunkt t:

$$\begin{aligned}
&\lambda_{(1)}^{(c)} \cdots \lambda_{(n_t^{(c)}-1)}^{(c)} && \text{für die erste Beobachtung } \boldsymbol{x}_{(1)}^{(c)} \text{ in Klasse } c, \\
&\qquad\vdots \\
&\lambda_{(n_t^{(c)}-1)}^{(c)} && \text{für die vorletzte Beobachtung } \boldsymbol{x}_{(n_t^{(c)}-1)}^{(c)} \text{ in Klasse } c.
\end{aligned}$$

Somit wird nicht der zeitliche Verlauf bzw. die Veränderung der interessierenden Größen Erwartungswertvektor und Kovarianzmatrix direkt modelliert, sondern der „Anteil" vergangener Beobachtungen wird in der Likelihood herabgewichtet, sodass sich diese unterschiedliche Gewichtung der Likelihood Terme auf die Schätzer auswirkt.

Zunächst wird der Fall *unveränderlicher Faktoren* $\lambda^{(c)} := \lambda_{(j)}^{(c)}$, $j = 1, \ldots, n_t^{(c)} - 1$, betrachtet. Für $\lambda^{(c)} := 1$ ist die gewichtete negative log-Likelihood (4.33) identisch mit der ungewichteten Version (4.32). Für $\lambda^{(c)} := 0$ wird die gesamte Datenhistorie des Datenstroms vergessen und nur die Likelihood basierend auf der aktuellen Beobachtung $\boldsymbol{x}_t$ betrachtet.

Durch Minimierung von (4.33) lassen sich die folgenden Schätzer bei Betrachtung von exponentiellem Vergessen herleiten (Herleitung in (B.2)/(B.3), (B.5)/(B.6) in Anhang B.2):

$$\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \frac{v_i^{(c)}}{N_t^{(c)}} \cdot \boldsymbol{x}_i, \qquad (4.34)$$

$$\tilde{\mathbf{\Pi}}_t^{(c)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \frac{v_i^{(c)}}{N_t^{(c)}} \cdot \boldsymbol{x}_i \boldsymbol{x}_i^T, \qquad (4.35)$$

$$\tilde{\boldsymbol{\Sigma}}_t^{(c)} = \tilde{\boldsymbol{\Pi}}_t^{(c)} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)^T, \tag{4.36}$$

wobei

$$N_t^{(c)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} v_i^{(c)} \text{ mit } v_t^{(c)} := 1 \text{ (bzw. Gewicht für Beob. mit größtem Index ist 1)} \quad (**) \tag{4.37}$$

eine Normierungskonstante für die Gewichte $v_i^{(c)}$ (s. (4.33)) darstellt.

Die entsprechenden rekursiven Schätzer, welche in Datenströmen online bestimmt werden können, haben die folgende Form (Herleitung in (B.4), (B.7), (B.8) und (B.9) in Anhang B.2):

$$\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} = \begin{cases} \tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)}, & \text{falls } c_t \neq c, \\ \left(1 - \dfrac{1}{N_t^{(c)}}\right) \tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)} + \dfrac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t, & \text{falls } c_t = c \in \{1, \ldots, M\}, \\ \boldsymbol{x}_t, & \text{falls } c_t = c = M+1 \quad (**), \end{cases} \tag{4.38}$$

$$\tilde{\boldsymbol{\Pi}}_t^{(c)} = \begin{cases} \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}, & \text{falls } c_t \neq c, \\ \left(1 - \dfrac{1}{N_t^{(c)}}\right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)} + \dfrac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \boldsymbol{x}_t^T, & \text{falls } c_t = c \in \{1, \ldots, M\}, \\ \mathbf{0}, & \text{falls } c_t = c = M+1, \end{cases} \tag{4.39}$$

$$\tilde{\boldsymbol{\Sigma}}_t^{(c)} = \begin{cases} \tilde{\boldsymbol{\Sigma}}_{t-1}^{(c)}, & \text{falls } c_t \neq c, \\ \tilde{\boldsymbol{\Pi}}_t^{(c)} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)^T, & \text{falls } c_t = c \in \{1, \ldots, M\}, \\ \mathbf{0}, & \text{falls } c_t = c = M+1, \end{cases} \tag{4.40}$$

wobei

$$N_t^{(c)} = \begin{cases} N_{t-1}^{(c)}, & \text{falls } c_t \neq c, \\ \lambda_{(n_{t-1}^{(c)})}^{(c)} N_{t-1}^{(c)} + 1, & \text{falls } c_t = c \in \{1, \ldots, M\}, \\ 1, & \text{falls } c_t = c = M+1 \end{cases} \tag{4.41}$$

mit $\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)} := \mathbf{0}$, $\tilde{\boldsymbol{\Pi}}_0^{(c)} := \mathbf{0}$, $N_0^{(c)} := 0$, $\lambda_{(n_0^{(c)})}^{(c)}$ Zufallszahl aus $[\lambda_-, \lambda_+]$ (s. (4.49)) $(**)$.

Die Fallunterscheidungen wurden zusätzlich zu der Ausführung von Anagnostopoulos et al. (2012) formal ergänzt (vgl. Erläuterungen ab Seite 93 ff.).

Um den Faktor $\lambda^{(c)} := \lambda_{(j)}^{(c)}$ zu finden, schlagen Anagnostopoulos et al. (2012, S. 143 f.) einen datenbasierten Ansatz vor. Sie betrachten die negative log-Likelihood der folgenden Beobachtung, welche mit $J_{t+1}^{(c)}$ bezeichnet wird, um die Prognosegüte für die Beobachtung $\boldsymbol{x}_{t+1}$ auf Basis der Schätzer zum Zeitpunkt t zu bestimmen. Dies erfolgt nur, wenn die folgende Beobachtung in Klasse c realisiert wird, also $\boldsymbol{x}_{t+1} = \boldsymbol{x}_{(n_t^{(c)}+1)}^{(c)} = \boldsymbol{x}_{(n_{t+1}^{(c)})}^{(c)}$. Ansonsten

bleiben zunächst alle im Folgenden definierten Größen erhalten, bis eine Beobachtung in Klasse c auftritt und ein Update erfolgt ($J_{t+1}^{(c)} := J_t^{(c)}$, $\left(J_{t+1}^{(c)}\right)' := \left(J_t^{(c)}\right)'$, $\lambda^{(c)}_{(n_{t+1}^{(c)})} := \lambda^{(c)}_{(n_t^{(c)})}$ und $\alpha_{t+1}^{(c)} := \alpha_t^{(c)}$; vgl. dazu die Erklärung ab Seite 93 oder Algorithmus 3 auf Seite 87 bzw. Algorithmus 4 auf Seite 88).

Die Anpassung der folgenden Beobachtung durch die negative log-Likelihood wird als Gütemaß herangezogen:

$$\begin{aligned} J_{t+1}^{(c)} &= \mathcal{L}(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}, \tilde{\boldsymbol{\Sigma}}_t^{(c)}; \boldsymbol{x}_{t+1}) \\ &= \frac{1}{2}\log\left|\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right| + \frac{1}{2}\left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)^T \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right) + const. \end{aligned} \tag{4.42}$$

$J_{t+1}^{(c)}$ ist eine Funktion in Abhängigkeit des unveränderlichen Faktors $\lambda^{(c)}$, da die Schätzer $\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}$ und $\tilde{\boldsymbol{\Sigma}}_t^{(c)}$ in Abhängigkeit von $\lambda^{(c)}$ betrachtet werden können, sodass der Gradient gebildet werden kann (Herleitung in (B.10) in Anhang B.2):

$$\begin{aligned} \left(J_{t+1}^{(c)}\right)' &= \frac{\partial \mathcal{L}(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}, \tilde{\boldsymbol{\Sigma}}_t^{(c)}; \boldsymbol{x}_{t+1})}{\partial \lambda^{(c)}} \\ &= \frac{1}{2}\left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)^T \left(-2\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}\left(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)' + \left(\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}\right)'\left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)\right) \\ &\quad + \frac{1}{2}\left(\log\left|\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right|\right)', \end{aligned} \tag{4.43}$$

wobei

$$\left(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)' = \frac{\partial \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}}{\partial \lambda^{(c)}} = \left(1 - \frac{1}{N_t^{(c)}}\right)\left(\tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}}\right)' - \frac{\left(N_t^{(c)}\right)'}{\left(N_t^{(c)}\right)^2}\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}}\right), \quad \left(\tilde{\boldsymbol{m}}^{(c)}_{n_0^{(c)}}\right)' := \boldsymbol{0}, \tag{4.44}$$

$$\left(\tilde{\boldsymbol{\Pi}}_t^{(c)}\right)' = \frac{\partial \tilde{\boldsymbol{\Pi}}_t^{(c)}}{\partial \lambda^{(c)}} = \left(1 - \frac{1}{N_t^{(c)}}\right)\left(\tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}\right)' - \frac{\left(N_t^{(c)}\right)'}{\left(N_t^{(c)}\right)^2}\left(\boldsymbol{x}_t\,\boldsymbol{x}_t^T - \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}\right), \quad \left(\tilde{\boldsymbol{\Pi}}_0^{(c)}\right)' := \boldsymbol{0}, \tag{4.45}$$

$$\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)' = \frac{\partial \tilde{\boldsymbol{\Sigma}}_t^{(c)}}{\partial \lambda^{(c)}} = \left(\tilde{\boldsymbol{\Pi}}_t^{(c)}\right)' - \left(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)'\left(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)^T - \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\left(\left(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)'\right)^T, \tag{4.46}$$

$$\left(N_t^{(c)}\right)' = \frac{\partial N_t^{(c)}}{\partial \lambda^{(c)}} = \lambda^{(c)}_{(n_{t-1}^{(c)})}\left(N_{t-1}^{(c)}\right)' + N_{t-1}^{(c)}, \quad \left(N_0^{(c)}\right)' := 0 \quad (**). \tag{4.47}$$

Fallunterscheidungen für die Formeln sind in (4.70)–(4.73) auf Seite 94 aufgeführt.

Die Gradienten der Inverse der Kovarianzmatrix und der logarithmierten Determinante der Kovarianzmatrix in (4.43) können mithilfe der folgenden Rechenregeln bestimmt werden (Anagnostopoulos et al., 2012, S. 144; Harville, 2008, S. 309/311):

$$\left(\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}\right)' = -\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)'\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}, \quad \left(\log\left|\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right|\right)' = \operatorname{tr}\left(\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)'\right). \tag{4.48}$$

Der Faktor $\lambda^{(c)}_{(n_t^{(c)})}$ wird daraufhin mithilfe eines Gradientenabstiegs online adaptiv angepasst (*self-tuning*):

$$\lambda^{(c)}_{(n_{t+1}^{(c)})} = \min\left(\lambda_+,\ \max\left(\lambda_-,\ \lambda^{(c)}_{(n_t^{(c)})} - \alpha_t^{(c)}\left(J_{t+1}^{(c)}\right)'\right)\right) =: \left[\lambda^{(c)}_{(n_t^{(c)})} - \alpha_t^{(c)}\left(J_{t+1}^{(c)}\right)'\right]_{\lambda_-}^{\lambda_+} \tag{4.49}$$

mit Schrittweite $\alpha_t^{(c)} > 0$.

Anagnostopoulos et al. (2012, S. 146) nennen einen unteren Schwellenwert von $\lambda_- = 0.7$ und einen oberen von $\lambda_+ = 0.999$, um numerische Instabilitäten des Algorithmus zu vermeiden.

Hier ist zu beachten, dass der Faktor $\lambda^{(c)}_{(j)}$ nicht mehr unveränderlich ist und die Annahme $\lambda^{(c)} = \lambda^{(c)}_{(j)}\ \forall j$, unter derer die Formeln (4.43)–(4.47) hergeleitet wurden, nicht mehr gilt. Die Bestimmung des Gradienten (4.43) ist nur für den Fall unveränderlicher Faktoren exakt. Da die Schrittweite $\alpha_t^{(c)}$ jedoch im Allgemeinen klein gewählt wird und sich der Faktor $\lambda^{(c)}_{(j)}$ über die Zeit dementsprechend nur leicht verändert, kann (4.43) als approximativer Gradient angesehen werden und im Gradientenabstieg (4.49) zur adaptiven Anpassung des Faktors verwendet werden.

Die Schrittweite kann entweder je nach Anwendung auf einen kleinen, positiven Wert festgesetzt werden, wie z. B. $10^{-8} \leq \alpha_t^{(c)} \leq 10^{-6}$. Darüber hinaus existieren bereits Algorithmen zur Adaption der Schrittweite in Gradientenabstiegsverfahren. Anagnostopoulos et al. (2012, S. 146) verwenden den Algorithmus *RPROP* (*resilient propagation*) (Riedmiller und Braun, 1993), welcher auf dem aktuellen Gradienten und dem Gradienten des vorherigen Schrittes basiert. Falls die Gradienten in dieselbe Richtung zeigen, wird die Schrittweite minimal vergrößert, falls die Vorzeichen hingegen gegensätzlich sind, wird die Schrittweite verringert:

$$\alpha_t^{(c)} = \begin{cases} \left[1.01\,\alpha_{t-1}^{(c)}\right]_{\alpha_{\min}}^{\alpha_{\max}}, & \text{falls } \left|\left(J_t^{(c)}\right)'\right| > 10^{-7} \text{ und } \left(J_t^{(c)}\right)'\left(J_{t-1}^{(c)}\right)' > 0, \\ \left[0.99\,\alpha_{t-1}^{(c)}\right]_{\alpha_{\min}}^{\alpha_{\max}}, & \text{falls } \left|\left(J_t^{(c)}\right)'\right| > 10^{-7} \text{ und } \left(J_t^{(c)}\right)'\left(J_{t-1}^{(c)}\right)' \leq 0, \\ \alpha_{t-1}^{(c)}, & \text{falls } \left|\left(J_t^{(c)}\right)'\right| \leq 10^{-7}, \end{cases} \tag{4.50}$$

Algorithmus 1 G-AF (in Anlehnung an Anagnostopoulos et al. (2012, S. 149) (∗∗))

Require: $\boldsymbol{x}_t$, $\tilde{\boldsymbol{\theta}}_{t-1} := \Big\{\tilde{\boldsymbol{m}}_{n_{t-1}}, d_{t-1}, \boldsymbol{G}_{t-1}, \tilde{\boldsymbol{\Sigma}}_{t-1}, \tilde{\boldsymbol{\Pi}}_{t-1}, \tilde{\boldsymbol{m}}'_{n_{t-1}}, d'_{t-1}, \boldsymbol{G}'_{t-1}, \tilde{\boldsymbol{\Sigma}}'_{t-1}, \tilde{\boldsymbol{\Pi}}'_{t-1}, \lambda_{(n_{t-1})}, N_{t-1}, N'_{t-1}, J_{t-1}, J'_{t-1}, \alpha_{t-1}\Big\}$,
$\text{fix}_{\boldsymbol{m}} = \texttt{FALSE}$, $\boldsymbol{m}_0 = \emptyset$, $\text{fix}_{\boldsymbol{\Sigma}} = \texttt{FALSE}$, $\boldsymbol{\Sigma}_0 = \emptyset$, $d_0 = \emptyset$, $\boldsymbol{G}_0 = \emptyset$,
$\boldsymbol{\Pi}_0 = \emptyset$ (der Vollständigkeit halber), (λ_-, λ_+, $\alpha_{\min}$, $\alpha_{\max}$)

1: Aktualisiere N_t und N'_t durch (4.41) und (4.47).
2: **if** $\text{fix}_{\boldsymbol{m}}$ **then**
3: $\quad\tilde{\boldsymbol{m}}_{n_t} \leftarrow \boldsymbol{m}_0$; $\quad\tilde{\boldsymbol{m}}'_{n_t} \leftarrow \boldsymbol{0}$
4: **else**
5: $\quad$Aktualisiere $\tilde{\boldsymbol{m}}_{n_t}$ und $\tilde{\boldsymbol{m}}'_{n_t}$ durch (4.38) und (4.44).
6: **end if**
7: **if** $\text{fix}_{\boldsymbol{\Sigma}}$ **then**
8: $\quad\tilde{\boldsymbol{\Sigma}}_t \leftarrow \boldsymbol{\Sigma}_0$; $\quad\tilde{\boldsymbol{\Pi}}_t \leftarrow \boldsymbol{\Pi}_0$ (der Vollständigkeit halber); $\quad d_t \leftarrow d_0$; $\quad\boldsymbol{G}_t \leftarrow \boldsymbol{G}_0$
9: $\quad\tilde{\boldsymbol{\Sigma}}'_t \leftarrow \boldsymbol{0}$; $\quad\tilde{\boldsymbol{\Pi}}'_t \leftarrow \boldsymbol{0}$ (der Vollständigkeit halber); $\quad d'_t \leftarrow 0$; $\quad\boldsymbol{G}'_t \leftarrow \boldsymbol{0}$
10: **else**
11: $\quad$Aktualisiere $\tilde{\boldsymbol{\Pi}}_t$, $\tilde{\boldsymbol{\Sigma}}_t$, $\tilde{\boldsymbol{\Pi}}'_t$ und $\tilde{\boldsymbol{\Sigma}}'_t$ durch (4.39), (4.40), (4.45) und (4.46).
12: $\quad$Aktualisiere d_t, $\boldsymbol{G}_t$, d'_t und $\boldsymbol{G}'_t$.
13: **end if**
14: Aktualisiere den Gradienten J'_t durch (4.43).
15: Aktualisiere den Faktor $\lambda_{(n_t)} = \left[\lambda_{(n_{t-1})} - \alpha_{t-1} J'_t\right]_{\lambda_-}^{\lambda_+}$ durch (4.49).
16: Aktualisiere die Schrittweite α_t durch den *RPROP*-Algorithmus (4.50).
17: **return** $\tilde{\boldsymbol{\theta}}_t$

wobei $\alpha_{\min}$ und $\alpha_{\max}$ unterer bzw. oberer Schwellenwert für die Schrittweite sind (z. B. $\alpha_{\min} = 10^{-8}$ und $\alpha_{\max} = 10^{-6}$) und demnach

$$\begin{aligned}\left[1.01\,\alpha_{t-1}^{(c)}\right]_{\alpha_{\min}}^{\alpha_{\max}} &:= \min\left(\alpha_{\max},\ \max\left(\alpha_{\min},\ 1.01\,\alpha_{t-1}^{(c)}\right)\right) \text{ bzw.} \\ \left[0.99\,\alpha_{t-1}^{(c)}\right]_{\alpha_{\min}}^{\alpha_{\max}} &:= \min\left(\alpha_{\max},\ \max\left(\alpha_{\min},\ 0.99\,\alpha_{t-1}^{(c)}\right)\right).\end{aligned} \tag{4.51}$$

Der Startwert $\alpha_0^{(c)}$ kann zufällig aus dem Intervall $[\alpha_{\min}, \alpha_{\max}]$ gewählt werden.

All diese Schritte werden in einem Teilalgorithmus *G-AF* (*Gaussian adaptive forgetting*) (s. Algorithmus 1 mit vereinfachter Notation ohne Exponent c für Klasse) zusammengefasst. Hier ist zu beachten, dass die Größen $\text{fix}_{\boldsymbol{m}}$, $\boldsymbol{m}_0$, $\text{fix}_{\boldsymbol{\Sigma}}$, $\boldsymbol{\Sigma}_0$, d_0, $\boldsymbol{G}_0$ im Falle der Quadratischen Diskriminanzanalyse nicht relevant sind. Eine Erläuterung folgt ab Seite 86 bei Heranziehen des *G-AF* Algorithmus für eine Online Variante der Linearen Diskriminanzanalyse. Ebenso erfolgt eine Erklärung der Parameter $\boldsymbol{G}_t$, $\boldsymbol{G}'_t$, d_t und d'_t auf Seite 85.

Dasselbe Vorgehen wird für die a-priori Verteilung der Klassen betrachtet. Die folgenden Schritte beschreiben den Teilalgorithmus *M-AF* (*multinomial adaptive forgetting*) (s. Algorithmus 2). Die Verteilung der Klassen folgt einer Multinomialverteilung (3.2). Die negative log-Likelihood in Abhängigkeit der Anzahl an Beobachtungen in den einzelnen Klassen $n_t^{(1)}, \ldots, n_t^{(M)}$ bzw. alternativ und äquivalent in Abhängigkeit der Ausprägungen $c_1, \ldots, c_t$

Algorithmus 2 M-AF (in Anlehnung an Anagnostopoulos et al. (2012, S. 149) (∗∗))

Require: c_t, $\tilde{\boldsymbol{\theta}}_{t-1}^{(0)} := \left\{ \tilde{P}_{t-1}^{(1)}, \ldots, \tilde{P}_{t-1}^{(M)}, \left(\tilde{P}_{t-1}^{(1)}\right)', \ldots, \left(\tilde{P}_{t-1}^{(M)}\right)', \lambda_{t-1}^{(0)}, N_{t-1}^{(0)}, \left(N_{t-1}^{(0)}\right)', J_{t-1}^{(0)}, \left(J_{t-1}^{(0)}\right)', \alpha_{t-1}^{(0)} \right\}$, $(\lambda_-,\ \lambda_+,\ \alpha_{\min},\ \alpha_{\max})$

1: Aktualisiere $N_t^{(0)}$ durch (4.55) sowie die Schätzer für die a-priori Wahrscheinlichkeiten $\tilde{P}_t^{(c)}$, $c = 1, \ldots, M$, durch (4.54).
2: Aktualisiere $\left(\tilde{P}_t^{(c)}\right)'$, $c = 1, \ldots, M$, sowie $\left(N_t^{(0)}\right)'$ durch (4.59) und (4.60).
3: Aktualisiere den Gradienten $\left(J_t^{(0)}\right)'$ durch (4.58).
4: Aktualisiere den Faktor $\lambda_t^{(0)} = \left[\lambda_{t-1}^{(0)} - \alpha_{t-1}^{(0)} \left(J_t^{(0)}\right)'\right]_{\lambda_-}^{\lambda_+}$ durch (4.49).
5: Aktualisiere die Schrittweite $\alpha_t^{(0)}$ durch den *RPROP*-Algorithmus (4.50).
6: **return** $\tilde{\boldsymbol{\theta}}_t^{(0)}$

der Zielvariablen besitzt unter Beachtung der Nebenbedingung $\sum_{c=1}^{M} p^{(c)} = 1$ zum Zeitpunkt t die folgende Form (∗∗):

$$\mathcal{L}(p^{(1)}, \ldots, p^{(M)}; n_t^{(1)}, \ldots, n_t^{(M)}) = -\sum_{c=1}^{M} \left(n_t^{(c)} \log \left(\frac{p^{(c)}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) + const.$$

bzw. (4.52)

$$\mathcal{L}(p^{(1)}, \ldots, p^{(M)}; c_1, \ldots, c_t) = -\sum_{c=1}^{M} \left(\underbrace{\left(\sum_{i=1}^{t} \mathbb{1}_{\{c_i = c\}} \right)}_{=n_t^{(c)}} \log \left(\frac{p^{(c)}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) + const.$$

Auch an dieser Stelle wird angenommen, dass sich die Verteilung – in diesem Falle charakterisiert durch die a-priori Wahrscheinlichkeiten – über die Zeit ändern kann. Dementsprechend wird anstelle von (4.52) wieder eine gewichtete Summe von Likelihood Termen betrachtet, bei welcher der Einfluss vergangener Beobachtungen durch Einführung von Faktoren $\lambda_i^{(0)}$, $i = 1, \ldots, t-1$, exponentiell abklingt (∗∗):

$$\begin{aligned} \mathcal{L}^{(\boldsymbol{\lambda})}(p^{(1)}, \ldots, p^{(M)}; c_1, \ldots, c_t) = & \sum_{i=1}^{t-1} \left(\underbrace{\left(\prod_{j=i}^{t-1} \lambda_j^{(0)} \right)}_{=: v_i^{(0)}} \mathcal{L}(p^{(1)}, \ldots, p^{(M)}; c_i) \right) \\ & + \underbrace{v_t^{(0)}}_{:=1} \mathcal{L}(p^{(1)}, \ldots, p^{(M)}; c_t). \end{aligned} \tag{4.53}$$

Durch Minimierung der gewichteten Likelihood-Summe lassen sich die rekursiven Formeln für die Schätzer $\tilde{P}_t^{(c)}$ für die a-priori Wahrscheinlichkeiten $p^{(c)}$, $c = 1, \ldots, M$, bestimmen:

$$\tilde{P}_t^{(c)} = \left(1 - \frac{1}{N_t^{(0)}}\right) \tilde{P}_{t-1}^{(c)} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t = c\}}, \quad \tilde{P}_0^{(c)} := \frac{1}{M}, \tag{4.54}$$

$$N_t^{(0)} = \lambda_{t-1}^{(0)} N_{t-1}^{(0)} + 1, \quad N_0^{(0)} := 0 \text{ oder } N_0^{(0)} := 1, \tag{4.55}$$

wobei auch hier

$$N_t^{(0)} = \sum_{i=1}^{t} v_i^{(0)} \text{ mit } v_t^{(0)} := 1 \tag{4.56}$$

eine Normierungskonstante für die Gewichte $v_i^{(0)}$ (s. (4.53)) darstellt. Eine Herleitung der rekursiven Formeln ist in (B.11)–(B.13) in Anhang B.2 zu finden. Im Falle von $N_0^{(0)} = 0$ ist die Initialisierung von $\tilde{P}_0^{(c)}$ irrelevant, da dann $N_1^{(0)} = 1$ in (4.54). Falls die Initialisierung durch $N_0^{(0)} = 1$ erfolgt, ist die Initialisierung von $\tilde{P}_0^{(c)}$ aus (4.54) sinnvoll, um Schätzer für die a-priori Wahrscheinlichkeiten zu repräsentieren. In diesem Fall kann der Startwert $\lambda_0^{(0)}$ für (4.55) zufällig aus dem Intervall $[\lambda_-, \lambda_+]$ gezogen werden.

Die negative log-Likelihood für die folgende Beobachtung bzw. Klassenausprägung c_{t+1} basierend auf den rekursiv bestimmten Schätzern $\tilde{P}_t^{(1)}, \ldots, \tilde{P}_t^{(M)}$ zum Zeitpunkt t im Datenstrom hat die folgende Form (∗∗):

$$J_{t+1}^{(0)} = \mathcal{L}(\tilde{P}_t^{(1)}, \ldots, \tilde{P}_t^{(M)}; c_{t+1}) = -\sum_{c=1}^{M} \mathbb{1}_{\{c_{t+1}=c\}} \left(\log \tilde{P}_t^{(c)} - \log \left(\sum_{k=1}^{M} \tilde{P}_t^{(k)} \right) \right). \tag{4.57}$$

Da sich auch die Schätzer $\tilde{P}_t^{(c)}$, $c = 1, \ldots, M$, in Abhängigkeit von $\lambda^{(0)} := \lambda_i^{(0)}$ betrachten lassen, lässt sich der Gradient von $J_{t+1}^{(0)}$ bilden (Herleitung in (B.14) in Anhang B.2):

$$\left(J_{t+1}^{(0)}\right)' = \frac{\partial \mathcal{L}(\tilde{P}_t^{(1)}, \ldots, \tilde{P}_t^{(M)}; c_{t+1})}{\partial \lambda^{(0)}} = -\sum_{c=1}^{M} \left(\left(\mathbb{1}_{\{c_{t+1}=c\}} - \tilde{P}_t^{(c)} \right) \frac{\left(\tilde{P}_t^{(c)}\right)'}{\tilde{P}_t^{(c)}} \right), \tag{4.58}$$

wobei

$$\left(\tilde{P}_t^{(c)}\right)' = \frac{\partial \tilde{P}_t^{(c)}}{\partial \lambda^{(0)}} = \left(1 - \frac{1}{N_t^{(0)}}\right) \left(\tilde{P}_{t-1}^{(c)}\right)' - \frac{\left(N_t^{(0)}\right)'}{\left(N_t^{(0)}\right)^2} \left(\mathbb{1}_{\{c_t=c\}} - \tilde{P}_{t-1}^{(c)} \right), \tag{4.59}$$

$$\left(N_t^{(0)}\right)' = \frac{\partial N_t^{(0)}}{\partial \lambda^{(0)}} = \lambda_{t-1}^{(0)} \left(N_{t-1}^{(0)}\right)' + N_{t-1}^{(0)} \quad (**) \tag{4.60}$$

die Ableitungen von (4.54) und (4.55) sind.

Der Faktor $\lambda_i^{(0)}$ wird ebenfalls mithilfe des Gradientenabstiegs (4.49) online angepasst, wobei der Gradient $\left(J_{t+1}^{(0)}\right)'$ betrachtet wird und alle Exponenten (c) in der Notation durch (0) zu ersetzen sind. Die Schrittweite $\alpha_t^{(0)}$ wird ebenfalls mithilfe des *RPROP* Algorithmus (4.50) aktualisiert. Zu beachten sei auch an dieser Stelle, dass der Gradient (4.58) nicht exakt ist, da durch den Aktualisierungsschritt die Annahme eines unveränderlichen Faktors $\lambda^{(0)} = \lambda_i^{(0)}$ $\forall i$ verletzt ist. Die Begründung dafür, dass $\left(J_{t+1}^{(0)}\right)'$ dennoch im Gradientenabstieg zur adaptiven Anpassung des Faktors verwendet werden kann, ist dieselbe wie zuvor: Da die Schrittweite $\alpha_t^{(0)}$ in jedem Schritt klein genug ist, sodass sich $\lambda_t^{(0)}$ nur

leicht verändert, kann die Annahme unveränderlicher Faktoren *approximativ* angenommen werden.

Anagnostopoulos et al. (2012, S. 145 f.) weisen darauf hin, dass der Teilalgorithmus *G-AF* aufgrund der Bestimmung des Gradienten $\left(J_{t+1}^{(c)}\right)'$ (s. (4.43)) basierend auf den Größen $d_t^{(c)} := \log\left|\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right|$ und $\boldsymbol{G}_t^{(c)} := \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}$ eine Komplexität von $\mathcal{O}(p^3)$ aufweist. Sie schlagen daher alternative rekursive Formeln mithilfe von Rechenregeln der Matrix Algebra für $d_t^{(c)}$, $\boldsymbol{G}_t^{(c)}$ und ihre Gradienten vor, sodass die Komplexität des gesamten *G-AF* Algorithmus auf $\mathcal{O}(p^2)$ reduziert und gleichzeitig die numerische Stabilität erhöht werden kann. Es wird also folgender Gradient betrachtet:

$$\begin{aligned}\left(J_{t+1}^{(c)}\right)' = \frac{1}{2}\left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T \left(-2\boldsymbol{G}_t^{(c)}\left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)' + \left(\boldsymbol{G}_t^{(c)}\right)'\left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)\right) \\ + \frac{1}{2}\left(d_t^{(c)}\right)'.\end{aligned} \tag{4.61}$$

Für die rekursiven Formeln für $d_t^{(c)}$, $\boldsymbol{G}_t^{(c)}$ und ihre Gradienten sei auf Anagnostopoulos et al. (2012, S. 145) verwiesen.

Die Teilalgorithmen *G-AF* und *M-AF* dienen als Basis für eine Online Version der Linearen und Quadratischen Diskriminanzanalyse. Die Vorhersage bei der Kanonischen Diskriminanzanalyse erfolgt auf Basis des Bayes-Theorems (vgl. Abschnitt 3.1). Eine Beobachtung $\boldsymbol{x}^*$ wird jener Klasse mit größter a-posteriori Wahrscheinlichkeit bzw. folgender minimaler Diskriminanzfunktion zugeordnet (vgl. (3.9) und (3.32)/(3.33)):

$$\begin{aligned}\hat{c}^* &= \underset{c=1,\ldots,M}{\arg\min}\left(\mathcal{L}(\boldsymbol{m}_{n_t^{(c)}}^{(c)}, \boldsymbol{S}_t^{(c)}; \boldsymbol{x}^*) - \log P_t^{(c)}\right) \\ &= \underset{c=1,\ldots,M}{\arg\min}\left(\frac{1}{2}\log\left|\boldsymbol{S}_t^{(c)}\right| + \frac{1}{2}\left(\boldsymbol{x}^* - \boldsymbol{m}_{n_t^{(c)}}^{(c)}\right)^T \left(\boldsymbol{S}_t^{(c)}\right)^{-1}\left(\boldsymbol{x}^* - \boldsymbol{m}_{n_t^{(c)}}^{(c)}\right) - \log P_t^{(c)}\right) \\ &\Leftrightarrow \underset{c=1,\ldots,M}{\arg\max}\left(-\frac{1}{2}\log\left|\boldsymbol{S}_t^{(c)}\right| - \frac{1}{2}\left(\boldsymbol{x}^* - \boldsymbol{m}_{n_t^{(c)}}^{(c)}\right)^T \left(\boldsymbol{S}_t^{(c)}\right)^{-1}\left(\boldsymbol{x}^* - \boldsymbol{m}_{n_t^{(c)}}^{(c)}\right) + \log P_t^{(c)}\right).\end{aligned} \tag{4.62}$$

Die anschauliche Idee dahinter ist, dass unter Auftreten einer neuen Beobachtung $\boldsymbol{x}^*$ die a-posteriori Wahrscheinlichkeit der Kombination $(\boldsymbol{x}^*, \hat{c}^*)$ mithilfe des Modells zum Zeitpunkt t maximiert wird. Zu beachten ist, dass bei der Linearen Diskriminanzanalyse die Annahme konstanter Kovarianzmatrizen, also $\boldsymbol{\Sigma}^{(1)} = \ldots = \boldsymbol{\Sigma}^{(M)}$ getroffen wird. Folglich wird als Schätzer für die „gemeinsame“ Kovarianzmatrix die gepoolte Kovarianzmatrix $\boldsymbol{S}_t$ (s. (4.5)) anstelle von $\boldsymbol{S}_t^{(c)}$ in (4.62) betrachtet.

In der Online Version des Algorithmus werden die ML-Schätzer (4.3) und (4.4) bzw. (4.5) durch ihre rekursiv bestimmten Varianten $\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}$, $\tilde{\boldsymbol{\Sigma}}_t^{(c)}$ und $\tilde{P}_t^{(c)}$ (vgl. (4.38), (4.40) und (4.54)) aus *G-AF* und *M-AF* ersetzt. Für die Quadratische Diskriminanzanalyse wird

demnach im Datenstrom die Klasse $\tilde{c}_t$ für die Beobachtung $\boldsymbol{x}_t$ zum Zeitpunkt t auf Basis der aktuellen Schätzer vorhergesagt:

$$\tilde{c}_t = \underset{c=1,\ldots,M}{\arg\min} \Bigg(\underbrace{\frac{1}{2} d_{t-1}^{(c)} + \frac{1}{2} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)} \right)^T \boldsymbol{G}_{t-1}^{(c)} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)} \right)}_{= J_t^{(c)} - const. = \mathcal{L}(\tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)}, \tilde{\boldsymbol{\Sigma}}_{t-1}^{(c)}; \boldsymbol{x}_t) - const.} - \log \tilde{P}_{t-1}^{(c)} \Bigg). \tag{4.63}$$

Im Falle der Linearen Diskriminanzanalyse wird die gepoolte Kovarianzmatrix als ein einzelner Schätzer für die Kovarianzmatrix betrachtet:

$$\tilde{c}_t = \underset{c=1,\ldots,M}{\arg\min} \left(\frac{1}{2} d_{t-1}^{(P)} + \frac{1}{2} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)} \right)^T \boldsymbol{G}_{t-1}^{(P)} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)} \right) - \log \tilde{P}_{t-1}^{(c)} \right). \tag{4.64}$$

Die Bestimmung von $d^{(P)}$ und $\boldsymbol{G}^{(P)}$ wird im Folgenden noch erläutert. An dieser Stelle sei jedoch darauf hingewiesen, dass die Bestimmung von $d^{(P)}$ streng genommen für die Klassifikationsregel irrelevant ist, da der Summand als additive Konstante bei jeder Diskriminanzfunktion der M Klassen gleichermaßen eingeht und daher bei der Minimierung unerheblich ist. Auf die Berechnung könnte daher verzichtet werden, was den Algorithmus noch effizienter machen würde.

Sobald das wahre Klassenlabel c_t zum nächsten Zeitpunkt $t+1$ auftritt, werden daraufhin die Größen (4.38)–(4.41), (4.44)–(4.47) und $d_t^{(c)}$, $\left(d_t^{(c)}\right)'$, $\boldsymbol{G}_t^{(c)}$, $\left(\boldsymbol{G}_t^{(c)}\right)'$ bzw. (4.54), (4.55) und (4.59), (4.60) sowie jeweils die Gradienten $\left(J_t^{(c)}\right)'$, $c = 1, \ldots, M$, durch (4.43) bzw. $\left(J_t^{(0)}\right)'$ durch (4.58), die Faktoren $\lambda_{(n_t^{(c)})}^{(c)}$, $c = 1, \ldots, M$, bzw. $\lambda_t^{(0)}$ und die Schrittweiten $\alpha_t^{(c)}$, $c = 1, \ldots, M$, bzw. $\alpha_t^{(0)}$ durch die Teilalgorithmen *G-AF* und *M-AF* aktualisiert. Diese Schritte werden für den Datenstrom abwechselnd durchgeführt, um eine Online Variante der Quadratischen Diskriminanzanalyse zu erhalten: Die Vorhersage der Klasse und darauf folgend ein Durchgang des *M-AF* Algorithmus und ein Durchlauf des *G-AF* Algorithmus für die entsprechende Klasse. Dies ist in Algorithmus 3 zusammengefasst.

Bei der Online Variante der Linearen Diskriminanzanalyse wird aufgrund der gepoolten Kovarianzmatrix ein weiterer Durchlauf des *G-AF* Algorithmus durchgeführt. Daher werden in Summe für jeden Aktualisierungsschritt des Klassifikationsmodells zwei Durchläufe des *G-AF* Algorithmus und ein Durchgang des *M-AF* Algorithmus sowie $M+2$ Parametervektoren $\tilde{\boldsymbol{\theta}}_{t-1}^{(c)}$, $c = 1, \ldots, M$, $\tilde{\boldsymbol{\theta}}_{t-1}^{(0)}$ sowie $\tilde{\boldsymbol{\theta}}_{t-1}^{(P)}$ benötigt. Der Ablauf ist im Algorithmus 4 beschrieben. Im Folgenden wird dieser Algorithmus durch über die Beschreibung von Anagnostopoulos et al. (2012) hinausgehende ergänzende Erklärungen erläutert.

Die Idee ist, dass zunächst in einem *G-AF* Durchlauf für die Aktualisierung des Mittelwertvektors die Kovarianzmatrix konstant gehalten wird durch (vgl. Zeilen 3 und 6 in Algorithmus 4 und Zeilen 8 und 9 in Algorithmus 1)

$$\boldsymbol{\Sigma}_0 := \tilde{\boldsymbol{\Sigma}}_t^{(c)} := \tilde{\boldsymbol{\Sigma}}_{t-1}^{(P)}, \quad \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)' := \mathbf{0}, \quad \text{wobei} \quad \tilde{\boldsymbol{\Sigma}}_0^{(P)} := \mathbf{0}.$$

Algorithmus 3 QDA-AF (in Anlehnung an Anagnostopoulos et al. (2012, S. 148)) (∗∗)

Require: $\boldsymbol{x}_t$, c_t, $\tilde{\boldsymbol{\theta}}_{t-1}^{(0)} := \Big\{\tilde{P}_{t-1}^{(1)}, \ldots, \tilde{P}_{t-1}^{(M)}, \Big(\tilde{P}_{t-1}^{(1)}\Big)', \ldots, \Big(\tilde{P}_{t-1}^{(M)}\Big)', \lambda_{t-1}^{(0)}, N_{t-1}^{(0)}, \Big(N_{t-1}^{(0)}\Big)', J_{t-1}^{(0)}, \Big(J_{t-1}^{(0)}\Big)', \alpha_{t-1}^{(0)}\Big\}$, $(\lambda_-,\ \lambda_+,\ \alpha_{\min},\ \alpha_{\max})$,

$\tilde{\boldsymbol{\theta}}_{t-1}^{(c)} := \Big\{\tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)}, d_{t-1}^{(c)}, \boldsymbol{G}_{t-1}^{(c)}, \tilde{\boldsymbol{\Sigma}}_{t-1}^{(c)}, \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}, \Big(\tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)}\Big)', \Big(d_{t-1}^{(c)}\Big)', \Big(\boldsymbol{G}_{t-1}^{(c)}\Big)', \Big(\tilde{\boldsymbol{\Sigma}}_{t-1}^{(c)}\Big)', \Big(\tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}\Big)', \lambda_{(n_{t-1}^{(c)})}^{(c)}, N_{t-1}^{(c)}, \Big(N_{t-1}^{(c)}\Big)', J_{t-1}^{(c)}, \Big(J_{t-1}^{(c)}\Big)', \alpha_{t-1}^{(c)}\Big\}, \quad c = 1, \ldots, M$

1: Sage die Klasse $\tilde{c}_t$ durch (4.63) vorher.
2: Führe den *M-AF* Algorithmus für c_t und $\tilde{\boldsymbol{\theta}}_{t-1}^{(0)}$ durch und erhalte $\tilde{\boldsymbol{\theta}}_t^{(0)}$.
3: **for** $c = 1$ **to** M **do**
4: **if** $c_t = c$ **then**
5: Führe den *G-AF* Algorithmus für $\boldsymbol{x}_t$ und $\tilde{\boldsymbol{\theta}}_{t-1}^{(c)}$ durch und erhalte $\tilde{\boldsymbol{\theta}}_t^{(c)}$.
6: **else**
7: $\tilde{\boldsymbol{\theta}}_t^{(c)} \leftarrow \tilde{\boldsymbol{\theta}}_{t-1}^{(c)}$
8: **end if**
9: **end for**
10: **return** $\tilde{c}_t$, $\tilde{\boldsymbol{\theta}}_t^{(0)}$, $\tilde{\boldsymbol{\theta}}_t^{(c)}$, $c = 1, \ldots, M$

Dies erfolgt, damit sich eine temporäre Veränderung der klassenspezifischen Kovarianzmatrix nicht auf die Anpassung des Faktors $\lambda_{(n_t^{(c)})}^{(c)}$ im self-tuning Schritt (4.49) auswirkt, weil die Kovarianzmatrix im Gradienten $\Big(J_t^{(c)}\Big)'$ betrachtet wird. Dies soll vermieden werden, weil die klassenbasierten Kovarianzmatrizen in der Linearen Diskriminanzanalyse keinen direkten Einfluss bei der Klassifikationsregel haben.

Wenn die effizienteren Schätzer $d_t^{(c)}$ und $\boldsymbol{G}_t^{(c)}$ und ihre Gradienten anstelle von $\tilde{\boldsymbol{\Sigma}}_t^{(c)}$ im Algorithmus verwendet werden, werden zusätzlich diese festgehalten (vgl. Zeilen 3 und 6 in Algorithmus 4):

$$d_0 := d_t^{(c)} := d_{t-1}^{(P)}, \quad \Big(d_t^{(c)}\Big)' := 0, \quad \boldsymbol{G}_0 := \boldsymbol{G}_t^{(c)} := \boldsymbol{G}_{t-1}^{(P)}, \quad \Big(\boldsymbol{G}_t^{(c)}\Big)' := \mathbf{0}.$$

Die Beobachtung $\boldsymbol{x}_t$ wird daraufhin durch den ML-Schätzer $\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}$, welcher zuvor im *G-AF* Durchlauf bestimmt wurde, zentriert (Zeile 7 in Algorithmus 4):

$$\boldsymbol{\xi}_t = \boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}. \tag{4.65}$$

Auf Basis dieser zentrierten Beobachtung wird ein weiterer Durchlauf des *G-AF* Algorithmus zur Aktualisierung der gepoolten Kovarianzmatrix inklusive eines eigenen Faktors $\lambda_t^{(P)}$ durchgeführt, wobei alle Schätzer mit Exponent (P) betrachtet werden. Hierbei wird nun hingegen der Mittelwertvektor und dessen Gradient konstant auf $\mathbf{0}$ gesetzt (vgl. Zeilen 12 und 13 in Algorithmus 4 und Zeile 3 in Algorithmus 1):

$$\boldsymbol{m}_0 := \tilde{\boldsymbol{m}}_{n_t}^{(P)} := \mathbf{0}, \quad \Big(\tilde{\boldsymbol{m}}_{n_t}^{(P)}\Big)' := \mathbf{0}.$$

Algorithmus 4 LDA-AF (in Anlehnung an Anagnostopoulos et al. (2012, S. 150)) (∗∗)

Require: $\boldsymbol{x}_t$, c_t, $\tilde{\boldsymbol{\theta}}_{t-1}^{(0)} := \Big\{\tilde{P}_{t-1}^{(1)}, \ldots, \tilde{P}_{t-1}^{(M)}, \Big(\tilde{P}_{t-1}^{(1)}\Big)', \ldots, \Big(\tilde{P}_{t-1}^{(M)}\Big)', \lambda_{t-1}^{(0)}, N_{t-1}^{(0)}, \Big(N_{t-1}^{(0)}\Big)', J_{t-1}^{(0)}, \Big(J_{t-1}^{(0)}\Big)', \alpha_{t-1}^{(0)}\Big\}$, $(\lambda_-,\, \lambda_+,\, \alpha_{\min},\, \alpha_{\max})$,

$\tilde{\boldsymbol{\theta}}_{t-1}^{(c)} := \Big\{\tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)}, d_{t-1}^{(c)}, \boldsymbol{G}_{t-1}^{(c)}, \tilde{\boldsymbol{\Sigma}}_{t-1}^{(c)}, \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}, \Big(\tilde{\boldsymbol{m}}_{n_{t-1}^{(c)}}^{(c)}\Big)', \Big(d_{t-1}^{(c)}\Big)', \Big(\boldsymbol{G}_{t-1}^{(c)}\Big)', \Big(\tilde{\boldsymbol{\Sigma}}_{t-1}^{(c)}\Big)', \Big(\tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}\Big)', \lambda_{(n_{t-1}^{(c)})}^{(c)}, N_{t-1}^{(c)}, \Big(N_{t-1}^{(c)}\Big)', J_{t-1}^{(c)}, \Big(J_{t-1}^{(c)}\Big)', \alpha_{t-1}^{(c)}\Big\}, \quad c = 1, \ldots, M,$

$\tilde{\boldsymbol{\theta}}_{t-1}^{(P)} := \Big\{\tilde{\boldsymbol{m}}_{n_{t-1}}^{(P)}, d_{t-1}^{(P)}, \boldsymbol{G}_{t-1}^{(P)}, \tilde{\boldsymbol{\Sigma}}_{t-1}^{(P)}, \tilde{\boldsymbol{\Pi}}_{t-1}^{(P)}, \Big(\tilde{\boldsymbol{m}}_{n_{t-1}}^{(P)}\Big)', \Big(d_{t-1}^{(P)}\Big)', \Big(\boldsymbol{G}_{t-1}^{(P)}\Big)', \Big(\tilde{\boldsymbol{\Sigma}}_{t-1}^{(P)}\Big)', \Big(\tilde{\boldsymbol{\Pi}}_{t-1}^{(P)}\Big)', \lambda_{t-1}^{(P)}, N_{t-1}^{(P)}, \Big(N_{t-1}^{(P)}\Big)', J_{t-1}^{(P)}, \Big(J_{t-1}^{(P)}\Big)', \alpha_{t-1}^{(P)}\Big\}$

1: Sage die Klasse $\tilde{c}_t$ durch (4.64) vorher.
2: Führe den *M-AF* Algorithmus für c_t und $\tilde{\boldsymbol{\theta}}_{t-1}^{(0)}$ durch und erhalte $\tilde{\boldsymbol{\theta}}_t^{(0)}$.
3: $\text{fix}_{\boldsymbol{\Sigma}} \leftarrow$ `TRUE`; $\boldsymbol{\Sigma}_0 \leftarrow \tilde{\boldsymbol{\Sigma}}_{t-1}^{(P)}$; $\boldsymbol{\Pi}_0 \leftarrow \tilde{\boldsymbol{\Pi}}_{t-1}^{(P)}$ (der Vollständigkeit halber); $d_0 \leftarrow d_{t-1}^{(P)}$; $\boldsymbol{G}_0 \leftarrow \boldsymbol{G}_{t-1}^{(P)}$
4: **for** $c = 1$ **to** M **do**
5: **if** $c_t = c$ **then**
6: Führe den *G-AF* Algorithmus für $\boldsymbol{x}_t$, $\tilde{\boldsymbol{\theta}}_{t-1}^{(c)}$, $\text{fix}_{\boldsymbol{\Sigma}}$, $\boldsymbol{\Sigma}_0$, d_0, $\boldsymbol{G}_0$ und $\boldsymbol{\Pi}_0$ (der Vollständigkeit halber) durch und erhalte $\tilde{\boldsymbol{\theta}}_t^{(c)}$.
7: $\boldsymbol{\xi}_t \leftarrow \boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}$
8: **else**
9: $\tilde{\boldsymbol{\theta}}_t^{(c)} \leftarrow \tilde{\boldsymbol{\theta}}_{t-1}^{(c)}$
10: **end if**
11: **end for**
12: $\text{fix}_{\boldsymbol{m}} \leftarrow$ `TRUE`; $\boldsymbol{m}_0 \leftarrow \boldsymbol{0}$
13: Führe den *G-AF* Algorithmus für $\boldsymbol{\xi}_t$, $\tilde{\boldsymbol{\theta}}_{t-1}^{(P)}$, $\text{fix}_{\boldsymbol{m}}$ und $\boldsymbol{m}_0$ durch und erhalte $\tilde{\boldsymbol{\theta}}_t^{(P)}$. (Aktualisierung der gepoolten Kovarianzmatrix)
14: **return** $\tilde{c}_t$, $\tilde{\boldsymbol{\theta}}_t^{(0)}$, $\tilde{\boldsymbol{\theta}}_t^{(c)}$, $c = 1, \ldots, M$, $\tilde{\boldsymbol{\theta}}_t^{(P)}$

Der Grund dafür ist, dass sonst mithilfe des *G-AF* Algorithmus der Mittelwertvektor basierend auf allen Beobachtungen aktualisiert werden würde, dieser hingegen nicht benötigt wird. Zudem wird nur so ein aktueller Schätzer für die gepoolte Kovarianzmatrix produziert, da der Algorithmus für die zentrierten Beobachtungen durchgeführt wird. Zum Zeitpunkt t gilt für den *G-AF* Durchlauf zur Schätzung der gepoolten Kovarianzmatrix:

$$
\begin{aligned}
N_t^{(P)} &= \lambda_{t-1}^{(P)} N_{t-1}^{(P)} + 1 \quad \text{(vgl. (4.41))}, \quad N_0^{(P)} := 0, \\
\Big(N_t^{(P)}\Big)' &= \lambda_{t-1}^{(P)} \Big(N_{t-1}^{(P)}\Big)' + N_{t-1}^{(P)} \quad \text{(vgl. (4.47))}, \\
\tilde{\boldsymbol{\Pi}}_t^{(P)} &= \left(1 - \frac{1}{N_t^{(P)}}\right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(P)} + \frac{1}{N_t^{(P)}} \cdot \boldsymbol{\xi}_t \boldsymbol{\xi}_t^T \quad \text{(vgl. (4.39))} \\
&\overset{(4.65)}{=} \left(1 - \frac{1}{N_t^{(P)}}\right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(P)} + \frac{1}{N_t^{(P)}} \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right) \left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T, \quad \tilde{\boldsymbol{\Pi}}_0^{(P)} := \boldsymbol{0}, \\
\tilde{\boldsymbol{\Sigma}}_t^{(P)} &= \tilde{\boldsymbol{\Pi}}_t^{(P)} - \tilde{\boldsymbol{m}}_{n_t}^{(P)} \Big(\tilde{\boldsymbol{m}}_{n_t}^{(P)}\Big)^T = \tilde{\boldsymbol{\Pi}}_t^{(P)} - \boldsymbol{0}\,\boldsymbol{0}^T = \tilde{\boldsymbol{\Pi}}_t^{(P)} \quad \text{(vgl. (4.40))},
\end{aligned}
$$

was ein analoger rekursiver Schätzer für den Schätzer

$$\tilde{\boldsymbol{\Sigma}}_t^{(P)} = \sum_{c=1}^{M} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \frac{v_i^{(P)}}{N_t^{(P)}} \left(\boldsymbol{x}_i - \tilde{\boldsymbol{m}}_{n_i^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \tilde{\boldsymbol{m}}_{n_i^{(c)}}^{(c)} \right)^T \tag{4.66}$$

der gewichteten gepoolten Kovarianzmatrix ist.

Hier ist zu beachten, dass dieser Schätzer (4.66) eine größere Varianz aufweist als der gewichtete Schätzer

$$\boldsymbol{\Sigma}_t^{(P)} = \sum_{c=1}^{M} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \frac{v_i^{(P)}}{N_t^{(P)}} \left(\boldsymbol{x}_i - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)^T \tag{4.67}$$

der gepoolten Kovarianzmatrix, welcher nicht im Datenstrom aktualisiert wird, da der Schätzer $\tilde{\boldsymbol{m}}_{n_i^{(c)}}^{(c)}$ eine größere Varianz aufweist als $\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}$, welcher zum Zeitpunkt t auf allen Beobachtungen aus Klasse c basiert. Im Falle von exponentiellem Vergessen wird das Gewicht $v_i^{(P)}$ für vergangene (zentrierte) Beobachtungen jedoch exponentiell kleiner durch $v_i^{(P)} = \prod_{j=i}^{t-1} \lambda_i^{(P)}$, sodass die höhere Varianz des Schätzers $\tilde{\boldsymbol{m}}_{n_i^{(c)}}^{(c)}$ relativiert wird. Der Schätzer (4.67) kann nicht im Datenstrom aktualisiert bzw. rekursiv bestimmt werden, da $\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}$ zu früheren Zeitpunkten $< t$ noch nicht berechnet werden kann.

Der Vorteil dieser zusätzlichen Einführung eines Durchlaufs des *G-AF* Teilalgorithmus im *LDA-AF* Algorithmus überwiegt, da zunächst wie bereits oben erwähnt der irrelevante Einfluss einer Änderung der klassenbedingten Kovarianzmatrix bei der Aktualisierung des Faktors $\lambda^{(c)}$ ausgeschaltet wird.

Insgesamt stellt demnach die Schätzung von (4.66) einen guten Kompromiss dar, um zu einer Online Variante der Linearen Diskriminanzanalyse mit adaptivem Vergessen beizutragen.

Anmerkungen Die Formel der gewichteten log-Likelihood (4.33) wurde korrigiert, da in Anagnostopoulos et al. (2012, S. 143) ein Notationsfehler vorliegt. Die Summe der Likelihood Terme wird dort (ausformuliert in der Notation dieser Arbeit) durch

$$\begin{aligned} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)})}^{(c)}) &= \lambda_{(n_t^{(c)}-1)}^{(c)} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)}-1)}^{(c)}) \\ &\quad + \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_t) \\ &= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \left(\left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_i) \right) \end{aligned}$$

beschrieben. Die zweite Gleichheit ist hier falsch, da die Summe nicht bis zum Zeitindex t laufen kann. Das Produkt ist dann für $i = t$ nicht definiert, da $n_t^{(c)} \leq j \leq n_t^{(c)} - 1$ aufgrund

der Annahme geordneter Zeitpunkte nicht definiert ist. In dem Produkt sollen fortlaufende Faktoren $\lambda_{(j)}^{(c)}$ für exponentielles Vergessen multipliziert werden. Wenn die erste Gleichung umgeformt wird, führt dies zudem zu

$$
\begin{aligned}
&\lambda_{(n_t^{(c)}-1)}^{(c)} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)}-1)}^{(c)}) + \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_t) \\
&= \lambda_{(n_t^{(c)}-1)}^{(c)} \left(\lambda_{(n_t^{(c)}-2)}^{(c)} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)}-2)}^{(c)}) + \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(n_t^{(c)}-1)}^{(c)}) \right) \\
&\quad + \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_t) \\
&= \lambda_{(n_t^{(c)}-1)}^{(c)} \left(\lambda_{(n_t^{(c)}-2)}^{(c)} \left(\lambda_{(n_t^{(c)}-3)}^{(c)} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)}-3)}^{(c)}) \right. \right. \\
&\qquad \left. \left. + \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(n_t^{(c)}-2)}^{(c)}) \right) + \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(n_t^{(c)}-1)}^{(c)}) \right) \\
&\quad + \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_t) \\
&= \lambda_{(n_t^{(c)}-1)}^{(c)} \lambda_{(n_t^{(c)}-2)}^{(c)} \lambda_{(n_t^{(c)}-3)}^{(c)} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)}-3)}^{(c)}) \\
&\quad + \lambda_{(n_t^{(c)}-1)}^{(c)} \lambda_{(n_t^{(c)}-2)}^{(c)} \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)}-2)}^{(c)}) \\
&\quad + \lambda_{(n_t^{(c)}-1)}^{(c)} \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)}-1)}^{(c)}) \\
&\quad + \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_t) \\
&= \ldots \\
&= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t-1}} \left(\left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_i) \right) + \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_t),
\end{aligned}
$$

also der Gleichung (4.33).

In dieser Notation muss der letzte Summand, die Likelihood basierend auf der aktuellen Beobachtung $\boldsymbol{x}_t$, explizit ausformuliert werden, da dieser das „Gewicht“ $\lambda_{(n_t^{(c)})}^{(c)} := v_t^{(c)} := 1$ erhält. Die anderen Summanden vergangener Beobachtungen erhalten durch das Produkt der Faktoren $\lambda_{(j)}^{(c)}$ ein zunehmend geringeres Gewicht (exponentielles Vergessen). Dieses Produkt wird als $v_i^{(c)} := \prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)}$ bezeichnet (s. (4.33)). Bei der Normierungskonstante (4.37) gilt dann entsprechend $v_t^{(c)} := 1$ für den Summanden des Zeitpunktes t.

Des Weiteren wurde die rekursive Update-Formel des Gradienten der Normierungskonstanten $\left(N_t^{(0)}\right)'$ (s. (4.60)) korrigiert. Anagnostopoulos et al. (2012, S. 145) definieren diese als

$$\left(N_t^{(0)}\right)' = \lambda_t^{(0)} \left(N_{t-1}^{(0)}\right)' + N_{t-1}^{(0)}.$$

Der aktuelle Faktor $\lambda_t^{(0)}$ liegt jedoch zu diesem Zeitpunkt noch nicht vor. Die Bildung des Gradienten (mit unveränderlichen Faktoren $\lambda^{(0)}$) von (4.55) mithilfe der Produktregel führt zu

$$\left(N_t^{(0)}\right)' = \frac{\partial\left(\lambda_{t-1}^{(0)} N_{t-1}^{(0)} + 1\right)}{\partial\lambda^{(0)}} = \lambda_{t-1}^{(0)}\left(N_{t-1}^{(0)}\right)' + \left(\lambda_{t-1}^{(0)}\right)' N_{t-1}^{(0)} = \lambda_{t-1}^{(0)}\left(N_{t-1}^{(0)}\right)' + N_{t-1}^{(0)}.$$

Weiter wurde der Pseudocode zum *G-AF* Algorithmus (s. Algorithmus 1 auf Seite 82) angepasst, da jener von Anagnostopoulos et al. (2012, S. 149) die Beschreibung des Algorithmus nicht in korrekter Reihenfolge darstellt. In dem Pseudocode werden dort zunächst alle Aktualisierungsschritte der interessierenden Größen vorgenommen, also Zeilen 1, 5, 11 und 12 aus Algorithmus 1. Erst im Anschluss daran werden für den *LDA-AF* Algorithmus die optionalen Fallunterscheidungen aus Zeilen 2 und 3 bzw. Zeilen 7–9 aufgeführt. Dies spiegelt allerdings nicht exakt die Idee bei der Verwendung des *G-AF* Algorithmus innerhalb des *LDA-AF* Algorithmus wider, bei dem die optionalen Fallunterscheidungen relevant sind.

Werden zunächst alle Größen aktualisiert, würde sich im Durchlauf für die gepoolte Kovarianzmatrix auf Basis der zentrierten Beobachtung $\boldsymbol{\xi}_t = \boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}$ und der Reihe von Schätzern $\tilde{\boldsymbol{\theta}}_{t-1}^{(P)}$ Folgendes ergeben:

$$\begin{aligned}
N_t^{(P)} &= \lambda_{t-1}^{(P)} N_{t-1}^{(P)} + 1, \quad N_0^{(P)} := 0,\\
\left(N_t^{(P)}\right)' &= \lambda_{t-1}^{(P)}\left(N_{t-1}^{(P)}\right)' + N_{t-1}^{(P)},\\
\tilde{\boldsymbol{m}}_{n_t}^{(P)} &= \left(1 - \frac{1}{N_t^{(P)}}\right)\underbrace{\tilde{\boldsymbol{m}}_{n_{t-1}}^{(P)}}_{=\mathbf{0}} + \frac{1}{N_t^{(P)}}\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right) = \frac{1}{N_t^{(P)}}\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right),\\
\tilde{\boldsymbol{\Pi}}_t^{(P)} &= \left(1 - \frac{1}{N_t^{(P)}}\right)\tilde{\boldsymbol{\Pi}}_{t-1}^{(P)} + \frac{1}{N_t^{(P)}}\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T, \quad \tilde{\boldsymbol{\Pi}}_0^{(P)} := \mathbf{0},\\
\tilde{\boldsymbol{\Sigma}}_t^{(P)} &= \tilde{\boldsymbol{\Pi}}_t^{(P)} - \tilde{\boldsymbol{m}}_{n_t}^{(P)}\left(\tilde{\boldsymbol{m}}_{n_t}^{(P)}\right)^T\\
&= \left(1 - \frac{1}{N_t^{(P)}}\right)\tilde{\boldsymbol{\Pi}}_{t-1}^{(P)} + \frac{1}{N_t^{(P)}}\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T\\
&\quad - \frac{1}{\left(N_t^{(P)}\right)^2}\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T\\
&= \left(1 - \frac{1}{N_t^{(P)}}\right)\left(\tilde{\boldsymbol{\Pi}}_{t-1}^{(P)} + \frac{1}{N_t^{(P)}}\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T\right),
\end{aligned}$$

was kein Schätzer für eine gepoolte Kovarianzmatrix ist. Erst im Nachhinein wird hier $\tilde{\boldsymbol{m}}_{n_t}^{(P)}$ konstant auf $\mathbf{0}$ gesetzt.

Auf der anderen Seite wird beim Durchlauf für die Aktualisierung des Mittelwertvektors trotzdem zusätzlich zunächst die klassenbedingte Kovarianzmatrix aktualisiert bzw. die

Größen $d_t^{(c)}$, $\left(d_t^{(c)}\right)'$, $\boldsymbol{G}_t^{(c)}$ und $\left(\boldsymbol{G}_t^{(c)}\right)'$ werden berechnet, bevor diese direkt im Anschluss konstant auf 0 bzw. $\mathbf{0}$ gesetzt werden, was ineffizient ist. Die korrigierte Version ist in Algorithmus 1 zu finden, welcher im Algorithmus 4 (*LDA-AF*) aufgerufen wird. Die Erläuterung erfolgt ab Seite 86.

Auch dieser Algorithmus *LDA-AF* wird im Pseudocode von Anagnostopoulos et al. (2012, S. 150) formal nicht ganz korrekt definiert. Zusätzlich zu der nicht korrekten Durchführung des Algorithmus *G-AF* zur Aktualisierung der Parameter werden die Zuweisungen $\text{fix}_{\boldsymbol{\Sigma}} \leftarrow \text{OPTIMAL}$ und $\text{fix}_{\boldsymbol{\mu}} \leftarrow \text{OPTIMAL}$ verwendet, wobei OPTIMAL nirgends beschrieben wird. Die Durchführung der Online Variante der Linearen Diskriminanzanalyse mit adaptivem Vergessen (*LDA-AF*) wird daher ab Seite 86 ausführlicher mit Bezug auf den korrigierten Pseudocode in Algorithmus 4 beschrieben.

Bezüglich des Algorithmus *M-AF* bzw. der Aktualisierung der Verteilung der Klassen wurden ebenfalls einige Verbesserungen und ausführlichere Herleitungen vorgenommen, da im Originalpaper von Anagnostopoulos et al. (2012) Notationsfehler vorliegen oder die Formeln theoretisch nicht ganz eindeutig definiert sind.

Der Zusammenhang der beiden Varianten der negativen log-Likelihood aus (4.52) wird in dem Originalpaper von Anagnostopoulos et al. (2012, S. 144) nicht erläutert. Die Autoren beschreiben die allgemeine Form der negativen log-Likelihood für eine beliebige Klassenausprägung c_i in Anpassung an die Notation dieser Arbeit durch

$$\mathcal{L}(p^{(1)}, \ldots, p^{(M)}; c_i) = -\sum_{c=1}^{M} \left(\mathbb{1}_{\{c_i = c\}} \cdot \log\left(\frac{p^{(c)}}{\sum_{k=1}^{M} p^{(k)}} \right) \right).$$

Allerdings fehlt der Übergang zur gesamten negativen log-Likelihood für alle Klassenausprägungen der Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t$ und daraufhin die konkrete Formulierung der gewichteten Summe der Likelihood Terme (4.53). Es werden lediglich die aus der Minimierung dieser gewichteten Summe der Likelihood Terme resultierenden Schätzer $\tilde{P}_t^{(c)}$ aufgeführt (Anagnostopoulos et al., 2012, S. 144).

Die negative log-Likelihood für die folgende Klassenausprägung c_{t+1} basierend auf den rekursiv bestimmten Schätzern $\tilde{P}_t^{(1)}, \ldots, \tilde{P}_t^{(M)}$ zum Zeitpunkt t im Datenstrom definieren sie (in Notation dieser Arbeit) durch (Anagnostopoulos et al., 2012, S. 145)

$$J_{t+1}^{(0)} = \mathcal{L}(\tilde{P}_{t+1}^{(1)}, \ldots, \tilde{P}_{t+1}^{(M)}; \boldsymbol{x}_{t+1}) = -\sum_{c=1}^{M} \mathbb{1}_{\{c_{t+1} = c\}} \left(\log \tilde{P}_t^{(c)} - \log\left(\sum_{k=1}^{M} \tilde{P}_t^{(k)} \right) \right).$$

Natürlich soll dies aber die NLL für die Klassenausprägung c_{t+1} und nicht die Beobachtung $\boldsymbol{x}_{t+1}$ sein. Zudem werden die aktuellen Schätzer $\tilde{P}_t^{(1)}, \ldots, \tilde{P}_t^{(M)}$ und nicht jene des kommenden Zeitpunktes $t+1$ betrachtet, wie anhand der rechten Seite der Formel auch deutlich wird. Diese Notationsfehler wurden in (4.57) behoben.

Neue Erweiterung Der Algorithmus ist bisher so angelegt, dass von Beginn an die Anzahl der auftretenden Klassen im Datenstrom bekannt sein muss. Dies ist für praktische Anwendungen jedoch ungünstig, da es durchaus sein kann, dass im Laufe der Zeit neue Klassenausprägungen hinzukommen und die Anzahl demnach nicht bereits bei der Initialisierung der Methode bekannt ist.

Im Teilalgorithmus *M-AF* basieren die rekursiven Formeln der Schätzer für die a-priori Wahrscheinlichkeiten sowie ihrer Gradienten auf den Klassenausprägungen. Die Formeln werden entgegen (4.54) und (4.59) folgendermaßen angepasst bzw. erweitert für den Fall, dass eine neue Klasse $c_t = M + 1$ auftritt.

$$\tilde{P}_t^{(c)} = \begin{cases} \left(1 - \frac{1}{N_t^{(0)}}\right) \tilde{P}_{t-1}^{(c)} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t = c\}}, & \text{falls } c_t = c \in \{1, \ldots, M\} \text{ oder } c_t \neq c, \\ \frac{1}{N_t^{(0)}}, & \text{falls } c_t = c = M + 1, \end{cases} \tag{4.68}$$

$$\left(\tilde{P}_t^{(c)}\right)' = \begin{cases} \frac{N_t^{(0)} - 1}{N_t^{(0)}} \cdot \left(\tilde{P}_{t-1}^{(c)}\right)' - \frac{\left(N_t^{(0)}\right)'}{\left(N_t^{(0)}\right)^2} \left(\mathbb{1}_{\{c_t = c\}} - \tilde{P}_{t-1}^{(c)}\right), & \text{falls } c_t = c \in \{1, \ldots, M\} \\ & \text{oder } c_t \neq c, \\ 0, & \text{falls } c_t = c = M + 1. \end{cases} \tag{4.69}$$

Im Falle von $c_t = M + 1$ wird somit der Parametervektor $\boldsymbol{\theta}_t^{(0)}$ (vgl. Algorithmus 2 auf Seite 83) um die Elemente $\tilde{P}_t^{(M+1)}$ und $\left(\tilde{P}_t^{(M+1)}\right)'$ erweitert.

Mittels des Algorithmus *G-AF* werden die klassenbedingten Verteilungen rekursiv modelliert. Alle Formeln basieren auf der neuen Beobachtung und indirekt auf ihrer Klassenausprägung. Für den Fall $c_t = M + 1$ wird somit ab diesem Zeitpunkt ein weiterer Parametervektor

$$\begin{aligned} \tilde{\boldsymbol{\theta}}_t^{(M+1)} := \Big\{ & \tilde{\boldsymbol{m}}_{n_t^{(M+1)}}^{(M+1)}, d_t^{(M+1)}, \boldsymbol{G}_t^{(M+1)}, \tilde{\boldsymbol{\Sigma}}_t^{(M+1)}, \tilde{\boldsymbol{\Pi}}_t^{(M+1)}, \left(\tilde{\boldsymbol{m}}_{n_t^{(M+1)}}^{(M+1)}\right)', \left(d_t^{(M+1)}\right)', \\ & \left(\boldsymbol{G}_t^{(M+1)}\right)', \left(\tilde{\boldsymbol{\Sigma}}_t^{(M+1)}\right)', \left(\tilde{\boldsymbol{\Pi}}_t^{(M+1)}\right)', \lambda_{(n_t^{(M+1)})}^{(M+1)}, N_t^{(M+1)}, \left(N_t^{(M+1)}\right)', J_t^{(M+1)}, \\ & \left(J_t^{(M+1)}\right)', \alpha_t^{(M+1)} \Big\} \end{aligned}$$

in *G-AF* (vgl. Algorithmus 1 auf Seite 82) und *QDA-AF* (vgl. Algorithmus 3 auf Seite 87) bzw. *LDA-AF* (vgl. Algorithmus 4 auf Seite 88) betrachtet. Die negative log-Likelihood $J_t^{(M+1)}$ wird eigentlich nicht benötigt und muss daher nicht aktualisiert werden. Die Größe fließt nur der Vollständigkeit halber als Input in die Algorithmen ein. Die Größen d_t, d_t', $\boldsymbol{G}_t$ und $\boldsymbol{G}_t'$ sind optional und können ohne Aktualisierung auch folgendermaßen bestimmt bzw. initialisiert werden: $d_t^{(M+1)} := \log\left|\tilde{\boldsymbol{\Sigma}}_t^{(M+1)}\right|$, $\left(d_t^{(M+1)}\right)' := \left(\log\left|\tilde{\boldsymbol{\Sigma}}_t^{(M+1)}\right|\right)'$, $\boldsymbol{G}_t^{(M+1)} := \left(\tilde{\boldsymbol{\Sigma}}_t^{(M+1)}\right)^{-1}$, $\left(\boldsymbol{G}_t^{(M+1)}\right)' := \left(\left(\tilde{\boldsymbol{\Sigma}}_t^{(M+1)}\right)^{-1}\right)'$.

Für alle anderen bisher eingeführten Formeln (zusätzlich zu (4.38)–(4.41)) wird im Folgenden eine Fallunterscheidung integriert, die zusätzlich den Fall einer bisher nicht aufgetretenen Klasse betrachtet und die aktuellen Größen für alle Klassen definiert, aus der keine aktuelle Beobachtung auftritt:

- Die entsprechenden Gradienten (vgl. (4.44)–(4.47)):

$$\left(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)' = \begin{cases} \left(\tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}}\right)', & \text{falls } c_t \neq c, \\ \left(1 - \frac{1}{N_t^{(c)}}\right)\left(\tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}}\right)' - \frac{\left(N_t^{(c)}\right)'}{\left(N_t^{(c)}\right)^2}\left(\boldsymbol{x}_t - \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}}\right), & \text{falls } c_t = c \in \{1,\ldots,M\}, \\ \mathbf{0}, & \text{falls } c_t = c = M+1, \end{cases} \tag{4.70}$$

$$\left(\tilde{\boldsymbol{\Pi}}_t^{(c)}\right)' = \begin{cases} \left(\tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}\right)', & \text{falls } c_t \neq c, \\ \left(1 - \frac{1}{N_t^{(c)}}\right)\left(\tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}\right)' - \frac{\left(N_t^{(c)}\right)'}{\left(N_t^{(c)}\right)^2}\left(\boldsymbol{x}_t\,\boldsymbol{x}_t^T - \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)}\right), & \text{falls } c_t = c \in \{1,\ldots,M\}, \\ \mathbf{0}, & \text{falls } c_t = c = M+1, \end{cases} \tag{4.71}$$

$$\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)' = \begin{cases} \left(\tilde{\boldsymbol{\Sigma}}_{t-1}^{(c)}\right)', & \text{falls } c_t \neq c, \\ \left(\tilde{\boldsymbol{\Pi}}_t^{(c)}\right)' - \left(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)'\left(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)^T - \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\left(\left(\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right)'\right)^T, & \text{falls } c_t = c \in \{1,\ldots,M\}, \\ \mathbf{0}, & \text{falls } c_t = c = M+1, \end{cases} \tag{4.72}$$

$$\left(N_t^{(c)}\right)' = \begin{cases} \left(N_{t-1}^{(c)}\right)', & \text{falls } c_t \neq c, \\ \lambda^{(c)}_{(n_{t-1}^{(c)})}\left(N_{t-1}^{(c)}\right)' + N_{t-1}^{(c)}, & \text{falls } c_t = c \in \{1,\ldots,M\}, \\ 0, & \text{falls } c_t = c = M+1. \end{cases} \tag{4.73}$$

- Gradient der negativen NLL:

$$\left(J_t^{(c)}\right)' = \begin{cases} \left(J_{t-1}^{(c)}\right)', & \text{falls } c_t \neq c, \\ \text{(4.43) bzw. (4.61) mit Index } t, & \text{falls } c_t = c \in \{1,\ldots,M\}, \\ 0, & \text{falls } c_t = c = M+1. \end{cases}$$

- Faktor:

$$\lambda^{(c)}_{(n^{(c)}_t)} = \begin{cases} \lambda^{(c)}_{(n^{(c)}_{t-1})}, & \text{falls } c_t \neq c, \\ (4.49) \text{ mit Index } t, & \text{falls } c_t = c \in \{1, \ldots, M\}, \\ \text{Zufallszahl aus } [\lambda_-, \lambda_+], & \text{falls } c_t = c = M+1. \end{cases}$$

- Schrittweite:

$$\alpha^{(c)}_t = \begin{cases} \alpha^{(c)}_{t-1}, & \text{falls } c_t \neq c, \\ (4.50), & \text{falls } c_t = c \in \{1, \ldots, M\}, \\ \text{Zufallszahl aus } [\alpha_{\min}, \alpha_{\max}], & \text{falls } c_t = c = M+1. \end{cases}$$

Rechenbeispiele für den Algorithmus *QDA-AF* und *LDA-AF* auf einem eindimensionalen Datenstrom sind in Anhang A.3 bzw. A.4 zu finden.

4.5 Zusammenfassung: exakt vs. approximativ

Bei der *Sequential Incremental LDA* und *Chunk Incremental LDA* (vgl. Abschnitt 4.2) handelt es sich jeweils um eine exakte Methode. Dies bedeutet, dass es sich bei allen Formeln für die Größen, die bei der Methode schrittweise aktualisiert werden, um die Klassifikationsregel anzupassen, um exakte Updates handelt. Nach jedem Aktualisierungsschritt sind die Parameter identisch zu jenen, die bei einem LDA-Modell resultieren, welches auf allen Beobachtungen gemeinsam angepasst wird (Batch-Methode der linearen Diskriminanzanalyse nach Fisher). Folglich sind auch die Klassifikationsregel und Prognosen für neue Beobachtungen nach jedem Aktualisierungsschritt identisch zu denen der Batch-Methode der Fisher LDA. Es erfolgt keine Anpassung an einen eventuell vorliegenden concept drift im Datenstrom, alle Beobachtungen fließen mit identischem Gewicht in die Klassifikationsregel ein.

Ebenso handelt es sich bei den einzelnen Schritten von *OLDC (Online Linear Discriminant Classifier)* (vgl. Abschnitt 4.3) um exakte Updates. In dem Fall einer festen Lernrate von $\lambda = 1/2$ ist das resultierende Modell sogar äquivalent zu jenem der Batch-Methode der Kanonischen LDA. Wird eine andere feste Lernrate oder eine adaptive Lernrate betrachtet, so sind die einzelnen Aktualisierungsschritte bzw. Update-Formeln zwar exakt, das resultierende Klassifikationsmodell nach jedem Aktualisierungsschritt ist jedoch nicht identisch zu jenem basierend auf allen Beobachtungen. Durch die Lernrate werden Gewichtungen für die einzelnen Beobachtungen bei den Updates eingeführt, sodass diese nicht mehr mit identischem Gewicht in die Klassifikationsregel einfließen. Dadurch kann eine Anpassung an einen eventuell vorliegenden concept drift erfolgen.

Die *Online Diskriminanzanalyse mit adaptivem Vergessen* (vgl. Abschnitt 4.4) ist keine exakte Methode. Erstens findet eine Anpassung an einen eventuell vorliegenden concept drift statt, indem eine Summe von Likelihood Termen betrachtet wird, bei der Summanden zeitlich vergangener Beobachtungen ein zunehmend geringeres Gewicht erhalten. Daher sind die einzelnen Parameter bzw. ist das resultierende Klassifikationsmodell nach jedem Aktualisierungsschritt nicht identisch zu jenem basierend auf allen Beobachtungen. Das Modell ist demnach in keinem Fall äquivalent zu jenem der Batch-Methode der Kanonischen LDA, da die Faktoren λ sich im Laufe der Aktualisierungen durch neue Beobachtungen selbst adaptiv anpassen (*self-tuning*) und diese nicht für den gesamten Datenstrom festgesetzt werden können wie zum Beispiel bei *OLDC*.

Innerhalb der Teilalgorithmen *G-AF* und *M-AF* handelt es sich bei den Update-Formeln für $N_t^{(0)}$, $\tilde{P}_t^{(c)}$, $\lambda_t^{(0)}$, $\alpha_t^{(0)}$, $N_t^{(c)}$, $\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}$, $\tilde{\boldsymbol{\Pi}}_t^{(c)}$, $\tilde{\boldsymbol{\Sigma}}_t^{(c)}$, $\lambda_{(n_t^{(c)})}^{(c)}$, $\alpha_t^{(c)}$ sowie bei *LDA-AF* zusätzlich $N_t^{(P)}$, $\tilde{\boldsymbol{\Pi}}_t^{(P)}$, $\tilde{\boldsymbol{\Sigma}}_t^{(P)}$, $\lambda_t^{(P)}$, $\alpha_t^{(P)}$ um exakte Updates in dem Sinne, dass keine Zufallszahlen einwirken oder die Formeln approximativ sind. Die Summe von Likelihood Termen $\mathcal{L}^{(\boldsymbol{\lambda})}$ aus (4.33) kann nach $\boldsymbol{\mu}^{(c)}$ und $\boldsymbol{\Sigma}^{(c)}$ bzw. jene aus (4.53) nach $p^{(c)}$ abgeleitet werden, um die entsprechenden ML-Schätzer bei Betrachtung von exponentiellem Vergessen zu erhalten. Auch die rekursiven Formeln für die Schätzer lassen sich exakt aus den Batch Varianten herleiten.

Da die Faktoren $\lambda^{(c)} \neq \lambda_{(j)}^{(c)}\ \forall j,\ \lambda^{(0)} \neq \lambda_t^{(0)},\ \lambda^{(P)} \neq \lambda_t^{(P)}\ \forall t$ und damit nicht konstant für alle Zeitpunkte sind, sondern sich selbst adaptiv anpassen, handelt es sich bei allen Ableitungen nach $\lambda^{(c)}$ bzw. $\lambda^{(0)}$ oder $\lambda^{(P)}$ nur um approximative Annäherungen. $\left(J_{t+1}^{(c)}\right)'$ bzw. $\left(J_{t+1}^{(0)}\right)'$ sowie $\left(J_{t+1}^{(P)}\right)'$ (bei *LDA-AF*) sind demnach approximative Gradienten. Ebenso sind die folgenden Gradienten in den Algorithmen lediglich approximativ, da diese unter Annahme eines konstanten Faktors $\lambda^{(c)}$, $\lambda^{(0)}$ oder $\lambda^{(P)}$ hergeleitet werden: $\left(N_t^{(0)}\right)'$, $\left(\tilde{P}_t^{(c)}\right)'$, $\left(N_t^{(c)}\right)'$, $\left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)'$, $\left(\tilde{\boldsymbol{\Pi}}_t^{(c)}\right)'$ und $\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)'$ sowie bei *LDA-AF* zusätzlich $\left(N_t^{(P)}\right)'$, $\left(\tilde{\boldsymbol{\Pi}}_t^{(P)}\right)'$ und $\left(\tilde{\boldsymbol{\Sigma}}_t^{(P)}\right)'$. Nur im Falle unveränderlicher Faktoren $\lambda^{(c)} := \lambda_{(j)}^{(c)}$ bzw. $\lambda^{(0)} := \lambda_t^{(0)}$ oder $\lambda^{(P)} := \lambda_t^{(P)}$ ist die Gradientenbildung exakt.

Zusätzlich ist zu beachten, dass für die Schätzer für $d_t^{(c)} := \log\left|\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right|$ und $\boldsymbol{G}_t^{(c)} := \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}$ im Datenstrom approximative Startwerte von $d_0 = -\delta$ und $\boldsymbol{G}_0 = \delta \boldsymbol{I}_{p\times p}$ mit großem δ verwendet werden. Der Grund liegt darin, dass für (4.38)–(4.41) die Startwerte so gewählt werden, dass die Schätzer im Falle fester Faktoren $\lambda^{(c)} := \lambda_{(j)}^{(c)} := 1$ den bekannten ML-Schätzern gleichen. In diesem Fall ist zu Beginn jedoch $\boldsymbol{\Sigma}_0 = \boldsymbol{0}$, weswegen die Inverse und Determinante nicht gebildet werden können und Approximationen verwendet werden.

5 Erweiterung der Methoden

Im Folgenden wird hergeleitet wie sich die Methode *OLDC* von Kuncheva und Plumpton (2008) (vgl. Abschnitt 4.3) sowie die Online Diskriminanzanalyse mit adaptivem Vergessen von Anagnostopoulos et al. (2012) (vgl. Abschnitt 4.4) auf die Betrachtung von Chunks erweitern lässt. Anstelle von Updates der interessierenden Größen für die LDA durch jeweils eine einzelne neue Beobachtung ist es mit den im Folgenden entwickelten Formeln möglich eine ganze Reihe von neuen Beobachtungen bei der Aktualisierung gleichzeitig zu betrachten. In Abschnitt 5.1 wird zunächst der Fall ohne Lernrate λ bei *OLDC* betrachtet. Darauf aufbauend wird in Abschnitt 5.2 zusätzlich die Lernrate in die Formeln integriert. Abschnitt 5.3 befasst sich mit der Erweiterung von *LDA-AF* und *QDA-AF*.

5.1 Erweiterung der Methode OLDC auf Chunks

In den Formeln der Chunk Methode werden im Gegensatz zur Aktualisierung durch eine einzelne Beobachtung nicht t und $t+1$ betrachtet, sondern allgemein die Zeitpunkte t_1 und t_2, wobei $t_2 > t_1$ (vgl. Seite 61). Es sei zudem n_{t_1} die Gesamtanzahl der Beobachtungen zum ersten Zeitpunkt t_1 und n_{t_2} jene nach Update durch einen Chunk von neuen Beobachtungen, also zum Zeitpunkt t_2. Mit $n_{t_1:t_2}$ wird die Anzahl der Beobachtungen aus dem neuen Chunk bezeichnet. Es gilt dabei: $n_{t_1} + n_{t_1:t_2} = n_{t_2}$. Für die Anzahl der Beobachtungen einer Klasse c wird analog die gleiche Notation verwendet, wobei ein (c) im Exponent bezeichnet, dass es sich um die Klasse c handelt. Zudem sei wie bei der *Chunk Incremental LDA* aus Abschnitt 4.2 erneut im Folgenden mit $T := \{t_1 + 1, \ldots, t_2\}$ die Menge der Zeitpunkte bzw. Beobachtungen des aktuellen Chunks bezeichnet.

Herleitung der Mittelwertvektoren und Schätzer für a-priori Wahrscheinlichkeiten

Der Mittelwertvektor $\boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}}$ der Klasse c lässt sich durch einen Chunk der Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$ zum Zeitpunkt t_2 folgendermaßen aktualisieren (vgl. (4.15)):

$$\boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} = \begin{cases} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}}, & \text{falls } \forall\, i \in T : g(\boldsymbol{x}_i) \neq c, \\ \dfrac{n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}}{n^{(c)}_{t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \in \{1, \ldots, M\}, \\ \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \notin \{1, \ldots, M\}. \end{cases} \tag{5.1}$$

Dazu müssen zunächst die neuen Mittelwertvektoren $\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}$, $c = 1, \ldots, M[, \ldots]$, des Chunks mithilfe der Formel (4.16) von Seite 66 wie bei der Methode *Chunk Incremental LDA* bestimmt und die Anzahl der neuen Beobachtungen in jeder Klasse $n^{(c)}_{t_1:t_2}$ erfasst werden.

Zur Schätzung der a-priori Wahrscheinlichkeiten der einzelnen Klassen können (Update-) Formeln für die relativen Häufigkeiten herangezogen werden, die jeweils auf der absoluten Anzahl an Beobachtungen aus dem vorherigen Aktualisierungsschritt basieren (vgl. (4.18) von *Chunk ILDA*):

$$P^{(c)}_{t_2} = \frac{n^{(c)}_{t_2}}{n_{t_2}} = \begin{cases} \dfrac{n^{(c)}_{t_1}}{n_{t_1} + n_{t_1:t_2}}, & \text{falls } \forall\, i \in T : g(\boldsymbol{x}_i) \neq c, \\ \dfrac{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}}{n_{t_1} + n_{t_1:t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \in \{1, \ldots, M\}, \\ \dfrac{n^{(c)}_{t_1:t_2}}{n_{t_1} + n_{t_1:t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \notin \{1, \ldots, M\}. \end{cases} \tag{5.2}$$

Zusätzlich zur Anzahl der Beobachtungen in jeder Klasse wird hier die Größe des Chunks $n_{t_1:t_2}$ benötigt.

Herleitung der aktualisierten inversen Kovarianzmatrix

Die inverse Kovarianzmatrix lässt sich durch einen Chunk der Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$ zum Zeitpunkt t_2 folgendermaßen mithilfe der Sherman-Morrison-Woodbury Formel aus Satz 2 (s. Seite 71) aktualisieren:

$$\boldsymbol{S}^{-1}_{t_2} = \frac{n_{t_2}}{n_{t_1}} \left(\boldsymbol{S}_{t_1} + \boldsymbol{U}\boldsymbol{V}^T\right)^{-1} = \frac{n_{t_2}}{n_{t_1}} \left(\boldsymbol{S}^{-1}_{t_1} - \boldsymbol{S}^{-1}_{t_1}\boldsymbol{U}\left(\boldsymbol{I}_{d\times d} + \boldsymbol{V}^T\boldsymbol{S}^{-1}_{t_1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}^T\boldsymbol{S}^{-1}_{t_1}\right).$$

Dabei sind die Matrizen $\boldsymbol{U}$ und $\boldsymbol{V}$ wie in der folgenden Formel (5.9) zu wählen. Die gesamte Herleitung wird im Folgenden erläutert.

Kuncheva und Plumpton (2008) nutzen die Sherman-Morrison-Woodbury Formel für ein Rang-1-Update, um direkt die inverse Kovarianzmatrix in jedem Schritt zu aktualisieren, wenn eine einzelne neue Beobachtung betrachtet wird (vgl. Satz 3 auf Seite 71). Wird eine ganze Reihe von neuen Beobachtungen betrachtet (Chunk), so entspricht dies analog einem Update höheren Rangs bei der Bestimmung der neuen inversen Kovarianzmatrix. Die Sherman-Morrison-Woodbury Formel für Änderungen vom Rang d wurde in Satz 2 auf Seite 71 vorgestellt. Zur Erinnerung (Golub und Van Loan, 1996, S. 50):

Für eine reguläre $(p \times p)$-Matrix $\boldsymbol{S}$ und $(p \times d)$-Matrizen $\boldsymbol{U}, \boldsymbol{V}$, wobei $\boldsymbol{I}_{d\times d} + \boldsymbol{V}^T\boldsymbol{S}^{-1}\boldsymbol{U}$ nicht-singulär sei, gilt:

$$\left(\boldsymbol{S} + \boldsymbol{U}\boldsymbol{V}^T\right)^{-1} = \boldsymbol{S}^{-1} - \boldsymbol{S}^{-1}\boldsymbol{U}\left(\boldsymbol{I}_{d\times d} + \boldsymbol{V}^T\boldsymbol{S}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}^T\boldsymbol{S}^{-1}. \tag{5.3}$$

Anhand dieser Formel ist bereits zu sehen, dass das Problem der Aktualisierung der Inversen der Kovarianzmatrix in diesem Fall nicht zwingend weniger komplex bei Nutzung der rechten Seite anstelle der linken Seite der Gleichung ist. Im Gegensatz zur Formel (4.25) beim Rang-1-Update lässt sich nämlich hier eine Invertierung nicht gänzlich vermeiden. Anstelle der Invertierung der aktualisierten Kovarianzmatrix bzw. allgemein $\boldsymbol{S} + \boldsymbol{U}\boldsymbol{V}^T$ muss der Term $\boldsymbol{I}_{d\times d} + \boldsymbol{V}^T\boldsymbol{S}^{-1}\boldsymbol{U}$ invertiert werden. Beim Rang-1-Update vereinfacht sich dies zu einer Division.

An dieser Stelle stellt sich demnach die Frage, wie $\boldsymbol{U}$ und $\boldsymbol{V}$ aussehen können, damit die Auswertung der rechten Seite von (5.3) weniger komplex und teuer ist als die Auswertung der linken Seite.

Zunächst wird dazu die Kovarianzmatrix $\boldsymbol{S}_{t_2}$ zum Zeitpunkt t_2 bei Update durch einen Chunk von $n_{t_1:t_2}$ neuen Beobachtungen hergeleitet. Es wird dabei über alle möglichen Klassen c zum Zeitpunkt t_2 summiert, wobei nun mehr als M Klassen vorliegen können, wenn die Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$ des aktuellen Chunks aus (teilweise) neuen Klassen stammen:

$$
\begin{aligned}
\boldsymbol{S}_{t_2} &= \frac{1}{n_{t_2}} \sum_c \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t_1}} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right)^T \right) \right. \\
&\qquad \left. + \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ t_1<i\le t_2}} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right)^T \right) \right) \\
&= \frac{1}{n_{t_2}} \sum_c \left(\underbrace{\sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right)}_{\text{Umformung in (B.15)}} \right. \\
&\qquad \left. + \underbrace{\sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right)}_{\text{Umformung in (B.16)}} \right) \\
&\overset{\text{(B.15)/(B.16)}}{=} \frac{1}{n_{t_2}} \sum_c \left(\sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right. \\
&\qquad + \frac{n^{(c)}_{t_1} \left(n^{(c)}_{t_1:t_2} \right)^2}{\left(n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2} \right)^2} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \\
&\qquad + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&\qquad \left. + \frac{\left(n^{(c)}_{t_1} \right)^2 n^{(c)}_{t_1:t_2}}{\left(n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2} \right)^2} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \right). \qquad (5.4)
\end{aligned}
$$

Da, wie oben bereits erwähnt,

$$n_{t_2} = n_{t_1} + n_{t_1:t_2} \quad \text{und} \quad n_{t_2}^{(c)} = n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)}$$

gilt, folgt insgesamt:

$$\begin{aligned}
\boldsymbol{S}_{t_2} = \frac{1}{n_{t_2}} \sum_c \Bigg(& \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
& + \frac{n_{t_1}^{(c)} \left(n_{t_1:t_2}^{(c)} \right)^2}{\left(n_{t_2}^{(c)} \right)^2} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \\
& + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
& + \frac{\left(n_{t_1}^{(c)} \right)^2 n_{t_1:t_2}^{(c)}}{\left(n_{t_2}^{(c)} \right)^2} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \Bigg) \\
= \frac{1}{n_{t_2}} \sum_c \Bigg(& \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
& + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
& + \frac{\left(n_{t_1}^{(c)} \right)^2 n_{t_1:t_2}^{(c)} + n_{t_1}^{(c)} \left(n_{t_1:t_2}^{(c)} \right)^2}{\left(n_{t_2}^{(c)} \right)^2} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \Bigg) \\
= \frac{1}{n_{t_2}} \sum_c \Bigg(& \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
& + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
& + \frac{\left(n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)} \right) n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{\left(n_{t_2}^{(c)} \right)^2} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \Bigg) \\
= \frac{1}{n_{t_2}} \sum_c \Bigg(& \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
& + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
& + \frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_2}^{(c)}} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \Bigg)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n_{t_2}} \left(\sum_c \left(\sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right) \right. \\
&\qquad + \sum_c \left(\sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right. \\
&\qquad\qquad \left. \left. + \frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \right) \right) \\
&\overset{\substack{n^{(c)}_{t_1}=0,\\ c>M}}{=} \frac{1}{n_{t_2}} \left(\sum_{c=1}^{M} \left(\sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right) \right. \\
&\qquad + \sum_c \left(\sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right. \\
&\qquad\qquad \left. \left. + \frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \right) \right) \\
&= \frac{n_{t_1}}{n_{t_2}} \left(\boldsymbol{S}_{t_1} + \frac{1}{n_{t_1}} \sum_c \left(\frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \right. \right. \\
&\qquad\qquad \left. \left. + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right) \right).
\end{aligned}
\tag{5.5}
$$

Zur Anwendung der Sherman-Morrison-Woodbury Formel (5.3) sind nun zwei Matrizen $\boldsymbol{U} \in \mathbb{R}^{p \times d}$ und $\boldsymbol{V} \in \mathbb{R}^{p \times d}$ zu finden, sodass:

$$
\begin{aligned}
\boldsymbol{U}\boldsymbol{V}^T &\overset{!}{=} \frac{1}{n_{t_1}} \sum_c \left(\frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \right. \\
&\qquad \left. + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right) \\
&= \sum_c \left(\frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n_{t_1} n^{(c)}_{t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \right. \\
&\qquad \left. + \sum_{i=t_1+1}^{t_2} \left(\frac{1}{n_{t_1}} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right).
\end{aligned}
\tag{5.6}
$$

Für zwei $(p \times d)$-Matrizen $\boldsymbol{U}$ und $\boldsymbol{V}$ gilt:

$$
\begin{aligned}
\boldsymbol{U}\boldsymbol{V}^T &= \begin{pmatrix} u_{11} & \dots & u_{1d} \\ \vdots & & \vdots \\ u_{p1} & \dots & u_{pd} \end{pmatrix} \begin{pmatrix} v_{11} & \dots & v_{p1} \\ \vdots & & \vdots \\ v_{1d} & \dots & v_{pd} \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_1 & \dots & \boldsymbol{u}_d \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_d^T \end{pmatrix} \\
&= \boldsymbol{u}_1 \boldsymbol{v}_1^T + \ldots + \boldsymbol{u}_d \boldsymbol{v}_d^T = \sum_{i=1}^{d} \boldsymbol{u}_i \boldsymbol{v}_i^T.
\end{aligned}
\tag{5.7}
$$

Die Sherman-Morrison-Woodbury Formel (5.3) für Änderungen vom Rang d lässt sich somit alternativ formulieren:

$$\begin{aligned}&\left(\boldsymbol{S}+\boldsymbol{U}\boldsymbol{V}^T\right)^{-1} \overset{(5.7)}{=} \left(\boldsymbol{S}+\left(\boldsymbol{u}_1\boldsymbol{v}_1^T+\ldots+\boldsymbol{u}_d\boldsymbol{v}_d^T\right)\right)^{-1}\\ &\overset{(5.3)}{=} \boldsymbol{S}^{-1}-\boldsymbol{S}^{-1}\left(\boldsymbol{u}_1\ldots\boldsymbol{u}_d\right)\left(\boldsymbol{I}_{d\times d}+\left(\boldsymbol{v}_1^T\ldots\boldsymbol{v}_d^T\right)^T\boldsymbol{S}^{-1}\left(\boldsymbol{u}_1\ldots\boldsymbol{u}_d\right)\right)^{-1}\left(\boldsymbol{v}_1^T\ldots\boldsymbol{v}_d^T\right)^T\boldsymbol{S}^{-1}.\end{aligned}$$

Zur Bestimmung der Matrizen $\boldsymbol{U}$ und $\boldsymbol{V}$ wird die Summe (5.6) in ihre einzelnen Summanden aufgeteilt:

$$\begin{aligned}\boldsymbol{U}\boldsymbol{V}^T=&\left(\frac{n_{t_1}^{(1)}n_{t_1:t_2}^{(1)}}{n_{t_1}n_{t_2}^{(1)}}\left(\boldsymbol{m}_{n_{t_1}^{(1)}}^{(1)}-\boldsymbol{m}_{n_{t_1:t_2}^{(1)}}^{(1)}\right)\left(\boldsymbol{m}_{n_{t_1}^{(1)}}^{(1)}-\boldsymbol{m}_{n_{t_1:t_2}^{(1)}}^{(1)}\right)^T\right.\\ &\left.+\sum_{i=t_1+1}^{t_2}\left(\frac{1}{n_{t_1}}\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_{t_1:t_2}^{(1)}}^{(1)}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_{t_1:t_2}^{(1)}}^{(1)}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=1\}}\right)\right)\\ &+\ldots+\\ &\left(\frac{n_{t_1}^{(M)}n_{t_1:t_2}^{(M)}}{n_{t_1}n_{t_2}^{(M)}}\left(\boldsymbol{m}_{n_{t_1}^{(M)}}^{(M)}-\boldsymbol{m}_{n_{t_1:t_2}^{(M)}}^{(M)}\right)\left(\boldsymbol{m}_{n_{t_1}^{(M)}}^{(M)}-\boldsymbol{m}_{n_{t_1:t_2}^{(M)}}^{(M)}\right)^T\right.\\ &\left.+\sum_{i=t_1+1}^{t_2}\left(\frac{1}{n_{t_1}}\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_{t_1:t_2}^{(M)}}^{(M)}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_{t_1:t_2}^{(M)}}^{(M)}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=M\}}\right)\right)\\ &+\sum_{c=M+1}^{M+m}\sum_{i=t_1+1}^{t_2}\left(\frac{1}{n_{t_1}}\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right).\end{aligned}\tag{5.8}$$

Der jeweils zweite bzw. die letzten m Summanden (bei m neuen Klassen)

$$\sum_{i=t_1+1}^{t_2}\left(\frac{1}{n_{t_1}}\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right),\quad c=1,\ldots,M+m,$$

bestehen aus $n_{t_1:t_2}^{(c)}$ einzelnen Summanden.

Allerdings muss im neuen Chunk nicht jede der M bisherigen Klassen auftreten. Die gesamte Summe (5.8) lässt sich demnach in höchstens

$$1+n_{t_1:t_2}^{(1)}+\ldots+1+n_{t_1:t_2}^{(M)}+\sum_{c=M+1}^{M+m}n_{t_1:t_2}^{(c)}=M+\sum_{c=1}^{M+m}n_{t_1:t_2}^{(c)}$$

einzelne Summanden aufteilen. Genauer:

$$M+\sum_{c=1}^{M+m}n_{t_1:t_2}^{(c)}\geq\sum_{c=1}^{M}\left(\mathbb{1}\left(\left(\sum_{j=t_1+1}^{t_2}\mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}}\right)\neq 0\right)\right)+\underbrace{\sum_{c=1}^{M+m}n_{t_1:t_2}^{(c)}}_{=n_{t_1:t_2}},$$

wobei $n_{t_1:t_2}^{(c)}=0$ für Klassen c, die nicht auftreten.

Die Matrizen $\boldsymbol{U}$ und $\boldsymbol{V}$ sind in diesem Fall identisch, da für alle einzelnen Summanden in (5.8) $\boldsymbol{u}_i\boldsymbol{u}_i^T$ gilt. Mithilfe von (5.7) lässt sich die Matrix $\boldsymbol{U} =: \boldsymbol{V}$ aus den einzelnen Summanden wie folgt in (5.9) aufstellen. Bezeichne dazu mit $\boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_{t_1:t_2}^{(c)})}^{(c)}$ die $n_{t_1:t_2}^{(c)}$ Beobachtungen aus $\boldsymbol{x}_j$, $j = t_1+1, \ldots, t_2$, die im neuen Chunk jeweils in Klasse c realisiert werden. Für diese gilt in der Summe (5.8) also $\mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}} = 1$.

Es gilt:

$$\begin{aligned} \boldsymbol{U} &= \begin{pmatrix} \boldsymbol{U}_1 & \ldots & \boldsymbol{U}_{M+m} \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_1 & \ldots & \boldsymbol{u}_d \end{pmatrix} = \boldsymbol{V}, \\ \boldsymbol{V}^T &= \begin{pmatrix} \boldsymbol{V}_1^T & \ldots & \boldsymbol{V}_{M+m}^T \end{pmatrix}^T = \begin{pmatrix} \boldsymbol{v}_1^T & \ldots & \boldsymbol{v}_d^T \end{pmatrix}^T = \boldsymbol{U}^T, \end{aligned} \tag{5.9}$$

wobei $\boldsymbol{U}$ und $\boldsymbol{V}$ $p \times \left(\sum_{c=1}^{M} \left(\mathbb{1} \left(\left(\sum_{j=t_1+1}^{t_2} \mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}} \right) \neq 0 \right) \right) + n_{t_1:t_2} \right)$-Matrizen sind und

$$\boldsymbol{U}_c := \begin{cases} \Bigg(\underbrace{\sqrt{\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_1} n_{t_2}^{(c)}}} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)}_{=:\boldsymbol{u}_1=:\boldsymbol{v}_1 \text{ für } c=1}, \\ \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(1)}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right), \ldots, \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(n_{t_1:t_2}^{(c)})}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \Bigg), & c = 1, \ldots, M, \\ \Bigg(\sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(1)}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right), \ldots, \underbrace{\sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(n_{t_1:t_2}^{(c)})}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)}_{=:\boldsymbol{u}_d=:\boldsymbol{v}_d \text{ für } c=M+m} \Bigg), & c = M+1, \ldots, M+m, \end{cases}$$

$$\boldsymbol{V}_c^T := \begin{cases} \Bigg(\underbrace{\sqrt{\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_1} n_{t_2}^{(c)}}} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T}_{=:\boldsymbol{v}_1^T=:\boldsymbol{u}_1^T \text{ für } c=1}, \\ \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(1)}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T, \ldots, \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(n_{t_1:t_2}^{(c)})}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \Bigg)^T, & c = 1, \ldots, M, \\ \Bigg(\sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(1)}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T, \ldots, \underbrace{\sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(n_{t_1:t_2}^{(c)})}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T}_{=:\boldsymbol{v}_d^T=:\boldsymbol{u}_d^T \text{ für } c=M+m} \Bigg)^T, & c = M+1, \ldots, M+m. \end{cases}$$

Mit diesen Matrizen $\boldsymbol{U}$ und $\boldsymbol{V}$ gilt (vgl. (5.6)):

$$\begin{aligned} \boldsymbol{U}\boldsymbol{V}^T = \sum_c \Bigg(& \frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_1} n_{t_2}^{(c)}} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \\ & + \sum_{i=t_1+1}^{t_2} \left(\frac{1}{n_{t_1}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \Bigg). \end{aligned}$$

Nach Formulierung der Kovarianzmatrix durch $\boldsymbol{S}_{t_2} = \frac{n_{t_1}}{n_{t_2}} \left(\boldsymbol{S}_{t_1} + \boldsymbol{U}\boldsymbol{V}^T \right)$ (vgl. (5.5)) lässt sich die neue Inverse mithilfe der Sherman-Morrison-Woodbury Formel (5.3) aktualisieren durch

$$\boldsymbol{S}_{t_2}^{-1} = \frac{n_{t_2}}{n_{t_1}} \left(\boldsymbol{S}_{t_1}^{-1} - \boldsymbol{S}_{t_1}^{-1}\boldsymbol{U} \left(\boldsymbol{I}_{d\times d} + \boldsymbol{V}^T\boldsymbol{S}_{t_1}^{-1}\boldsymbol{U} \right)^{-1} \boldsymbol{V}^T\boldsymbol{S}_{t_1}^{-1} \right), \tag{5.10}$$

wobei $\boldsymbol{I}_{d\times d}$ die Einheitsmatrix mit folgender Dimension d ist:

$$d := \underbrace{\sum_{c=1}^{M} \left(\mathbb{1} \left(\left(\sum_{j=t_1+1}^{t_2} \mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}} \right) \neq 0 \right) \right)}_{\text{Anzahl aus den bisherigen } M \text{ Klassen im Chunk } \boldsymbol{x}_{t_1+1},\ldots,\boldsymbol{x}_{t_2}} + \underbrace{n_{t_1:t_2}}_{\text{Chunk-Größe}}.$$

Übergang auf die Aktualisierung durch eine einzelne neue Beobachtung

Wird nur eine einzelne neue Beobachtung $\boldsymbol{x}_{t_1+1}$ für das Update betrachtet, so kommt diese aus der Klasse $k \in \{1, \ldots, M+1\}$. Das bedeutet die Matrizen $\boldsymbol{U}$ und $\boldsymbol{V}^T$ aus (5.9) verringern sich zunächst zu $\boldsymbol{U}_k$ und $\boldsymbol{V}_k^T$ für $g(\boldsymbol{x}_{t_1+1}) = k$. Zudem vereinfacht sich der Mittelwertvektor der Klasse k des neuen Chunks zu $\boldsymbol{m}^{(k)}_{n^{(k)}_{t_1:t_2}} = \boldsymbol{x}_{t_1+1}$ sowie $\boldsymbol{x}^{(k)}_{(1)} = \boldsymbol{x}_{t_1+1}$, da $n^{(k)}_{t_1:t_2} = 1$. Außerdem gilt $n^{(k)}_{t_2} = n^{(k)}_{t_1} + 1$. Sei zunächst der Fall $k \in \{1, \ldots, M\}$ betrachtet. Dabei bleiben lediglich die folgenden Matrizen erhalten:

$$\begin{aligned}
\boldsymbol{U}_k &= \left(\sqrt{\frac{n^{(k)}_{t_1} n^{(k)}_{t_1:t_2}}{n_{t_1} n^{(k)}_{t_2}}} \left(\boldsymbol{m}^{(k)}_{n^{(k)}_{t_1}} - \boldsymbol{m}^{(k)}_{n^{(k)}_{t_1:t_2}} \right), \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}^{(k)}_{(1)} - \boldsymbol{m}^{(k)}_{n^{(k)}_{t_1:t_2}} \right) \right) \\
&= \left(\sqrt{\frac{n^{(k)}_{t_1}}{n_{t_1}\left(n^{(k)}_{t_1}+1\right)}} \left(\boldsymbol{m}^{(k)}_{n^{(k)}_{t_1}} - \boldsymbol{x}_{t_1+1} \right), \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{x}_{t_1+1} \right) \right), \\
\boldsymbol{V}_k^T &= \left(\sqrt{\frac{n^{(k)}_{t_1} n^{(k)}_{t_1:t_2}}{n_{t_1} n^{(k)}_{t_2}}} \left(\boldsymbol{m}^{(k)}_{n^{(k)}_{t_1}} - \boldsymbol{m}^{(k)}_{n^{(k)}_{t_1:t_2}} \right)^T, \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}^{(k)}_{(1)} - \boldsymbol{m}^{(k)}_{n^{(k)}_{t_1:t_2}} \right)^T \right)^T \\
&= \left(\sqrt{\frac{n^{(k)}_{t_1}}{n_{t_1}\left(n^{(k)}_{t_1}+1\right)}} \left(\boldsymbol{m}^{(k)}_{n^{(k)}_{t_1}} - \boldsymbol{x}_{t_1+1} \right)^T, \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{x}_{t_1+1} \right)^T \right)^T.
\end{aligned}$$

Der jeweils zweite Ausdruck fällt im Spezialfall also komplett weg, sodass nur

$$\begin{aligned}
\boldsymbol{U} &= \sqrt{\frac{n^{(k)}_{t_1}}{n_{t_1}\left(n^{(k)}_{t_1}+1\right)}} \left(\boldsymbol{m}^{(k)}_{n^{(k)}_{t_1}} - \boldsymbol{x}_{t_1+1} \right) \in \mathbb{R}^{p\times 1}, \\
\boldsymbol{V}^T &= \sqrt{\frac{n^{(k)}_{t_1}}{n_{t_1}\left(n^{(k)}_{t_1}+1\right)}} \left(\boldsymbol{m}^{(k)}_{n^{(k)}_{t_1}} - \boldsymbol{x}_{t_1+1} \right)^T \in \mathbb{R}^{1\times p}
\end{aligned}$$

betrachtet werden.

Die Anwendung der Sherman-Morrison-Woodbury Formel (5.3) resultiert in (vgl. (5.10))

$$\begin{aligned}
\boldsymbol{S}_{t_2}^{-1} &= \frac{n_{t_2}}{n_{t_1}} \left(\boldsymbol{S}_{t_1}^{-1} - \boldsymbol{S}_{t_1}^{-1} \boldsymbol{U} \left(\boldsymbol{I}_{1\times 1} + \boldsymbol{V}^T \boldsymbol{S}_{t_1}^{-1} \boldsymbol{U} \right)^{-1} \boldsymbol{V}^T \boldsymbol{S}_{t_1}^{-1} \right) = \frac{n_{t_2}}{n_{t_1}} \left(\boldsymbol{S}_{t_1}^{-1} - \frac{\boldsymbol{S}_{t_1}^{-1} \boldsymbol{U} \boldsymbol{V}^T \boldsymbol{S}_{t_1}^{-1}}{1 + \boldsymbol{V}^T \boldsymbol{S}_{t_1}^{-1} \boldsymbol{U}} \right) \\
&= \frac{n_{t_1}+1}{n_{t_1}} \left(\boldsymbol{S}_{t_1}^{-1} - \frac{\frac{n_{t_1}^{(k)}}{n_{t_1}\left(n_{t_1}^{(k)}+1\right)} \cdot \boldsymbol{S}_{t_1}^{-1} \left(\boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} - \boldsymbol{x}_{t_1+1} \right) \left(\boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} - \boldsymbol{x}_{t_1+1} \right)^T \boldsymbol{S}_{t_1}^{-1}}{1 + \frac{n_{t_1}^{(k)}}{n_{t_1}\left(n_{t_1}^{(k)}+1\right)} \left(\boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} - \boldsymbol{x}_{t_1+1} \right)^T \boldsymbol{S}_{t_1}^{-1} \left(\boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} - \boldsymbol{x}_{t_1+1} \right)} \right) \\
&= \frac{n_{t_1}+1}{n_{t_1}} \left(\boldsymbol{S}_{t_1}^{-1} - \frac{\boldsymbol{S}_{t_1}^{-1} \left(\boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} - \boldsymbol{x}_{t_1+1} \right) \left(\boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} - \boldsymbol{x}_{t_1+1} \right)^T \boldsymbol{S}_{t_1}^{-1}}{\frac{n_{t_1}\left(n_{t_1}^{(k)}+1\right)}{n_{t_1}^{(k)}} + \left(\boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} - \boldsymbol{x}_{t_1+1} \right)^T \boldsymbol{S}_{t_1}^{-1} \left(\boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} - \boldsymbol{x}_{t_1+1} \right)} \right) \\
&= \frac{n_{t_1}+1}{n_{t_1}} \left(\boldsymbol{S}_{t_1}^{-1} - \frac{\boldsymbol{S}_{t_1}^{-1} \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} \right) \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} \right)^T \boldsymbol{S}_{t_1}^{-1}}{\frac{n_{t_1}\left(n_{t_1}^{(k)}+1\right)}{n_{t_1}^{(k)}} + \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} \right)^T \boldsymbol{S}_{t_1}^{-1} \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{m}_{n_{t_1}^{(k)}}^{(k)} \right)} \right),
\end{aligned} \tag{5.11}$$

was der Anwendung der Sherman-Morrison-Woodbury Formel (4.25) (nach Vereinfachung) für Updates vom Rang 1 entspricht, wobei $t_1 = t$ und $t_1 + 1 = t_2 = t + 1$ (vgl. (4.26)).

Für den Fall $g(\boldsymbol{x}_{t_1+1}) = k = M + 1$ (neu auftretende Klasse) werden die Teile

$$\begin{aligned}
\boldsymbol{U}_{M+1} &:= \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(1)}^{(M+1)} - \boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)} \right) = \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)} \right) \\
&= \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{x}_{t_1+1} \right) = \boldsymbol{0}
\end{aligned}$$

und

$$\begin{aligned}
\boldsymbol{V}_{M+1}^T &:= \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{(1)}^{(M+1)} - \boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)} \right)^T = \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{m}_{n_{t_1:t_2}^{(M+1)}}^{(M+1)} \right)^T \\
&= \sqrt{\frac{1}{n_{t_1}}} \left(\boldsymbol{x}_{t_1+1} - \boldsymbol{x}_{t_1+1} \right)^T = \boldsymbol{0}^T
\end{aligned}$$

aus (5.9) betrachtet. Die Matrizen $\boldsymbol{U}$ und $\boldsymbol{V}^T$ vereinfachen sich somit zum Nullvektor. Die Anwendung der Sherman-Morrison-Woodbury Formel resultiert dann im Fall $k = M + 1$ in

$$\boldsymbol{S}_{t_2}^{-1} = \frac{n_{t_2}}{n_{t_1}} \left(\boldsymbol{S}_{t_1}^{-1} - \boldsymbol{S}_{t_1}^{-1} \boldsymbol{U} \left(\boldsymbol{I}_{1\times 1} + \boldsymbol{V}^T \boldsymbol{S}_{t_1}^{-1} \boldsymbol{U} \right)^{-1} \boldsymbol{V}^T \boldsymbol{S}_{t_1}^{-1} \right) = \frac{n_{t_2}}{n_{t_1}} \cdot \boldsymbol{S}_{t_1}^{-1},$$

was dem Update (4.26) der inversen Kovarianzmatrix im Falle einer neu auftretenden Klasse entspricht.

Abschätzung der Anzahl nötiger Rechenoperationen für inverse Kovarianzmatrix

In diesem Abschnitt erfolgt nun eine Abschätzung, welche Dimension $\boldsymbol{U}$ und $\boldsymbol{V}$ haben können, damit die Anwendung der Sherman-Morrison-Woodbury Formel zur Bestimmung der aktuellen inversen Kovarianzmatrix einen Effizienzvorteil darstellt. Es soll also die Frage beantwortet werden, in welchen Fällen die Auswertung der rechten Seite der Gleichung (5.3) weniger komplex und teuer ist als jene der linken Seite. Betrachtet wird sowohl der Fall des Updates durch eine einzelne neue Beobachtung, also ein Rang-1-Update der Kovarianzmatrix,

$$\boldsymbol{S}_{t+1}^{-1} \overset{(4.23)}{=} \begin{cases} \dfrac{n_t+1}{n_t}\left(\boldsymbol{S}_t + \dfrac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)} \cdot \boldsymbol{z}\boldsymbol{z}^T\right)^{-1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = k \in \{1,\dots,M\}, \\ \dfrac{n_t+1}{n_t} \cdot \boldsymbol{S}_t^{-1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = M+1, \end{cases} \tag{5.12}$$

$$\overset{(4.26)}{=} \begin{cases} \dfrac{n_t+1}{n_t}\left(\boldsymbol{S}_t^{-1} - \dfrac{\boldsymbol{S}_t^{-1}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{S}_t^{-1}}{\frac{n_t\left(n_t^{(k)}+1\right)}{n_t^{(k)}} + \boldsymbol{z}^T\boldsymbol{S}_t^{-1}\boldsymbol{z}}\right), & \text{falls } g(\boldsymbol{x}_{t+1}) = k \in \{1,\dots,M\}, \\ \dfrac{n_t+1}{n_t} \cdot \boldsymbol{S}_t^{-1}, & \text{falls } g(\boldsymbol{x}_{t+1}) = M+1 \end{cases} \tag{5.13}$$

mit $\boldsymbol{z} = \boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}$, als auch der Fall der Aktualisierung der Kovarianzmatrix durch ein ganzes Chunk von neuen Beobachtungen:

$$\begin{aligned} \boldsymbol{S}_{t_2}^{-1} &\overset{(5.5)}{=} \frac{n_{t_2}}{n_{t_1}}\left(\boldsymbol{S}_{t_1} + \frac{1}{n_{t_1}}\sum_{c=1}^{M+m}\left(\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_2}^{(c)}} \cdot \boldsymbol{z}_{\text{chunks}}\,\boldsymbol{z}_{\text{chunks}}^T \right.\right. \\ &\qquad \left.\left. + \sum_{i=t_1+1}^{t_2}\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_{t_1:t_2}^{(c)}}\right)\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_{t_1:t_2}^{(c)}}\right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right)\right)^{-1} \\ &\overset{(5.6)}{=} \frac{n_{t_2}}{n_{t_1}}\left(\boldsymbol{S}_{t_1} + \boldsymbol{U}\boldsymbol{V}^T\right)^{-1} \qquad (5.14) \\ &\overset{(5.10)}{=} \frac{n_{t_2}}{n_{t_1}}\left(\boldsymbol{S}_{t_1}^{-1} - \boldsymbol{S}_{t_1}^{-1}\boldsymbol{U}\left(\boldsymbol{I}_{d\times d} + \boldsymbol{V}^T\boldsymbol{S}_{t_1}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}^T\boldsymbol{S}_{t_1}^{-1}\right) \qquad (5.15) \end{aligned}$$

mit $\boldsymbol{z}_{\text{chunks}} := \boldsymbol{m}^{(c)}_{n_{t_1}^{(c)}} - \boldsymbol{m}^{(c)}_{n_{t_1:t_2}^{(c)}}$ und $\boldsymbol{U}$, $\boldsymbol{V}^T$ aus (5.9).

Zunächst werden dazu Grundlagen der elementaren Rechenoperationen bei Matrizenoperationen eingeführt:

Es seien $\boldsymbol{A} \in \mathbb{R}^{m\times m}$, $\boldsymbol{X} \in \mathbb{R}^{k\times m}$, $\boldsymbol{Y} \in \mathbb{R}^{m\times n}$ und $\boldsymbol{Z} \in \mathbb{R}^{m\times n}$ Matrizen sowie a ein Skalar. Tabelle 5.1 fasst die benötigten Anzahlen an elementaren Rechenoperationen für ausgewählte Matrix-Rechenoperationen zusammen.

Tabelle 5.1: Anzahl elementarer Rechenoperationen bei Matrizenoperationen.

	Additionen	Subtraktionen	Multiplikationen	Divisionen
$\boldsymbol{Y}+\boldsymbol{Z}$	mn	–	–	–
$\boldsymbol{Y}-\boldsymbol{Z}$	–	mn	–	–
$\boldsymbol{XY}$	$k(m-1)n$	–	kmn	–
$a\boldsymbol{X}$	–	–	km	–

Für die Invertierung einer reellwertigen Matrix $\boldsymbol{A}$ wird zum Beispiel in `R` bei Anwendung der Funktion `solve()` die LAPACK Routine `DGESV` zur Lösung eines linearen Gleichungssystems verwendet (Anderson et al., 1999). Nach The Numerical Algorithms Group Ltd (2012) werden dabei approximativ $\frac{2}{3}m^3+2m^2m=\frac{8}{3}m^3$ Operationen benötigt, wobei nicht zwischen den einzelnen Operationen aus Tabelle 5.1 unterschieden wird.

Die Transponierung einer Matrix benötigt keine der elementaren Rechenoperationen aus Tabelle 5.1. Hier werden stattdessen die Anzahlen an Vertauschungen betrachtet. Für die Bestimmung von $\boldsymbol{X}^T$ kann hergeleitet werden, dass $km-k-\left((k-1)^2+(k-1)\right)/2 = km-k-\left(k^2-k\right)/2 = k\left(2m-k-1\right)/2$ (für $k,m>1$, $k\leq m$) Vertauschungen benötigt werden. Falls eine der Dimensionen Eins ist, $\boldsymbol{X}$ also ein Vektor, werden hingegen keine Vertauschungen für die Transponierung benötigt.

Mithilfe der Zerlegung in elementare Rechenoperationen lässt sich die Bestimmung der aktualisierten inversen Kovarianzmatrix durch die Sherman-Morrison-Woodbury Formel und durch Invertierung der aktualisierten Kovarianzmatrix vergleichen.

Bei Aktualisierung der Kovarianzmatrix durch eine einzelne neue Beobachtung werden bei Nutzung der Sherman-Morrison-Woodbury Formel annähernd $8p^2+p+3$ elementare Rechenoperationen benötigt (bzw. $10p^2+5$, falls Additionen mit Faktor 2 betrachtet werden), während die direkte Invertierung der aktualisierten Kovarianzmatrix $\boldsymbol{S}_{t_2}$ approximativ $\frac{8}{3}p^3+4p^2+p+3$ Rechenoperationen (bzw. $\frac{8}{3}p^3+5p^2+p+5$) kostet. Dadurch ergibt sich bereits ab einer Dimension von $p=2$ ein Effizienzvorteil der Sherman-Morrison-Woodbury Formel, da

$$\begin{cases} \frac{8}{3}p^3+4p^2+p+3 > 8p^2+p+3 \Leftrightarrow 8p>12, & p\geq 2, \\ \frac{8}{3}p^3+4p^2+p+3 < 8p^2+p+3 \Leftrightarrow 8p<12, & p=1, \end{cases}$$

bzw. auch bei alternativer Sichtweise (Additionen mit Faktor 2)

$$\begin{cases} \frac{8}{3}p^3+5p^2+p+5 > 10p^2+5 \Leftrightarrow 8p^2-15p+3>0, & p\geq 2, \\ \frac{8}{3}p^3+5p^2+p+5 < 10p^2+5 \Leftrightarrow 8p^2-15p+3<0, & p=1. \end{cases}$$

Im Fall der Aktualisierung der inversen Kovarianzmatrix durch einen Chunk von neuen Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$ werden bei gleichwertiger Betrachtung aller elementarer Rechenoperationen $\frac{8}{3}d^3 + 4d^2p + 3d(2p^2 - p) + p^2 + 1$ Operationen bei Verwendung der Sherman-Morrison-Woodbury Formel benötigt und $\frac{8}{3}p^3 + p^2(2d+1) + 1$ Operationen ohne. Dies führt mithilfe von Simulationen für $p, d \leq 100$ annähernd zu den folgenden Ungleichungen für einen Vergleich der beiden Methoden:

$$\begin{cases} f_{\text{erg}}(d,p) > 0, & \text{falls } p \geq 2d + \left\lfloor \frac{d+1}{3} \right\rfloor \wedge d \leq 100 \text{ bzw. } d < \left\lfloor \frac{3p+1}{7} \right\rfloor \wedge p \leq 100, \\ f_{\text{erg}}(d,p) < 0, & \text{falls } p < 2d + \left\lfloor \frac{d+1}{3} \right\rfloor \wedge d \leq 100 \text{ bzw. } d \geq \left\lfloor \frac{3p+1}{7} \right\rfloor \wedge p \leq 100, \end{cases}$$

wobei $f_{\text{erg}}(d,p) = 8p^3 - 12dp^2 - 12d^2p + 9dp - 8d^3$.

Falls $f_{\text{erg}}(d,p) > 0$, so ist die Anwendung der Sherman-Morrison-Woodbury Formel von Vorteil, im anderen Fall die direkte Invertierung der aktualisierten Kovarianzmatrix.

Werden die Additionen erneut mit Faktor 2 betrachtet, so vergrößern sich die Anzahlen zu $\frac{8}{3}d^3 + 9dp^2 + 6d^2p - 6dp + 1$ (mit Sherman-Morrison-Woodbury Formel) und $\frac{8}{3}p^3 + 3dp^2 + p^2 + 1$ (ohne). Daraus ergibt sich die Formel $f_{\text{erg}}(d,p) = 8p^3 - 18dp^2 - 18d^2p + 18dp + 3p^2 - 8d^3$ für den Vergleich.

Die Anzahl der benötigten Rechenoperationen wird im Folgenden für jeweils beide Situationen (Update durch einzelne Beobachtung und durch Chunk) und beide Sichtweisen (gleichwertige Betrachtung der Rechenoperationen und Betrachtung der Additionen mit Faktor 2) hergeleitet.

Zunächst wird der Fall der Aktualisierung durch eine einzelne neue Beobachtung betrachtet. Mithilfe der Sherman-Morrison-Woodbury Formel lässt sich die neue inverse Kovarianzmatrix in diesem Fall durch (5.11) bzw. (5.13) bestimmen. Die Matrizen und Vektoren besitzen die folgenden Dimensionen: $\boldsymbol{S}_t^{-1} \in \mathbb{R}^{p \times p}$, $\boldsymbol{x}_{t+1} \in \mathbb{R}^{p \times 1}$, $\boldsymbol{m}^{(k)}_{n_t^{(k)}} \in \mathbb{R}^{p \times 1}$. Die Matrix $\boldsymbol{S}_t^{-1}$ ist aus dem vorherigen Iterationsschritt fest und muss nicht bestimmt werden. Zudem können sowohl $\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}$ als auch darauffolgend $\boldsymbol{S}_t^{-1}\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)$ einmalig berechnet und gespeichert werden und müssen für die Bestimmung von (5.13) nicht mehrfach bestimmt werden. Insgesamt ergeben sich daraus (vgl. Tabelle 5.2 (a))

- $p(p-1) + (p-1)p + (p-1) + p^2(1-1) + 1 + 1 + 1 = 2p^2 - p + 2$ Additionen,
- $p + p^2 = p^2 + p$ Subtraktionen,
- $p^2 + p^2 + p + p^2 + 1 + p^2 + p^2 = 5p^2 + p + 1$ Multiplikationen,
- $1 + 1 + 1 = 3$ Divisionen.

Da lediglich der Vektor $\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}$ transponiert wird, werden keine Tauschoperationen benötigt (vgl. Seite 107). Für eine Aufstellung der einzelnen Rechenoperationen, die sich jeweils durch die schrittweise Berechnung der Formeln ergibt, sei für alle folgenden Fälle auf Tabelle 5.2 verwiesen.

Tabelle 5.2: Schrittweise Aufstellung der einzelnen Rechenoperationen für die Aktualisierung durch
(a) *eine einzelne Beobachtung mit Sherman-Morrison-Woodbury Formel*:
$\frac{n_t+1}{n_t}\left(\boldsymbol{S}_t^{-1}-\frac{\boldsymbol{S}_t^{-1}\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T\boldsymbol{S}_t^{-1}}{\frac{n_t\left(n_t^{(k)}+1\right)}{n_t^{(k)}}+\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T\boldsymbol{S}_t^{-1}\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)}\right)$,
(b) *eine einzelne Beobachtung ohne Sherman-Morrison-Woodbury Formel*:
$\frac{n_t}{n_t+1}\left(\boldsymbol{S}_t+\frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)}\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T\right)$,
(c) *einen Chunk mithilfe der Sherman-Morrison-Woodbury Formel*:
$\frac{n_{t_2}}{n_{t_1}}\left(\boldsymbol{S}_{t_1}^{-1}-\boldsymbol{S}_{t_1}^{-1}\boldsymbol{U}\left(\boldsymbol{I}_{d\times d}+\boldsymbol{V}^T\boldsymbol{S}_{t_1}^{-1}\boldsymbol{U}\right)^{-1}\boldsymbol{V}^T\boldsymbol{S}_{t_1}^{-1}\right)$,
(d) *einen Chunk ohne Sherman-Morrison-Woodbury Formel*:
$\frac{n_{t_1}}{n_{t_2}}\left(\boldsymbol{S}_{t_1}+\boldsymbol{U}\boldsymbol{V}^T\right)$.

		Additionen	Subtraktionen	Multiplikationen	Divisionen
(a)	①: $\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}$		$1\cdot p$		
	②: $\boldsymbol{S}_{t_1}^{-1}\cdot$ ①	$p\cdot(p-1)\cdot 1$		$p\cdot p\cdot 1$	
	③: ①$^T\cdot\boldsymbol{S}_{t_1}^{-1}$	$1\cdot(p-1)\cdot p$		$1\cdot p\cdot p$	
	④: ③ $\cdot$ ①	$1\cdot(p-1)\cdot 1$		$1\cdot p\cdot 1$	
	⑤: ② $\cdot$ ③	$p\cdot(1-1)\cdot p$		$p\cdot 1\cdot p$	
	⑥: $n_t^{(k)}+1$	1			
	⑦: $n_t\cdot$ ⑥			1	
	⑧: ⑦$/n_t^{(k)}$				1
	⑨: ⑧ + ④	1			
	⑩: 1/⑨ $\cdot$ ⑤			$p\cdot p$	1
	⑪: $\boldsymbol{S}_t^{-1}-$ ⑩		$p\cdot p$		
	⑫: $n_{t_1}+1$	1			
	⑬: ⑫$/n_{t_1}$				1
	⑭: ⑬ $\cdot$ ⑪			$p\cdot p$	
(b)	①: $\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}$		$1\cdot p$		
	②: ① $\cdot$ ①^T	$p\cdot(1-1)\cdot p$		$p\cdot 1\cdot p$	
	③: $n_t^{(k)}+1$	1			
	④: $n_t\cdot$ ③			1	
	⑤: $n_t^{(k)}/$ ④				1
	⑥: ⑤ $\cdot$ ②			$p\cdot p$	
	⑦: $\boldsymbol{S}_t+$ ⑥	$p\cdot p$			
	⑧: n_t+1	1			
	⑨: $n_t/$ ⑧				1
	⑩: ⑨ $\cdot$ ⑦			$p\cdot p$	

Fortsetzung auf der nächsten Seite

		Additionen	Subtraktionen	Multiplikationen	Divisionen
(c)	①: $\boldsymbol{S}_{t_1}^{-1}\boldsymbol{U}$	$p\cdot(p-1)\cdot d$		$p\cdot p\cdot d$	
	②: $\boldsymbol{V}^T\boldsymbol{S}_{t_1}^{-1}$	$d\cdot(p-1)\cdot p$		$d\cdot p\cdot p$	
	③: ② $\cdot\,\boldsymbol{U}$	$d\cdot(p-1)\cdot d$		$d\cdot p\cdot d$	
	④: $\boldsymbol{I}_{d\times d}+$ ③	$d\cdot d$			
	⑤: ① $\cdot\,($④$)^{-1}$	$p\cdot(d-1)\cdot d$		$p\cdot d\cdot d$	
	⑥: ⑤ $\cdot$ ②	$p\cdot(d-1)\cdot p$		$p\cdot d\cdot p$	
	⑦: $\boldsymbol{S}_{t_1}^{-1}-$ ⑥		$p\cdot p$		
	⑧: $\frac{n_{t_2}}{n_{t_1}}$				1
	⑨: ⑧ $\cdot$ ⑦			$p\cdot p$	
(d)	①: $\boldsymbol{U}\boldsymbol{V}^T$	$p\cdot(d-1)\cdot p$		$p\cdot d\cdot p$	
	②: $\boldsymbol{S}_{t_1}+$ ①:	$p\cdot p$			
	③: $\frac{n_{t_1}}{n_{t_2}}$				1
	④: ③ $\cdot$ ②			$p\cdot p$	

Ohne die Sherman-Morrison-Woodbury Formel berechnet sich die neue inverse Kovarianzmatrix durch (5.12). Auch hier muss die Differenz $\boldsymbol{x}_{t+1}-\boldsymbol{m}^{(k)}_{n_t^{(k)}}$ nur einmal berechnet werden. Aufgrund der nötigen Invertierung setzt sich die Berechnung aus den folgenden elementaren Rechenoperationen zusammen (vgl. Tabelle 5.2 (b)):

- $p^2(1-1)+1+p^2+1=p^2+2$ Additionen,
- p Subtraktionen,
- $p^2+1+p^2+p^2=3p^2+1$ Multiplikationen,
- 2 Divisionen,
- $\frac{8}{3}p^3$ zusätzliche Operationen für die Invertierung.

Die benötigten Additionen, Subtraktionen und Multiplikationen können annähernd als gleichwertig betrachtet werden, lediglich die Durchführung einer Division ist weitaus komplexer. Bei zunächst gleichwertiger Betrachtung aller elementaren Rechenoperationen ergibt sich der Zusammenhang für die Gesamtanzahl an Rechenoperationen annähernd durch (ohne Divisionen)

$$\begin{cases} \frac{8}{3}p^3+4p^2+p+3 & > & 8p^2+p+3 & \Leftrightarrow & 8p>12, \quad p\geq 2, \\ \underbrace{\frac{8}{3}p^3+4p^2+p+3}_{\substack{\text{Anzahl Rechenoperationen } \textit{ohne}\\ \text{Sherman-Morrison-Woodbury Formel}}} & < & \underbrace{8p^2+p+3}_{\substack{\text{Anzahl Rechenoperationen } \textit{mit}\\ \text{Sherman-Morrison-Woodbury Formel}}} & \Leftrightarrow & 8p<12, \quad p=1. \end{cases} \tag{5.16}$$

Zwar wird bei der Berechnung durch die Sherman-Morrison-Woodbury Formel eine Division mehr benötigt, jedoch beträgt die obige Abschätzung bereits ab $p=2$ Einflussgrößen $42+\frac{1}{3}>37$ zugunsten der Nutzung der Sherman-Morrison-Woodbury Formel, sodass die

Divisionen nicht mehr stark ins Gewicht fallen. Daher lässt sich sagen, dass bei Aktualisierung durch eine einzelne neue Beobachtung die Anwendung der Sherman-Morrison-Woodbury Formel zur Aktualisierung der Kovarianzmatrix schon ab $p = 2$ sinnvoll ist.

Eine alternative Betrachtung ist jene, dass Additionen vergleichsweise etwas teurer sind und demnach mit dem Faktor 2 bei der Zusammenfassung der Rechenoperationen betrachtet werden können. In diesem Fall gilt für die Aktualisierung durch eine einzelne Beobachtung mithilfe der Sherman-Morrison-Woodbury Formel (vgl. Auflistung auf Seite 108):

- $2\left(2p^2 - p + 2\right)$ Additionen,
- $p^2 + p$ Subtraktionen,
- $5p^2 + p + 1$ Multiplikationen,
- 3 Divisionen.

Ohne Verwendung der Sherman-Morrison-Woodbury Formel werden dann hingegen

- $2\left(p^2 + 2\right)$ Additionen,
- p Subtraktionen,
- $3p^2 + 1$ Multiplikationen,
- 2 Divisionen,
- $\frac{8}{3}p^3$ zusätzliche Operationen für die Invertierung

benötigt (vgl. Auflistung auf Seite 110).

Der Zusammenhang für die Gesamtanzahl an elementaren Rechenoperationen ergibt sich dann entgegen (5.16) durch (ohne Divisionen):

$$\begin{cases} \underbrace{\frac{8}{3}p^3 + 5p^2 + p + 5}_{\substack{\text{Anzahl Rechenoperationen } \textit{ohne} \\ \text{Sherman-Morrison-Woodbury Formel}}} & > & \underbrace{10p^2 + 5}_{\substack{\text{Anzahl Rechenoperationen } \textit{mit} \\ \text{Sherman-Morrison-Woodbury Formel}}} & \Leftrightarrow & 8p^2 - 15p + 3 > 0, & p \geq 2, \\ \underbrace{\frac{8}{3}p^3 + 5p^2 + p + 5}_{\substack{\text{Anzahl Rechenoperationen } \textit{ohne} \\ \text{Sherman-Morrison-Woodbury Formel}}} & < & \underbrace{10p^2 + 5}_{\substack{\text{Anzahl Rechenoperationen } \textit{mit} \\ \text{Sherman-Morrison-Woodbury Formel}}} & \Leftrightarrow & 8p^2 - 15p + 3 < 0, & p = 1. \end{cases}$$

Diese Sichtweise ändert demnach nichts an der Schlussfolgerung, dass die Nutzung der Sherman-Morrison-Woodbury Formel zur Aktualisierung der Kovarianzmatrix ab einer Dimension von $p = 2$ Vorteile bietet.

Im Fall des Updates durch einen Chunk von neuen Beobachtungen hat die Sherman-Morrison-Woodbury Formel die Form (5.10) bzw. (5.15). Es gilt $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{p \times d}$ wie in (5.9). Die Matrix $\boldsymbol{V}$ muss transponiert werden. Die Anzahl der nötigen Vertauschungen beträgt:

$$\text{Anzahl Vertauschungen} = \begin{cases} p\left(2d - p - 1\right)/2, & \text{falls } p \leq d, \\ d\left(2p - d - 1\right)/2, & \text{falls } p > d, \end{cases} \tag{5.17}$$

wobei die Anzahl der Spalten d der Matrizen wie oben ist (vgl. Seite 104):

$$d := \underbrace{\sum_{c=1}^{M} \left(\mathbb{1} \left(\left(\sum_{j=t_1+1}^{t_2} \mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}} \right) \neq 0 \right) \right)}_{\text{Anzahl aus den bisherigen } M \text{ Klassen im Chunk } \boldsymbol{x}_{t_1+1},\ldots,\boldsymbol{x}_{t_2}} + \underbrace{n_{t_1:t_2}}_{\text{Chunk-Größe}}. \tag{5.18}$$

Die Anzahl der benötigten Rechenoperationen beträgt (vgl. Tabelle 5.2 (c)):

- $p(p-1)d + d(p-1)p + d(p-1)d + d^2 + p(d-1)d + p(d-1)p = 3dp^2 + p\left(2d^2 - 3d\right) - p^2$ Additionen,
- p^2 Subtraktionen,
- $dp^2 + dp^2 + d^2p + d^2p + dp^2 + p^2 = p^2(3d+1) + 2d^2p$ Multiplikationen,
- 1 Division,
- $\frac{8}{3}d^3$ zusätzliche Operationen für die Invertierung,
- Anzahl Vertauschungen wie in (5.17).

Ohne Anwendung der Sherman-Morrison-Woodbury Formel berechnet sich die neue inverse Kovarianzmatrix durch Aktualisierung durch einen Chunk an Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$ durch (5.14). Die Anzahl der benötigten elementaren Rechenoperationen setzt sich hierbei zusammen aus (vgl. Tabelle 5.2 (d)):

- $p(d-1)p + p^2 = dp^2$ Additionen,
- $dp^2 + p^2 = p^2(d+1)$ Multiplikationen,
- 1 Division,
- $\frac{8}{3}p^3$ zusätzliche Operationen für die Invertierung,
- Anzahl Vertauschungen wie in (5.17).

Die Anzahlen benötigter Additionen, Subtraktionen und Multiplikationen werden auch hier zunächst gleichwertig betrachtet. In beiden Formeln tritt nur genau eine Division auf, daher kann diese beim Vergleich vernachlässigt werden. Der Vergleich beider Methoden bezüglich der Gesamtanzahl an elementaren Rechenoperationen ist durch die folgenden Ungleichungen gegeben:

$$\begin{aligned} \underbrace{\tfrac{8}{3}p^3 + p^2\,(2d+1)}_{\substack{\text{Anzahl Rechenoperationen } \textit{ohne} \\ \text{Sherman-Morrison-Woodbury Formel}}} &> \underbrace{\tfrac{8}{3}d^3 + 4d^2p + 3d\left(2p^2 - p\right) + p^2}_{\substack{\text{Anzahl Rechenoperationen } \textit{mit} \\ \text{Sherman-Morrison-Woodbury Formel}}}, \\ \underbrace{\tfrac{8}{3}p^3 + p^2\,(2d+1)}_{\substack{\text{Anzahl Rechenoperationen } \textit{ohne} \\ \text{Sherman-Morrison-Woodbury Formel}}} &< \underbrace{\tfrac{8}{3}d^3 + 4d^2p + 3d\left(2p^2 - p\right) + p^2}_{\substack{\text{Anzahl Rechenoperationen } \textit{mit} \\ \text{Sherman-Morrison-Woodbury Formel}}}. \end{aligned} \tag{5.19}$$

Gilt die erste Ungleichung, so ist die Anwendung der Sherman-Morrison-Woodbury Formel von Vorteil, im anderen Fall die direkte Invertierung der aktualisierten Kovarianzmatrix $\boldsymbol{S}_{t_2}$.

Aufgrund der zwei Parameter p und d ist der Zusammenhang hier weitaus komplexer als beim Vergleich der beiden Methoden zur Bestimmung der aktualisierten inversen Kovari-

anzmatrix bei Betrachtung einer einzelnen neuen Beobachtung. Für einen anschaulichen Zusammenhang zwischen p und d lässt sich daher nur für Einschränkungen zeigen, wann die Berechnung mithilfe der Sherman-Morrison-Woodbury Formel weniger Rechenoperationen benötigt. Anhand von Simulationen mit $p, d \leq 100$ für alle Kombinationen von p und d lassen sich annähernd folgende Bedingungen für p und d abschätzen:

$$\begin{cases} f_{\text{erg}}(d,p) > 0, & \text{falls } p \geq 2d + \left\lfloor \frac{d+1}{3} \right\rfloor \wedge d \leq 100 \text{ bzw. } d < \left\lfloor \frac{3p+1}{7} \right\rfloor \wedge p \leq 100, \\ f_{\text{erg}}(d,p) < 0, & \text{falls } p < 2d + \left\lfloor \frac{d+1}{3} \right\rfloor \wedge d \leq 100 \text{ bzw. } d \geq \left\lfloor \frac{3p+1}{7} \right\rfloor \wedge p \leq 100, \end{cases}$$

wobei

$$f_{\text{erg}}(d,p) = 8p^3 - 12dp^2 - 12d^2p + 9dp - 8d^3. \tag{5.20}$$

Die Gleichung $f_{\text{erg}} = 0$ lässt sich nicht in analytisch geschlossener Form darstellen.

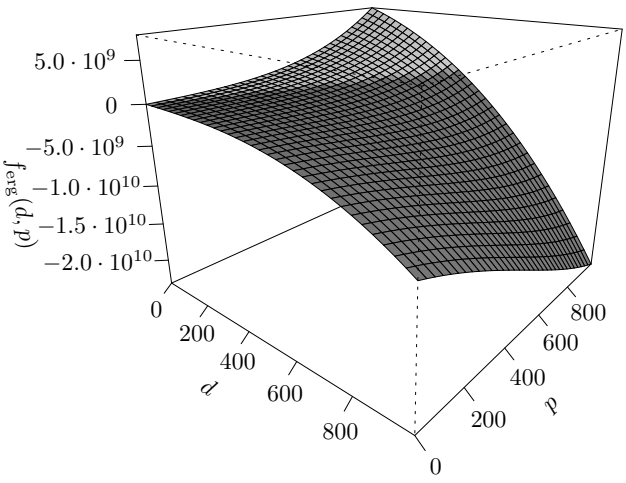
Anhand dieser Abschätzung wird bereits deutlich, dass die Nutzung der Sherman-Morrison-Woodbury Formel bei einem Update der inversen Kovarianzmatrix durch einen Chunk an neuen Beobachtungen in Hinblick auf Effizienz wenig vorteilhaft ist. Beispielsweise werden bei $d = 100$ erst ab $p \geq 233$ Einflussparametern bei Anwendung der Sherman-Morrison-Woodbury Formel weniger Rechenoperationen benötigt als bei Aktualisierung der Kovarianzmatrix selbst und anschließender Invertierung. In praktischen Anwendungen wird die Dimension p meist nicht so groß sein, der Parameter d, in welchen die Chunk-Größe einfließt (vgl. (5.18)), hängt jedoch sehr stark von der Anwendung ab.

Durch Einsetzen in die beiden Ungleichungen (5.19) ist zu sehen, dass für $d = 1000$ beispielsweise schon $p \geq 2329$ sein muss, damit die Sherman-Morrison-Woodbury Formel der direkten Invertierung der aktualisierten Kovarianzmatrix vorzuziehen ist. In Abbildung 5.1 (a) ist die Funktion f_{erg} für den Wertebereich $d, p \in \{1, \ldots, 1000\}$ dreidimensional veranschaulicht. Falls $f_{\text{erg}} > 0$, ist die Sherman-Morrison-Woodbury von Vorteil. Es wird jedoch deutlich, dass der Großteil der Fläche im negativen Wertebereich liegt, was für die direkte Invertierung der Kovarianzmatrix spricht. In Abbildung 5.1 (b) ist das Problem auf zwei Dimensionen vereinfacht. Die dunkelgraue Linie approximiert die Grenze $f_{\text{erg}} = 0$. Es zeigt sich, dass in niedrigen Dimensionen ($p, d < 1000$) die Trennung von f_{erg} im Nullpunkt approximativ linear ist.

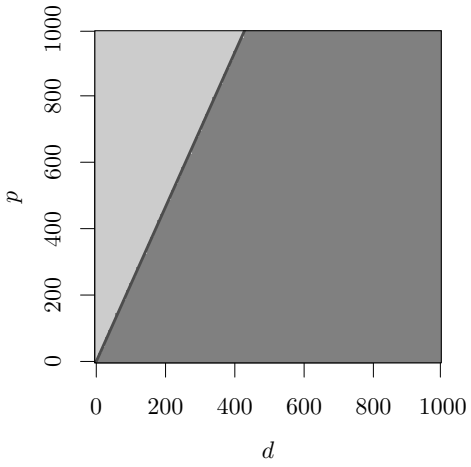
Abbildungen 5.1 (c) und (d) veranschaulichen analog die Funktion

$$f_{\text{erg}}(d,p) = 8p^3 - 18dp^2 - 18d^2p + 18dp + 3p^2 - 8d^3 \tag{5.21}$$

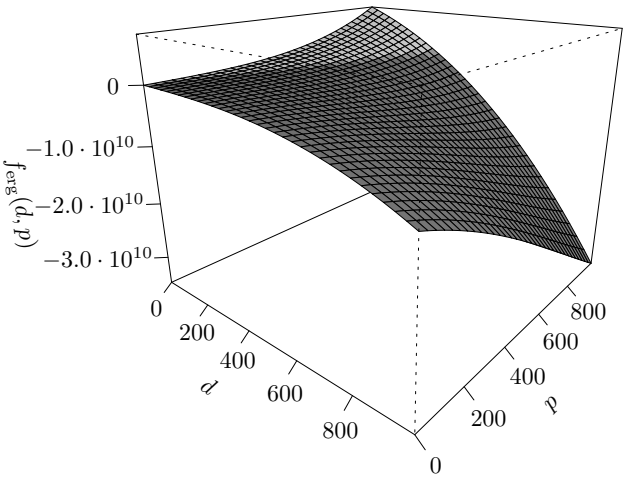
zum Vergleich der benötigten Rechenoperationen mit und ohne Verwendung der Sherman-Morrison-Woodbury Formel, welche sich durch die alternative Erfassung der jeweiligen Gesamtanzahl an elementaren Rechenoperationen ergibt. Bei dieser werden Additionen mit dem Faktor 2 betrachtet. Die Funktion $f_{\text{erg}}(d,p)$ ergibt sich aus der Differenz der Gesamtanzahl an Rechenoperationen zur Bestimmung der aktualisierten inversen Kovarianzmatrix



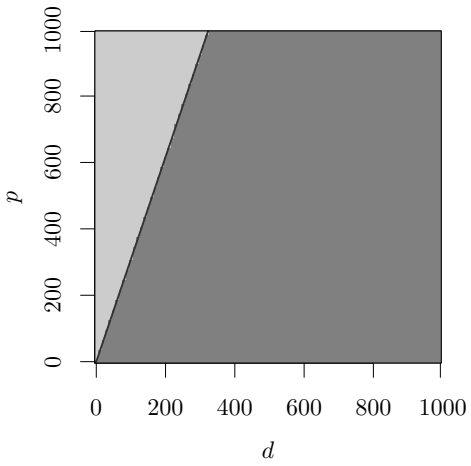
(a) Dreidimensionale Darstellung von f_{erg} aus (5.20).



(b) Zweidimensionale Darstellung von f_{erg} aus (5.20).



(c) Dreidimensionale Darstellung von f_{erg} aus (5.21).



(d) Zweidimensionale Darstellung von f_{erg} aus (5.21).

Abbildung 5.1: Darstellung von f_{erg}. Die Linie markiert approximativ $f_{\text{erg}} = 0$.

ohne Sherman-Morrison-Woodbury Formel und jener *mit* Sherman-Morrison-Woodbury Formel:

$$\begin{array}{ccc} \frac{8}{3}p^3 + 3dp^2 + p^2 & > & \frac{8}{3}d^3 + 9dp^2 + 6d^2p - 6dp, \\ \underbrace{\frac{8}{3}p^3 + 3dp^2 + p^2}_{\substack{\text{Anzahl Rechenoperationen } \textit{ohne} \\ \text{Sherman-Morrison-Woodbury Formel}}} & < & \underbrace{\frac{8}{3}d^3 + 9dp^2 + 6d^2p - 6dp}_{\substack{\text{Anzahl Rechenoperationen } \textit{mit} \\ \text{Sherman-Morrison-Woodbury Formel}}}. \end{array}$$

Auch hier wird die einzelne Division auf beiden Seiten für den Vergleich vernachlässigt. Gilt die erste Ungleichung und somit $f_{\text{erg}}(d,p) > 0$ so ist die Anwendung der Sherman-Morrison-Woodbury Formel von Vorteil, im anderen Fall die direkte Invertierung der aktualisierten Kovarianzmatrix $\boldsymbol{S}_{t_2}$.

Die einzelnen Anzahlen an Rechenoperationen (mit Faktor 2 für Additionen) betragen unter Nutzung der Sherman-Morrison-Woodbury Formel (vgl. Auflistung auf Seite 112):

- $2\left(3dp^2 + p\left(2d^2 - 3d\right) - p^2\right) = 6dp^2 + 4d^2p - 6dp - 2p^2$ Additionen,
- p^2 Subtraktionen,
- $p^2(3d+1) + 2d^2p = 3dp^2 + 2d^2p + p^2$ Multiplikationen,
- 1 Division,
- $\frac{8}{3}d^3$ zusätzliche Operationen für die Invertierung.

Ohne Verwendung der Sherman-Morrison-Woodbury Formel werden (vgl. Auflistung auf Seite 112) folgende Anzahlen an Rechenoperationen (mit Faktor 2 für Additionen) benötigt:

- $2dp^2$ Additionen,
- $p^2(d+1)$ Multiplikationen,
- 1 Division,
- $\frac{8}{3}p^3$ zusätzliche Operationen für die Invertierung.

Anhand Abbildungen 5.1 (c) und (d) wird deutlich, dass der relevante Wertebereich, in dem die Sherman-Morrison-Woodbury Formel einen Vorteil bietet ($f_{\text{erg}} > 0$), unter der alternativen Sichtweise nun noch geringer ist.

Insgesamt lässt sich schlussfolgern, dass in den meisten praktischen Anwendungen die Nutzung der Sherman-Morrison-Woodbury Formel bei Aktualisierung der Linearen Diskriminanzanalyse durch einen Chunk von Beobachtungen keinen Vorteil hat, sondern eher nachteilig ist, sodass die Update-Formel (5.5) für die Kovarianzmatrix betrachtet wird ($\boldsymbol{z}_{\text{chunks}}$ auf Seite 106 definiert):

$$\boldsymbol{S}_{t_2} = \frac{n_{t_1}}{n_{t_2}}\left(\boldsymbol{S}_{t_1} + \frac{1}{n_{t_1}}\sum_{c=1}^{M+m}\left(\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_2}^{(c)}} \cdot \boldsymbol{z}_{\text{chunks}}\, \boldsymbol{z}_{\text{chunks}}^T + \sum_{i=t_1+1}^{t_2}\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}\right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right)\right).$$

Diese wird anschließend invertiert bzw. es wird (5.14) herangezogen.

5.2 Erweiterung der Methode OLDC auf Chunks mit Lernrate

In einem nächsten Schritt wird wie bei der ursprünglichen Methode *OLDC* (vgl. Abschnitt 4.3 ab Seite 72) eine Lernrate λ in die Aktualisierungsformeln eingefügt. Dadurch soll eine Anpassung der Diskriminanzanalyse an einen concept drift ermöglicht werden. Die Idee besteht darin, dass in der Chunk Methode nun analog durch die Lernrate λ der aktuelle Chunk bei den Update-Formeln gewichtet wird.

Dabei werden die Eigenschaften der Lernrate der Update-Methode für eine einzelne Beobachtung beibehalten (vgl. Seite 72):

- $\lambda \to 0$: Der neue Chunk wird nicht betrachtet (bei $\lambda = 0$), es erfolgt keine Aktualisierung des Modells durch die neuen Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$.
- $\lambda = 1/2$: Keine Gewichtung; alle Formeln sind identisch mit jenen ohne Lernrate: (5.22) = (5.1), (5.23) = (5.2), (5.26)/(5.27) = (5.5).
- $\lambda \to 1$: Die Vergangenheit wird nicht mehr betrachtet (bei $\lambda = 1$), der ganze Fokus liegt auf den Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$ des neuen Chunks.

Herleitung der Mittelwertvektoren und Schätzer für a-priori Wahrscheinlichkeiten

Die Einführung der Lernrate erfolgt analog wie bei der Aktualisierungsformel (4.27) der Methode *OLDC* für eine einzelne Beobachtung. In der Update-Formel (5.1) wird die Summe der Beobachtungen des neuen Chunks $n_{t_1:t_2}^{(c)} \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}$ mit λ gewichtet, während die (gewichtete) „Summe aller alten Beobachtungen“ $n_{t_1}^{(c)} \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)}$ und somit der „alte“ Anteil der Beobachtungen das Gewicht $1 - \lambda$ erhält. Zur Normierung wird zusätzlich auch die bisherige Anzahl aller Beobachtungen $n_{t_2}^{(c)} = n_{t_1}^{(c)} + n_{t_1:t_2}^{(c)}$ in Klasse c zum Zeitpunkt t_2 durch die gewichtete Variante $(1-\lambda)n_{t_1}^{(c)} + \lambda n_{t_1:t_2}^{(c)}$ ersetzt:

$$\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} = \begin{cases} \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)}, & \text{falls } \forall\, i \in T : g(\boldsymbol{x}_i) \neq c, \\ \dfrac{(1-\lambda)n_{t_1}^{(c)} \boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} + \lambda n_{t_1:t_2}^{(c)} \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}}{(1-\lambda)n_{t_1}^{(c)} + \lambda n_{t_1:t_2}^{(c)}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \in \{1, \ldots, M\}, \\ \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \notin \{1, \ldots, M\}. \end{cases} \tag{5.22}$$

Die Gewichtungen bei den relativen Häufigkeiten als Schätzer für die a-priori Wahrscheinlichkeiten werden ähnlich gewählt. Die Anzahl n_{t_1} der bisherigen Beobachtungen sowie jene in Klasse c ($n_{t_1}^{(c)}$) wird entgegen (5.2) mit $1 - \lambda$ gewichtet, während der „Anteil“ des neuen Chunks $n_{t_1:t_2}$ bzw. die Anzahl neuer Beobachtungen in Klasse c ($n_{t_1:t_2}^{(c)}$) mit λ gewichtet wird:

$$P_{t_2}^{(c)} = \begin{cases} \dfrac{(1-\lambda)n_{t_1}^{(c)}}{(1-\lambda)n_{t_1} + \lambda n_{t_1:t_2}}, & \text{falls } \forall\, i \in T : g(\boldsymbol{x}_i) \neq c, \\ \dfrac{(1-\lambda)n_{t_1}^{(c)} + \lambda n_{t_1:t_2}^{(c)}}{(1-\lambda)n_{t_1} + \lambda n_{t_1:t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \in \{1, \ldots, M\}, \\ \dfrac{\lambda n_{t_1:t_2}^{(c)}}{(1-\lambda)n_{t_1} + \lambda n_{t_1:t_2}}, & \text{falls } \exists\, i \in T : g(\boldsymbol{x}_i) = c \notin \{1, \ldots, M\}. \end{cases} \tag{5.23}$$

Herleitung der aktualisierten inversen Kovarianzmatrix

Im vorherigen Abschnitt wurde gezeigt, dass die Sherman-Morrison-Woodbury Formel im Falle der Betrachtung von Chunks in den meisten Fällen zur Bestimmung der aktuellen inversen Kovarianzmatrix keinen Vorteil bietet. Daher wird die aktualisierte Kovarianzmatrix (vgl. (5.5))

$$\boldsymbol{S}_{t_2} = \frac{n_{t_1}}{n_{t_2}} \left(\boldsymbol{S}_{t_1} + \frac{1}{n_{t_1}} \sum_{c=1}^{M+m} \left(\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_2}^{(c)}} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right) \right) \tag{5.24}$$

im Algorithmus direkt invertiert.

Die gewichtete Variante der inversen Kovarianzmatrix zum Zeitpunkt t_2 nach Aktualisierung durch den neuen Chunk der Beobachtungen $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$ sieht folgendermaßen aus:

$$\boldsymbol{S}_{t_2}^{-1} = \frac{(1-\lambda) n_{t_1} + \lambda n_{t_1:t_2}}{n_{t_1}} \cdot \left((1-\lambda) \boldsymbol{S}_{t_1} + \frac{\lambda}{n_{t_1}} \sum_{c=1}^{M} \left(\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_2}^{(c)}} \cdot \boldsymbol{z}_{\text{chunks}}^{*} \boldsymbol{z}_{\text{chunks}}^{*T} \right) + \frac{\lambda n_{t_1:t_2}}{n_{t_1}} \cdot \boldsymbol{S}_{t_1:t_2} \right)^{-1} \tag{5.25}$$

$$\text{mit } \boldsymbol{z}_{\text{chunks}}^{*} := \frac{\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} \left((1-\lambda) n_{t_1}^{(c)} + \lambda n_{t_1:t_2}^{(c)} \right) - \lambda n_{t_1:t_2}^{(c)} \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}}{(1-\lambda) n_{t_1}^{(c)}} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}.$$

Zur Herleitung dieser Aktualisierungsformel für die gewichtete Variante wird zunächst die Kovarianzmatrix $\boldsymbol{S}_{t_2}$ selbst aus (5.24) betrachtet und analog die Idee von Kuncheva und Plumpton (2008, S. 514) übernommen (vgl. Seite 73):

- Der hintere „neue“ Teil in $\boldsymbol{S}_{t_2}$ wird mit λ gewichtet.
- Die alte Kovarianzmatrix $\boldsymbol{S}_{t_1}$ in der Formel für $\boldsymbol{S}_{t_2}$ wird mit $1-\lambda$ gewichtet.
- Die gesamte Anzahl der Beobachtungen $n_{t_2} = n_{t_1} + n_{t_1:t_2}$ im Vorfaktor wird durch die gewichtete Variante $(1-\lambda) n_{t_1} + \lambda n_{t_1:t_2}$ ersetzt.

Dadurch ergibt sich die vorläufige durch die Lernrate λ gewichtete Variante:

$$\boldsymbol{S}_{t_2} = \frac{n_{t_1}}{(1-\lambda) n_{t_1} + \lambda n_{t_1:t_2}} \cdot \left((1-\lambda) \boldsymbol{S}_{t_1} + \frac{\lambda}{n_{t_1}} \sum_{c} \left(\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_2}^{(c)}} \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{m}_{n_{t_1}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right) \right).$$

Mit $\boldsymbol{z}_{\text{chunks}} := \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}$ gilt:

$$\begin{aligned}\boldsymbol{S}_{t_2} = \frac{n_{t_1}}{(1-\lambda)n_{t_1} + \lambda n_{t_1:t_2}} \cdot \\ \left((1-\lambda)\boldsymbol{S}_{t_1} + \frac{\lambda}{n_{t_1}} \sum_c \left(\frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}_{\text{chunks}} \boldsymbol{z}^T_{\text{chunks}} \right.\right. \\ \left.\left. + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right) \right).\end{aligned} \tag{5.26}$$

In einem nächsten Schritt werden noch weitere Anpassungen der Formel vorgenommen, damit die auf Seite 116 erwähnten Eigenschaften für die Grenzwerte der Lernrate λ erhalten bleiben. Eine Umformung von $\boldsymbol{S}_{t_2}$ führt (bei m neuen Klassen im aktuellen Chunk $\boldsymbol{x}_{t_1+1}, \ldots, \boldsymbol{x}_{t_2}$) zu:

$$\begin{aligned}\boldsymbol{S}_{t_2} &= \frac{n_{t_1}}{(1-\lambda)n_{t_1} + \lambda n_{t_1:t_2}} \cdot \\ &\quad \left((1-\lambda)\boldsymbol{S}_{t_1} + \frac{\lambda}{n_{t_1}} \sum_{c=1}^{M+m} \left(\frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}_{\text{chunks}} \boldsymbol{z}^T_{\text{chunks}} \right.\right. \\ &\qquad \left.\left. + \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right) \right) \\ &\overset{\substack{n^{(c)}_{t_1}=0,\\ c>M}}{=} \frac{n_{t_1}}{(1-\lambda)n_{t_1} + \lambda n_{t_1:t_2}} \cdot \\ &\quad \left((1-\lambda)\boldsymbol{S}_{t_1} + \frac{\lambda}{n_{t_1}} \sum_{c=1}^{M} \left(\frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}_{\text{chunks}} \boldsymbol{z}^T_{\text{chunks}} \right) \right. \\ &\qquad \left. + \frac{\lambda}{n_{t_1}} \sum_{c=1}^{M+m} \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \right) \\ &= \frac{n_{t_1}}{(1-\lambda)n_{t_1} + \lambda n_{t_1:t_2}} \cdot \\ &\quad \left((1-\lambda)\boldsymbol{S}_{t_1} + \frac{\lambda}{n_{t_1}} \sum_{c=1}^{M} \left(\frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}_{\text{chunks}} \boldsymbol{z}^T_{\text{chunks}} \right) \right. \\ &\qquad \left. + \frac{\lambda n_{t_1:t_2}}{n_{t_1}} \cdot \underbrace{\frac{1}{n_{t_1:t_2}} \sum_{c=1}^{M+m} \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right)}_{=\boldsymbol{S}_{t_1:t_2}} \right) \\ &= \frac{n_{t_1}}{(1-\lambda)n_{t_1} + \lambda n_{t_1:t_2}} \cdot \\ &\quad \left((1-\lambda)\boldsymbol{S}_{t_1} + \frac{\lambda}{n_{t_1}} \sum_{c=1}^{M} \left(\frac{n^{(c)}_{t_1} n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}_{\text{chunks}} \boldsymbol{z}^T_{\text{chunks}} \right) + \frac{\lambda n_{t_1:t_2}}{n_{t_1}} \cdot \boldsymbol{S}_{t_1:t_2} \right).\end{aligned}$$

Es gilt:

$$\boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \overset{(5.22)}{=} \frac{(1-\lambda)n^{(c)}_{t_1}\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + \lambda n^{(c)}_{t_1:t_2}\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}}{(1-\lambda)n^{(c)}_{t_1} + \lambda n^{(c)}_{t_1:t_2}}$$

$$\iff \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} = \frac{\boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}}\left((1-\lambda)n^{(c)}_{t_1} + \lambda n^{(c)}_{t_1:t_2}\right) - \lambda n^{(c)}_{t_1:t_2}\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}}{(1-\lambda)n^{(c)}_{t_1}}.$$

Ersetze $\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}}$ in obiger Formel, d. h. anstelle von $\boldsymbol{z}_{\text{chunks}}$ wird $\boldsymbol{z}^*_{\text{chunks}}$ betrachtet mit:

$$\boldsymbol{z}^*_{\text{chunks}} := \frac{\boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}}\left((1-\lambda)n^{(c)}_{t_1} + \lambda n^{(c)}_{t_1:t_2}\right) - \lambda n^{(c)}_{t_1:t_2}\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}}{(1-\lambda)n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}.$$

Damit lässt sich die Formel für die aktualisierte Kovarianzmatrix $\boldsymbol{S}_{t_2}$ analog umformulieren:

$$\boldsymbol{S}_{t_2} = \frac{n_{t_1}}{(1-\lambda)n_{t_1} + \lambda n_{t_1:t_2}} \cdot \left((1-\lambda)\boldsymbol{S}_{t_1} + \frac{\lambda}{n_{t_1}}\sum_{c=1}^{M}\left(\frac{n^{(c)}_{t_1}n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}^*_{\text{chunks}}\boldsymbol{z}^{*T}_{\text{chunks}}\right) + \frac{\lambda n_{t_1:t_2}}{n_{t_1}} \cdot \boldsymbol{S}_{t_1:t_2}\right). \tag{5.27}$$

Die Formel (5.27) erfüllt die geforderten Eigenschaften für Werte der Lernrate λ:

- $\lambda \to 0$:

$$\begin{aligned}\boldsymbol{S}_{t_2} &\underset{\lambda=0}{=} \underbrace{\frac{n_{t_1}}{(1-0)n_{t_1} + 0 \cdot n_{t_1:t_2}}}_{=1 \text{ (für } \lambda=0)} \cdot \\ &\left(\underbrace{(1-0)\boldsymbol{S}_{t_1}}_{=\boldsymbol{S}_{t_1}} + \underbrace{\frac{0}{n_{t_1}}\sum_{c=1}^{M}\left(\frac{n^{(c)}_{t_1}n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}^*_{\text{chunks}}\boldsymbol{z}^{*T}_{\text{chunks}}\right) + \frac{0 \cdot n_{t_1:t_2}}{n_{t_1}} \cdot \boldsymbol{S}_{t_1:t_2}}_{=\boldsymbol{0}}\right) \\ &= \boldsymbol{S}_{t_1}.\end{aligned}$$

- $\lambda = 1/2$:

$$\begin{aligned}\boldsymbol{S}_{t_2} &\underset{\lambda=\frac{1}{2}}{=} \frac{n_{t_1}}{\left(1-\frac{1}{2}\right)n_{t_1} + \frac{1}{2} \cdot n_{t_1:t_2}} \cdot \\ &\left(\left(1-\frac{1}{2}\right)\boldsymbol{S}_{t_1} + \frac{1}{2n_{t_1}}\sum_{c=1}^{M}\left(\frac{n^{(c)}_{t_1}n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}^*_{\text{chunks}}\boldsymbol{z}^{*T}_{\text{chunks}}\right) + \frac{n_{t_1:t_2}}{2n_{t_1}} \cdot \boldsymbol{S}_{t_1:t_2}\right) \\ &= \frac{n_{t_1}}{n_{t_1}+n_{t_1:t_2}}\left(\boldsymbol{S}_{t_1} + \frac{1}{n_{t_1}}\sum_{c=1}^{M}\left(\frac{n^{(c)}_{t_1}n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}^*_{\text{chunks}}\boldsymbol{z}^{*T}_{\text{chunks}}\right) + \frac{n_{t_1:t_2}}{n_{t_1}} \cdot \boldsymbol{S}_{t_1:t_2}\right) \\ &= \frac{n_{t_1}}{n_{t_2}}\left(\boldsymbol{S}_{t_1} + \frac{1}{n_{t_1}}\left(\sum_{c=1}^{M+m}\left(\frac{n^{(c)}_{t_1}n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_2}} \cdot \boldsymbol{z}_{\text{chunks}}\boldsymbol{z}^{T}_{\text{chunks}}\right) + n_{t_1:t_2}\boldsymbol{S}_{t_1:t_2}\right)\right) = (5.24).\end{aligned}$$

- $\lambda \to 1$:

$$\begin{aligned}
\boldsymbol{S}_{t_2} &\underset{\lambda=1}{=} \frac{n_{t_1}}{(1-1)n_{t_1} + 1 \cdot n_{t_1:t_2}} \cdot \\
&\left(\underbrace{(1-1)\boldsymbol{S}_{t_1}}_{\underset{\lambda=1}{=} \mathbf{0}_{p\times p}} + \frac{1}{n_{t_1}} \sum_{c=1}^{M} \left(\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_2}^{(c)}} \cdot \boldsymbol{z}_{\text{chunks}}^{*} \boldsymbol{z}_{\text{chunks}}^{*T} \right) + \frac{1 \cdot n_{t_1:t_2}}{n_{t_1}} \cdot \boldsymbol{S}_{t_1:t_2} \right) \\
&= \frac{n_{t_1}}{n_{t_1:t_2}} \left(\frac{1}{n_{t_1}} \sum_{c=1}^{M} \left(\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_2}^{(c)}} \cdot \boldsymbol{z}_{\text{chunks}}^{*} \boldsymbol{z}_{\text{chunks}}^{*T} \right) + \frac{n_{t_1:t_2}}{n_{t_1}} \cdot \boldsymbol{S}_{t_1:t_2} \right) \\
&= \underbrace{\frac{1}{n_{t_1:t_2}} \sum_{c=1}^{M} \left(\frac{n_{t_1}^{(c)} n_{t_1:t_2}^{(c)}}{n_{t_2}^{(c)}} \cdot \boldsymbol{z}_{\text{chunks}}^{*} \boldsymbol{z}_{\text{chunks}}^{*T} \right)}_{\underset{\lambda\to 1}{\to} \mathbf{0}_{p\times p},\ \text{da } \boldsymbol{z}_{\text{chunks}}^{*} \underset{\lambda\to 1}{\to} \mathbf{0}\ (*)} + \boldsymbol{S}_{t_1:t_2} \underset{\lambda\to 1}{\to} \boldsymbol{S}_{t_1:t_2},
\end{aligned}$$

$$\begin{aligned}
\text{da } \lim_{\lambda\to 1} \boldsymbol{z}_{\text{chunks}}^{*} &= \lim_{\lambda\to 1} \frac{\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} \left((1-\lambda) n_{t_1}^{(c)} + \lambda n_{t_1:t_2}^{(c)} \right) - \lambda n_{t_1:t_2}^{(c)} \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}}{(1-\lambda) n_{t_1}^{(c)}} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \\
&\overset{\text{l'Hospital}}{=} \lim_{\lambda\to 1} \frac{-n_{t_1}^{(c)} \boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} + n_{t_1:t_2}^{(c)} \boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} - n_{t_1:t_2}^{(c)} \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}}{-n_{t_1}^{(c)}} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \\
&= \frac{-n_{t_1}^{(c)} \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} + n_{t_1:t_2}^{(c)} \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} - n_{t_1:t_2}^{(c)} \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}}{-n_{t_1}^{(c)}} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} \\
&= \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)} = \mathbf{0},
\end{aligned}$$

da $\boldsymbol{m}_{n_{t_2}^{(c)}}^{(c)} = \boldsymbol{m}_{n_{t_1:t_2}^{(c)}}^{(c)}$ für $\lambda = 1$ (vgl. (5.22)) und demnach $\boldsymbol{z}_{\text{chunks}}^{*} \underset{\lambda\to 1}{\to} \mathbf{0}$ $(*)$.

Invertierung von (5.27) führt somit zu der Aktualisierungsformel aus (5.25) für die gewichtete Variante der inversen Kovarianzmatrix, welche die gewünschten Eigenschaften für die verschiedenen Werte der Lernrate λ aufweist.

5.3 Erweiterung von Online Diskriminanzanalyse mit exponentiellem Vergessen auf Chunks

Im Folgenden wird die von Anagnostopoulos et al. (2012) vorgestellte Online Diskriminanzanalyse mit exponentiellem Vergessen (Abschnitt 4.4) auf die Aktualisierung durch einen Chunk von neuen Beobachtungen erweitert. Dabei wird analog die Idee herangezogen, dass eine gewichtete negative log-Likelihood betrachtet wird, bei der Summanden von Likelihood Termen zeitlich vergangener Chunks durch die Faktoren $\lambda_i^{(c)} \in [0,1],\ i = t_1, \ldots, t_\tau$, ein zunehmend geringeres Gewicht erhalten (exponentielles Vergessen).

Als vereinfachende Situation wird angenommen, dass jedes der τ Chunks Beobachtungen aus Klasse c enthält. Die geordneten Beobachtungen aus Klasse c seien mit $\boldsymbol{x}^{(c)}_{(1)} = \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_0+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_1})}, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_1+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_2})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{\tau-1}+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_\tau})}$ bezeichnet, wobei die beliebige Anzahl $n^{(c)}_{t_{j-1}:t_j}$ an Beobachtungen $\boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{j-1}+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_j})}$ in Chunk j auftritt. Die Klassenlabels $c_{(t_0+1)} = c_{(1)}, \ldots, c_{(t_1)}, c_{(t_1+1)}, \ldots, c_{(t_2)}, \ldots, c_{(t_{\tau-1}+1)}, \ldots, c_{(t_\tau)}$ kommen zeitversetzt an: Für Chunk j mit den Beobachtungen $\boldsymbol{x}_{t_{j-1}+1}, \ldots, \boldsymbol{x}_{t_j}$ zum festen Zeitpunkt t_j kommen sie mit dem folgenden Chunk $j+1$ zum Zeitpunkt t_{j+1} an.

Mit diesen Annahmen lässt sich die gewichtete negative log-Likelihood der multivariaten Normalverteilung in Anlehnung an die sequentielle Variante (4.33) wie folgt beschreiben:

$$
\begin{aligned}
&\mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \underbrace{\boldsymbol{x}^{(c)}_{(1)}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_1})}}_{\text{Beob. aus } c \text{ in Chunk } 1}, \underbrace{\boldsymbol{x}^{(c)}_{(n^{(c)}_{t_1+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_2})}}_{\text{Beob. aus } c \text{ in Chunk } 2}, \ldots, \underbrace{\boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{\tau-1}+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_\tau})}}_{\text{Beob. aus } c \text{ in Chunk } \tau}) \\
&= \sum_{j=1}^{\tau-1} \left(\underbrace{\left(\prod_{t_j \le i \le t_{\tau-1}} \lambda_i^{(c)} \right)}_{=: v^{(c)}_{t_j}} \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{j-1}+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_j})}) \right) \\
&\quad + \underbrace{v^{(c)}_{t_\tau}}_{:=1} \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{\tau-1}+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_\tau})}) \\
&\overset{(4.32)}{=} -v^{(c)}_{t_1} \log f(\boldsymbol{x}^{(c)}_{(n^{(c)}_{t_0+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_1})}; \boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}) \\
&\quad - \ldots - v^{(c)}_{t_{\tau-1}} \log f(\boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{\tau-2}+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{\tau-1}})}; \boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}) \\
&\quad - \log f(\boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{\tau-1}+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_\tau})}; \boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}) \\
&\overset{(4.32)}{=} \lambda^{(c)}_{t_1} \cdots \lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \sum_{i=t_0+1}^{t_1} \left(\log \left| \boldsymbol{\Sigma}^{(c)} \right| + \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \ldots + \lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\log \left| \boldsymbol{\Sigma}^{(c)} \right| + \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \frac{1}{2} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\log \left| \boldsymbol{\Sigma}^{(c)} \right| + \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right) \right) + const.
\end{aligned}
\tag{5.28}
$$

Zur Bestimmung des ML-Schätzers für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)}$ von Klasse c wird diese gewichtete negative log-Likelihood (5.28) nach $\boldsymbol{\mu}^{(c)}$ abgeleitet. Dabei können die einzelnen Summanden getrennt abgeleitet werden:

$$
\frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}^{(c)}_{(1)}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_1})}, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_1+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_2})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{\tau-1}+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_\tau})})
$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \cdot \frac{1}{2} \cdot n_{t_0:t_1}^{(c)} \log \left| \boldsymbol{\Sigma}^{(c)} \right| \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \cdot \frac{1}{2} \sum_{i=t_0+1}^{t_1} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \ldots + \\
&\quad + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\lambda_{t_{\tau-1}}^{(c)} \cdot \frac{1}{2} \cdot n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \log \left| \boldsymbol{\Sigma}^{(c)} \right| \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\lambda_{t_{\tau-1}}^{(c)} \cdot \frac{1}{2} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\frac{1}{2} \cdot n_{t_{\tau-1}:t_\tau}^{(c)} \log \left| \boldsymbol{\Sigma}^{(c)} \right| \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\frac{1}{2} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \right) + 0 \\
&= \lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \cdot \frac{1}{2} \sum_{i=t_0+1}^{t_1} \left(\frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \right. \\
&\qquad\qquad - \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&\qquad\qquad - \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \\
&\qquad\qquad \left. + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \ldots + \\
&\quad + \lambda_{t_{\tau-1}}^{(c)} \cdot \frac{1}{2} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \right. \\
&\qquad\qquad - \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&\qquad\qquad - \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \\
&\qquad\qquad \left. + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \frac{1}{2} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \right. \\
&\qquad\qquad - \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&\qquad\qquad \left. - \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\left(\boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \right)
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(1)}{=} \lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \cdot \frac{1}{2} \sum_{i=t_0+1}^{t_1} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} + 2 \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&\quad + \ldots + \\
&\quad + \lambda_{t_{\tau-1}}^{(c)} \cdot \frac{1}{2} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} + 2 \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&\quad + \frac{1}{2} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} + 2 \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&= \lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \sum_{i=t_0+1}^{t_1} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} + \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&\quad + \ldots + \\
&\quad + \lambda_{t_{\tau-1}}^{(c)} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} + \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&\quad + \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} + \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&= \lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \sum_{i=t_0+1}^{t_1} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \\
&\quad + \ldots + \\
&\quad + \lambda_{t_{\tau-1}}^{(c)} \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \\
&\quad + \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right).
\end{aligned}
$$

Für die Ableitungen nach Vektoren werden dabei die folgenden Rechenregeln herangezogen:

Rechenregel (1). Seien $\boldsymbol{X} \in \mathbb{R}^{p \times p}$ eine quadratische Matrix sowie $\boldsymbol{a} \in \mathbb{R}^p$ und $\boldsymbol{b} \in \mathbb{R}^p$ zwei Vektoren, dann gilt (Harville, 2008, S. 299 f.; Petersen und Pedersen, 2012, S. 10 f.):

$$
\begin{aligned}
\frac{\partial \boldsymbol{b}^T \boldsymbol{X} \boldsymbol{a}}{\partial \boldsymbol{a}} &= \boldsymbol{X}^T \boldsymbol{b} \overset{\text{falls } \boldsymbol{X}=\boldsymbol{X}^T}{=} \boldsymbol{X} \boldsymbol{b}, \\
\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{b}}{\partial \boldsymbol{a}} &= \frac{\partial \left(\boldsymbol{X} \boldsymbol{b} \right)^T \boldsymbol{a}}{\partial \boldsymbol{a}} = \frac{\partial \boldsymbol{b}^T \boldsymbol{X}^T \boldsymbol{a}}{\partial \boldsymbol{a}} = \boldsymbol{X} \boldsymbol{b}, \\
\frac{\partial \boldsymbol{a}^T \boldsymbol{X} \boldsymbol{a}}{\partial \boldsymbol{a}} &= \left(\boldsymbol{X} + \boldsymbol{X}^T \right) \boldsymbol{a} \overset{\text{falls } \boldsymbol{X}=\boldsymbol{X}^T}{=} 2 \boldsymbol{X} \boldsymbol{a}.
\end{aligned}
$$

Gleichsetzen der Ableitung mit $\mathbf{0}$ ergibt zunächst:

$$\begin{aligned}
\mathbf{0} \overset{!}{=}\ & \lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \left(\boldsymbol{\Sigma}^{(c)}\right)^{-1} \sum_{i=t_0+1}^{t_1} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right) + \ldots + \\
& + \lambda_{t_{\tau-1}}^{(c)} \left(\boldsymbol{\Sigma}^{(c)}\right)^{-1} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right) + \left(\boldsymbol{\Sigma}^{(c)}\right)^{-1} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right) \\
\Leftrightarrow \mathbf{0} = & \lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \sum_{i=t_0+1}^{t_1} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right) + \ldots + \\
& + \lambda_{t_{\tau-1}}^{(c)} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right) + \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right).
\end{aligned}$$

Mit $v_{t_j}^{(c)} := \prod_{i=t_j}^{t_{\tau-1}} \lambda_i^{(c)}$ und $v_{t_\tau}^{(c)} := 1$ (vgl. (5.28)) ergibt weiteres Auflösen den aktuellen ML-Schätzer für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)}$:

$$\begin{aligned}
& v_{t_1}^{(c)} n_{t_0:t_1}^{(c)} \boldsymbol{\mu}^{(c)} + \ldots + v_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \boldsymbol{\mu}^{(c)} + v_{t_\tau}^{(c)} n_{t_{\tau-1}:t_\tau}^{(c)} \boldsymbol{\mu}^{(c)} \\
& \quad = v_{t_1}^{(c)} \sum_{i=t_0+1}^{t_1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} + \ldots + v_{t_{\tau-1}}^{(c)} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} + v_{t_\tau}^{(c)} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \\
& \quad \Leftrightarrow \\
& \boldsymbol{\mu}^{(c)} = \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)}}{\sum\limits_{k=1}^{\tau} v_{t_k}^{(c)} n_{t_{k-1}:t_k}^{(c)}} \\
\Rightarrow\ & \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} = \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)}}{N_{t_\tau}^{(c)}},
\end{aligned} \tag{5.29}$$

wobei

$$N_{t_\tau}^{(c)} := \sum_{j=1}^{\tau} v_{t_j}^{(c)} n_{t_{j-1}:t_j}^{(c)} = \sum_{j=1}^{\tau-1} \left(\left(\prod_{i=t_j}^{t_{\tau-1}} \lambda_i^{(c)} \right) \cdot n_{t_{j-1}:t_j}^{(c)} \right) + n_{t_{\tau-1}:t_\tau}^{(c)} \tag{5.30}$$

die entsprechende Normierungskonstante ist.

Zur Bestimmung des ML-Schätzers für die Kovarianzmatrix $\boldsymbol{\Sigma}^{(c)}$ wird die negative log-Likelihood (5.28) nach $\boldsymbol{\Sigma}^{(c)}$ abgeleitet. Dabei können auch hier die Summanden einzeln betrachtet werden. Es werden die folgenden zwei Rechenregeln für Ableitungen nach Matrizen benötigt:

Rechenregel (2). Sei $\boldsymbol{X} \in \mathbb{R}^{p \times p}$ eine quadratische und invertierbare Matrix, dann gilt (Harville, 2008, S. 310; Petersen und Pedersen, 2012, S. 9):

$$\frac{\partial \log |\boldsymbol{X}|}{\partial \boldsymbol{X}} = \left(\boldsymbol{X}^{-1}\right)^T = \left(\boldsymbol{X}^T\right)^{-1} \overset{\boldsymbol{X}^T=\boldsymbol{X}}{=} \boldsymbol{X}^{-1}.$$

Rechenregel (3). Seien $\boldsymbol{X} \in \mathbb{R}^{p \times p}$ eine quadratische und invertierbare Matrix sowie $\boldsymbol{a} \in \mathbb{R}^p$ und $\boldsymbol{b} \in \mathbb{R}^p$ zwei Vektoren, dann gilt (Petersen und Pedersen, 2012, S. 10):

$$\frac{\partial \boldsymbol{a}^T \boldsymbol{X}^{-1} \boldsymbol{b}}{\partial \boldsymbol{X}} = -\left(\boldsymbol{X}^{-1}\right)^T \boldsymbol{a}\boldsymbol{b}^T \left(\boldsymbol{X}^{-1}\right)^T \overset{\boldsymbol{X}^T=\boldsymbol{X}}{=} -\boldsymbol{X}^{-1}\boldsymbol{a}\boldsymbol{b}^T\boldsymbol{X}^{-1}.$$

Damit:

$$\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}^{(c)}_{(1)}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_1})}, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_1+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_2})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_{\tau-1}+1})}, \ldots, \boldsymbol{x}^{(c)}_{(n^{(c)}_{t_\tau})}) \\
&= \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \left(\lambda^{(c)}_{t_1} \cdots \lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \cdot n^{(c)}_{t_0:t_1} \log \left| \boldsymbol{\Sigma}^{(c)} \right| \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \left(\lambda^{(c)}_{t_1} \cdots \lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \sum_{i=t_0+1}^{t_1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \ldots + \\
&\quad + \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \left(\lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \cdot n^{(c)}_{t_{\tau-2}:t_{\tau-1}} \log \left| \boldsymbol{\Sigma}^{(c)} \right| \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \left(\lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \left(\frac{1}{2} \cdot n^{(c)}_{t_{\tau-1}:t_\tau} \log \left| \boldsymbol{\Sigma}^{(c)} \right| \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \left(\frac{1}{2} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right) \right) + 0 \\
&\overset{(2)/(3)}{=} \lambda^{(c)}_{t_1} \cdots \lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \cdot n^{(c)}_{t_0:t_1} \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \\
&\quad + \lambda^{(c)}_{t_1} \cdots \lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \sum_{i=t_0+1}^{t_1} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \right) \\
&\quad + \ldots + \\
&\quad + \lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \cdot n^{(c)}_{t_{\tau-2}:t_{\tau-1}} \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \\
&\quad + \lambda^{(c)}_{t_{\tau-1}} \cdot \frac{1}{2} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \right) \\
&\quad + \frac{1}{2} \cdot n^{(c)}_{t_{\tau-1}:t_\tau} \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \\
&\quad + \frac{1}{2} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}^{(c)}_{(n^{(c)}_i)} - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \right).
\end{aligned}$$

Gleichsetzen der Ableitung mit der Nullmatrix und Multiplikation beider Seiten mit dem Faktor 2 sowie der Matrix $\mathbf{\Sigma}^{(c)}$ von rechts ergibt zunächst:

$$
\begin{aligned}
\mathbf{0} \overset{!}{=}\ & \lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} n_{t_0:t_1}^{(c)} \\
& + \lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \left(\mathbf{\Sigma}^{(c)}\right)^{-1} \sum_{i=t_0+1}^{t_1} \left(- \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \right) \\
& + \ldots + \\
& + \lambda_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \\
& + \lambda_{t_{\tau-1}}^{(c)} \left(\mathbf{\Sigma}^{(c)}\right)^{-1} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(- \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \right) \\
& + n_{t_{\tau-1}:t_\tau}^{(c)} \\
& + \left(\mathbf{\Sigma}^{(c)}\right)^{-1} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(- \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \right).
\end{aligned}
$$

Mit $v_{t_j}^{(c)} := \prod_{i=t_j}^{t_{\tau-1}} \lambda_i^{(c)}$ und $v_{t_\tau}^{(c)} := 1$ (vgl. (5.28)) ergibt weiteres Auflösen den aktuellen ML-Schätzer für die Kovarianzmatrix $\mathbf{\Sigma}^{(c)}$:

$$
\begin{aligned}
& - v_{t_1}^{(c)} n_{t_0:t_1}^{(c)} - \ldots - v_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} - v_{t_\tau}^{(c)} n_{t_{\tau-1}:t_\tau}^{(c)} \\
& \quad = \left(\mathbf{\Sigma}^{(c)}\right)^{-1} v_{t_1}^{(c)} \sum_{i=t_0+1}^{t_1} \left(- \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \right) \\
& \qquad + \ldots + \left(\mathbf{\Sigma}^{(c)}\right)^{-1} v_{t_{\tau-1}}^{(c)} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(- \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \right) \\
& \qquad + \left(\mathbf{\Sigma}^{(c)}\right)^{-1} v_{t_\tau}^{(c)} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(- \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \right) \\
& \quad \Leftrightarrow \\
& \mathbf{\Sigma}^{(c)} \left(v_{t_1}^{(c)} n_{t_0:t_1}^{(c)} + \ldots + v_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} + v_{t_\tau}^{(c)} n_{t_{\tau-1}:t_\tau}^{(c)} \right) \\
& \quad = v_{t_1}^{(c)} \sum_{i=t_0+1}^{t_1} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \\
& \qquad + \ldots + v_{t_{\tau-1}}^{(c)} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \\
& \qquad + v_{t_\tau}^{(c)} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)} \right)^T \\
& \quad \Leftrightarrow
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{\Sigma}^{(c)} &= \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)}\right)\left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{\mu}^{(c)}\right)^T}{\sum\limits_{k=1}^{\tau} v_{t_k}^{(c)} n_{t_{k-1}:t_k}^{(c)}} \\
&= \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right)^T - 2\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{\mu}^{(c)}\right)^T + \boldsymbol{\mu}^{(c)} \left(\boldsymbol{\mu}^{(c)}\right)^T\right)}{\sum\limits_{k=1}^{\tau} v_{t_k}^{(c)} n_{t_{k-1}:t_k}^{(c)}} \\
&= \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right)^T - 2\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{\mu}^{(c)}\right)^T\right)}{\sum\limits_{k=1}^{\tau} v_{t_k}^{(c)} n_{t_{k-1}:t_k}^{(c)}} + \boldsymbol{\mu}^{(c)} \left(\boldsymbol{\mu}^{(c)}\right)^T \\
\Rightarrow \tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)} &= \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right)^T - 2\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right)^T\right)}{\sum\limits_{k=1}^{\tau} v_{t_k}^{(c)} n_{t_{k-1}:t_k}^{(c)}} + \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right)^T \\
&= \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right)^T}{\sum\limits_{k=1}^{\tau} v_{t_k}^{(c)} n_{t_{k-1}:t_k}^{(c)}} - 2\sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)}}{\sum\limits_{k=1}^{\tau} v_{t_k}^{(c)} n_{t_{k-1}:t_k}^{(c)}} \left(\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right)^T \\
&\quad + \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right)^T \\
&\overset{(5.29)}{=} \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right)^T}{\sum\limits_{k=1}^{\tau} v_{t_k}^{(c)} n_{t_{k-1}:t_k}^{(c)}} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right)^T \\
&\overset{(5.30)}{=} \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum\limits_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right)^T}{N_{t_\tau}^{(c)}} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right)^T. \qquad (5.31)
\end{aligned}
$$

Analoge rekursive Formeln für die ML-Schätzer (5.29) und (5.31) inklusive Startwerte für die Online Diskriminanzanalyse auf Chunks sehen folgendermaßen aus:

$$\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} = \left(1 - \frac{n_{t_{\tau-1}:t_\tau}^{(c)}}{N_{t_\tau}^{(c)}}\right) \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)}, \quad \tilde{\boldsymbol{m}}_{n_{t_0}^{(c)}}^{(c)} := \mathbf{0}, \qquad (5.32)$$

$$\tilde{\boldsymbol{\Pi}}_{t_\tau}^{(c)} = \left(1 - \frac{n_{t_{\tau-1}:t_\tau}^{(c)}}{N_{t_\tau}^{(c)}}\right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(c)} + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)}\right)^T, \quad \tilde{\boldsymbol{\Pi}}_{t_0}^{(c)} := \mathbf{0}, \qquad (5.33)$$

$$\tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)} = \tilde{\boldsymbol{\Pi}}_{t_\tau}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right)^T, \qquad (5.34)$$

wobei der Normierungsfaktor $N_{t_\tau}^{(c)}$ folgendermaßen aktualisiert werden kann:

$$N_{t_\tau}^{(c)} = \lambda_{t_{\tau-1}}^{(c)} N_{t_{\tau-1}}^{(c)} + n_{t_{\tau-1}:t_\tau}^{(c)}, \quad N_{t_0}^{(c)} := 0. \tag{5.35}$$

Für eine Herleitung sei auf (B.17)–(B.19) in Anhang B.3 verwiesen.

Analog zur sequentiellen Methode, bei welcher eine einzelne neue Beobachtung hinzugezogen wird, wird hier die negative log-Likelihood $J_{t_{\tau+1}}^{(c)}$ des kommenden Chunks mit Beobachtungen $\boldsymbol{x}_{t_\tau+1}^{(c)}, \ldots, \boldsymbol{x}_{t_{\tau+1}}^{(c)}$ betrachtet, um die Prognosegüte für die Beobachtungen dieses folgenden Chunks auf Basis der aktuellen Schätzer zum Zeitpunkt t_τ zu bewerten (vgl. Seite 80):

$$\begin{aligned} J_{t_{\tau+1}}^{(c)} &= \mathcal{L}(\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}, \tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)}; \boldsymbol{x}_{(n_{t_\tau+1}^{(c)})}^{(c)}, \ldots, \boldsymbol{x}_{(n_{t_{\tau+1}}^{(c)})}^{(c)}) \\ &= \frac{1}{2} \sum_{i=t_\tau+1}^{t_{\tau+1}} \left(\log \left| \tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)} \right| + \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T \left(\tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)} \right)^{-1} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \right) + const. \end{aligned}$$

Der Faktor $\lambda_{t_\tau}^{(c)}$ wird daraufhin mithilfe eines Gradientenabstiegs online adaptiv angepasst (*self-tuning*) (vgl. (4.49)):

$$\lambda_{t_{\tau+1}}^{(c)} = \min \left(\lambda_+, \ \max \left(\lambda_-, \ \lambda_{t_\tau}^{(c)} - \alpha_{t_\tau}^{(c)} \left(J_{t_{\tau+1}}^{(c)} \right)' \right) \right) =: \left[\lambda_{t_\tau}^{(c)} - \alpha_{t_\tau}^{(c)} \left(J_{t_{\tau+1}}^{(c)} \right)' \right]_{\lambda_-}^{\lambda_+} \tag{5.36}$$

mit Schrittweite $\alpha_{t_\tau}^{(c)} > 0$.

Die Schwellenwerte können wie von den Autoren für die sequentielle Methode vorgeschlagen gewählt werden (vgl. Seite 81): $\lambda_- = 0.7$, $\lambda_+ = 0.999$. Für die Bestimmung der Schrittweite kann auch hier der *RPROP* Algorithmus analog zu (4.50) und (4.51) verwendet werden:

$$\alpha_{t_\tau}^{(c)} = \begin{cases} \left[1.01 \, \alpha_{t_{\tau-1}}^{(c)} \right]_{\alpha_{\min}}^{\alpha_{\max}}, & \text{falls } \left| \left(J_{t_\tau}^{(c)} \right)' \right| > 10^{-7} \text{ und } \left(J_{t_\tau}^{(c)} \right)' \left(J_{t_{\tau-1}}^{(c)} \right)' > 0, \\ \left[0.99 \, \alpha_{t_{\tau-1}}^{(c)} \right]_{\alpha_{\min}}^{\alpha_{\max}}, & \text{falls } \left| \left(J_{t_\tau}^{(c)} \right)' \right| > 10^{-7} \text{ und } \left(J_{t_\tau}^{(c)} \right)' \left(J_{t_{\tau-1}}^{(c)} \right)' \leq 0, \\ \alpha_{t_{\tau-1}}^{(c)}, & \text{falls } \left| \left(J_{t_\tau}^{(c)} \right)' \right| \leq 10^{-7}. \end{cases} \tag{5.37}$$

Unter Annahme eines unveränderlichen Faktors $\lambda^{(c)} := \lambda_i^{(c)} \ \forall i$ wird zur Gradientenbildung die negative log-Likelihood $J_{t_{\tau+1}}^{(c)}$ nach dem Faktor $\lambda^{(c)}$ abgeleitet (vgl. (4.43)):

$$\begin{aligned} \left(J_{t_{\tau+1}}^{(c)} \right)' &= \frac{\partial \mathcal{L}(\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}, \tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)}; \boldsymbol{x}_{(n_{t_\tau+1}^{(c)})}^{(c)}, \ldots, \boldsymbol{x}_{(n_{t_{\tau+1}}^{(c)})}^{(c)})}{\partial \lambda^{(c)}} \\ &= \frac{\partial}{\partial \lambda^{(c)}} \left(\frac{1}{2} \cdot n_{t_\tau:t_{\tau+1}}^{(c)} \log \left| \tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)} \right| \right) \\ &\quad + \frac{\partial}{\partial \lambda^{(c)}} \left(\frac{1}{2} \sum_{i=t_\tau+1}^{t_{\tau+1}} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T \left(\tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)} \right)^{-1} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \right) \end{aligned} \tag{5.38}$$

$$
\begin{aligned}
&= \frac{1}{2} \cdot n^{(c)}_{t_\tau:t_{\tau+1}} \left(\log \left| \tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right| \right)' \\
&\quad + \frac{1}{2} \sum_{i=t_\tau+1}^{t_{\tau+1}} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right)^T \left(-2 \left(\tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right)^{-1} \left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right)' \right. \\
&\qquad \left. + \left(\left(\tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right)^{-1} \right)' \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right) \right)
\end{aligned}
$$

und dabei sind (vgl. (4.48))

$$
\left(\left(\tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right)^{-1} \right)' = - \left(\tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right)^{-1} \left(\tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right)' \left(\tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right)^{-1}, \quad \left(\log \left| \tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right| \right)' = \operatorname{tr} \left(\left(\tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right)^{-1} \left(\tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right)' \right). \tag{5.39}
$$

Die einzelnen Aktualisierungsformeln für die Gradienten ergeben sich durch Ableiten der Formeln (5.32)–(5.35):

$$
\begin{aligned}
\left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right)' &= \frac{\partial \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}}}{\partial \lambda^{(c)}} \qquad (5.40) \\
&= \left(1 - \frac{n^{(c)}_{t_{\tau-1}:t_\tau}}{N^{(c)}_{t_\tau}} \right) \left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_{\tau-1}}} \right)' + \frac{n^{(c)}_{t_{\tau-1}:t_\tau} \left(N^{(c)}_{t_\tau} \right)'}{\left(N^{(c)}_{t_\tau} \right)^2} \cdot \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_{\tau-1}}} \\
&\quad - \frac{\left(N^{(c)}_{t_\tau} \right)'}{\left(N^{(c)}_{t_\tau} \right)^2} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}^{(c)}_{(n_i^{(c)})} \\
&= \left(1 - \frac{n^{(c)}_{t_{\tau-1}:t_\tau}}{N^{(c)}_{t_\tau}} \right) \left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_{\tau-1}}} \right)' - \frac{\left(N^{(c)}_{t_\tau} \right)'}{\left(N^{(c)}_{t_\tau} \right)^2} \left(\sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}^{(c)}_{(n_i^{(c)})} - n^{(c)}_{t_{\tau-1}:t_\tau} \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_{\tau-1}}} \right), \\
\left(\tilde{\boldsymbol{\Pi}}^{(c)}_{t_\tau} \right)' &= \frac{\partial \tilde{\boldsymbol{\Pi}}^{(c)}_{t_\tau}}{\partial \lambda^{(c)}} \qquad (5.41) \\
&= \left(1 - \frac{n^{(c)}_{t_{\tau-1}:t_\tau}}{N^{(c)}_{t_\tau}} \right) \left(\tilde{\boldsymbol{\Pi}}^{(c)}_{t_{\tau-1}} \right)' + \frac{n^{(c)}_{t_{\tau-1}:t_\tau} \left(N^{(c)}_{t_\tau} \right)'}{\left(N^{(c)}_{t_\tau} \right)^2} \cdot \tilde{\boldsymbol{\Pi}}^{(c)}_{t_{\tau-1}} \\
&\quad - \frac{\left(N^{(c)}_{t_\tau} \right)'}{\left(N^{(c)}_{t_\tau} \right)^2} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}^{(c)}_{(n_i^{(c)})} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} \right)^T \\
&= \left(1 - \frac{n^{(c)}_{t_{\tau-1}:t_\tau}}{N^{(c)}_{t_\tau}} \right) \left(\tilde{\boldsymbol{\Pi}}^{(c)}_{t_{\tau-1}} \right)' \\
&\quad - \frac{\left(N^{(c)}_{t_\tau} \right)'}{\left(N^{(c)}_{t_\tau} \right)^2} \left(\sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}^{(c)}_{(n_i^{(c)})} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} \right)^T - n^{(c)}_{t_{\tau-1}:t_\tau} \cdot \tilde{\boldsymbol{\Pi}}^{(c)}_{t_{\tau-1}} \right), \\
\left(\tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau} \right)' &= \frac{\partial \tilde{\boldsymbol{\Sigma}}^{(c)}_{t_\tau}}{\partial \lambda^{(c)}} = \left(\tilde{\boldsymbol{\Pi}}^{(c)}_{t_\tau} \right)' - \left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right)' \left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right)^T - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \left(\left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right)' \right)^T, \qquad (5.42)
\end{aligned}
$$

$$\left(N_{t_\tau}^{(c)}\right)' = \frac{\partial N_{t_\tau}^{(c)}}{\partial \lambda^{(c)}} = \lambda_{t_{\tau-1}}^{(c)} \left(N_{t_{\tau-1}}^{(c)}\right)' + N_{t_{\tau-1}}^{(c)} \tag{5.43}$$

mit Startwerten $\left(\tilde{\boldsymbol{m}}_{n_{t_0}^{(c)}}^{(c)}\right)' := \mathbf{0}$, $\left(\tilde{\boldsymbol{\Pi}}_{t_0}^{(c)}\right)' := \mathbf{0}$, $\left(N_{t_0}^{(c)}\right)' := 0$.

Der Gradient der negativen log-Likelihood (5.38) und die abgeleiteten Formeln (5.40)–(5.43) sind zwar nur unter der Annahme eines unveränderlichen Faktors $\lambda^{(c)}$ exakt, da jedoch die Schrittweite $\alpha_{t_\tau}^{(c)}$ in (5.36) im Allgemeinen klein ist und sich der Faktor daher über die Zeit nur leicht verändert, kann auch im Fall der Betrachtung von Chunks $\left(J_{t_{\tau+1}}^{(c)}\right)'$ als approximativer Gradient im Gradientenabstieg (5.36) betrachtet werden (vgl. Seite 81).

Mit den hergeleiteten Formeln liegen alle nötigen Größen für den Teilalgorithmus *G-AF* (Gaussian adaptive forgetting, s. Algorithmus 1 auf Seite 82) im Falle der Betrachtung von Chunks zur Aktualisierung vor, wobei die Indizes der Größen in Algorithmus 1 auf Chunks anzupassen sind. Die noch fehlenden (optionalen) Größen $\boldsymbol{G}_{t_\tau}^{(c)}$, $\left(\boldsymbol{G}_{t_\tau}^{(c)}\right)'$, $d_{t_\tau}^{(c)}$ und $\left(d_{t_\tau}^{(c)}\right)'$ aus Algorithmus 1 können folgendermaßen bestimmt werden: $\boldsymbol{G}_{t_\tau}^{(c)} := \left(\tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)}\right)^{-1}$, $\left(\boldsymbol{G}_{t_\tau}^{(c)}\right)' := \left(\left(\tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)}\right)^{-1}\right)'$, $d_{t_\tau}^{(c)} = \log\left|\tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)}\right|$ und $\left(d_{t_\tau}^{(c)}\right)' = \left(\log\left|\tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)}\right|\right)'$. Für die Bestimmung der Gradienten können dabei die Rechenregeln (5.39) herangezogen werden. Da davon ausgegangen wird, dass in jedem Chunk Beobachtungen aus jeder Klasse vorliegen, werden für die Aktualisierung durch jedes Chunk M Durchläufe des *G-AF* Algorithmus betrachtet.

Im Falle der LDA wird ähnlich vorgegangen wie in der sequentiellen Methode (vgl. ab Seite 86). Es werden zunächst M Durchläufe des *G-AF* Algorithmus für die Aktualisierung der Mittelwertvektoren $\boldsymbol{m}_{n_{t_\tau}^{(c)}}^{(c)}$, $c = 1, \ldots, M$, vorgenommen, wobei $\boldsymbol{\Sigma}_0 := \tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)} := \tilde{\boldsymbol{\Sigma}}_{t_{\tau-1}}^{(P)}$ (Initialisierung $\tilde{\boldsymbol{\Sigma}}_{t_0}^{(P)} := \mathbf{0}$) und $\left(\tilde{\boldsymbol{\Sigma}}_{t_\tau}^{(c)}\right)' := \mathbf{0}$ konstant gehalten werden, sodass temporäre Veränderungen der klassenspezifischen Kovarianzmatrizen keinen Einfluss auf die Anpassung des Faktors $\lambda_{t_\tau}^{(c)}$ haben.

Auf Basis der zentrierten Beobachtungen $\boldsymbol{\xi}_i^{(c)} = \boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \boldsymbol{m}_{n_{t_\tau}^{(c)}}^{(c)}$, $c \in \{1, \ldots, M\}$, $i = t_{\tau-1}+1, \ldots, t_\tau$, erfolgt im Anschluss ein weiterer *G-AF* Durchlauf zur Aktualisierung der gepoolten Kovarianzmatrix inklusive eigenem Gradienten $\left(J_{t_\tau}^{(P)}\right)'$, Faktor $\lambda_{t_\tau}^{(P)}$ und eigener Schrittweite $\alpha_{t_\tau}^{(P)}$, wobei nun der Mittelwertvektor inklusive Gradient konstant gehalten wird: $\boldsymbol{m}_0 := \tilde{\boldsymbol{m}}_{n_{t_\tau}}^{(P)} := \mathbf{0}$, $\left(\tilde{\boldsymbol{m}}_{n_{t_\tau}}^{(P)}\right)' := \mathbf{0}$. Die einzelnen Größen in diesem Durchlauf werden folgendermaßen aktualisiert:

$$\begin{aligned} N_{t_\tau}^{(P)} &= \lambda_{t_{\tau-1}}^{(P)} N_{t_{\tau-1}}^{(P)} + n_{t_{\tau-1}:t_\tau}, \quad N_{t_0}^{(P)} := 0, \\ \left(N_{t_\tau}^{(P)}\right)' &= \lambda_{t_{\tau-1}}^{(P)} \left(N_{t_{\tau-1}}^{(P)}\right)' + N_{t_{\tau-1}}^{(P)}, \end{aligned}$$

$$\begin{aligned}
\tilde{\mathbf{\Pi}}_{t_\tau}^{(P)} &= \left(1 - \frac{n_{t_{\tau-1}:t_\tau}}{N_{t_\tau}^{(P)}}\right) \tilde{\mathbf{\Pi}}_{t_{\tau-1}}^{(P)} + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \left(\sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{\xi}_i^{(c)} \left(\boldsymbol{\xi}_i^{(c)}\right)^T \right) \\
&= \left(1 - \frac{n_{t_{\tau-1}:t_\tau}}{N_{t_\tau}^{(P)}}\right) \tilde{\mathbf{\Pi}}_{t_{\tau-1}}^{(P)} \\
&\quad + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right)^T, \ \tilde{\mathbf{\Pi}}_{t_0}^{(P)} := \mathbf{0}, \\
\tilde{\mathbf{\Sigma}}_{t_\tau}^{(P)} &= \tilde{\mathbf{\Pi}}_{t_\tau}^{(P)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}}^{(P)} \left(\tilde{\boldsymbol{m}}_{n_{t_\tau}}^{(P)}\right)^T = \tilde{\mathbf{\Pi}}_{t_\tau}^{(P)} - \mathbf{0} \cdot \mathbf{0}^T = \tilde{\mathbf{\Pi}}_{t_\tau}^{(P)}. \qquad (5.44)
\end{aligned}$$

$\tilde{\mathbf{\Sigma}}_{t_\tau}^{(P)}$ ist eine rekursive Variante von (vgl. (B.20) in Anhang B.3)

$$\sum_{c=1}^{M} \sum_{j=1}^{\tau} \left(\frac{v_{t_j}^{(P)}}{N_{t_\tau}^{(P)}} \sum_{i=t_{j-1}+1}^{t_j} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_j}^{(c)}}^{(c)}\right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_j}^{(c)}}^{(c)}\right)^T \right). \qquad (5.45)$$

Dieser Schätzer hat eine größere Varianz als die gewichtete gepoolte Kovarianzmatrix

$$\mathbf{\Sigma}_{t_\tau}^{(P)} = \sum_{c=1}^{M} \sum_{j=1}^{\tau} \left(\frac{v_{t_j}^{(P)}}{N_{t_\tau}^{(P)}} \sum_{i=t_{j-1}+1}^{t_j} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}\right)^T \right). \qquad (5.46)$$

Jedoch ist auch hier (5.46) nicht schrittweise aktualisierbar, da $\tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)}$ zu früheren Zeitpunkten noch nicht bekannt ist. Der Schätzer (5.45) ist eine gute Approximation, da das Gewicht $v_{t_j}^{(P)}$ für vergangene Chunks (aus zentrierten Beobachtungen) exponentiell kleiner wird durch $v_{t_j}^{(P)} = \prod_{i=t_j}^{t_{\tau-1}} \lambda_i^{(P)}$, sodass die höhere Varianz des Schätzers relativiert wird (vgl. Seite 89).

Die gewichtete negative log-Likelihood in Abhängigkeit der Ausprägungen der Zielvariablen für die a-priori Verteilung der Klassen basiert auf der Multinomialverteilung und kann analog zum sequentiellen Fall (4.53) folgendermaßen aufgestellt werden:

$$\begin{aligned}
&\mathcal{L}^{(\boldsymbol{\lambda})}(p^{(1)}, \ldots, p^{(M)}; \underbrace{c_{(1)}, \ldots, c_{(t_1)}}_{\text{Klassen in Chunk 1}}, \underbrace{c_{(t_1+1)}, \ldots, c_{(t_2)}}_{\text{Klassen in Chunk 2}}, \ldots, \underbrace{c_{(t_{\tau-1}+1)}, \ldots, c_{(t_\tau)}}_{\text{Klassen in Chunk } \tau}) \\
&\overset{(4.53)}{=} \sum_{j=1}^{\tau-1} \left(\underbrace{\left(\prod_{t_j \le i \le t_{\tau-1}} \lambda_i^{(0)} \right)}_{=: v_{t_j}^{(0)}} \mathcal{L}(p^{(1)}, \ldots, p^{(M)}; c_{(t_{j-1}+1)}, \ldots, c_{(t_j)}) \right) \\
&\quad + \underbrace{v_{t_\tau}^{(0)}}_{:=1} \mathcal{L}(p^{(1)}, \ldots, p^{(M)}; c_{(t_{\tau-1}+1)}, \ldots, c_{(t_\tau)})
\end{aligned}$$

$$
\begin{aligned}
&\overset{(4.52)}{=} -v_{t_1}^{(0)} \sum_{c=1}^{M} \left(\left(\sum_{i=t_0+1}^{t_1} \mathbb{1}_{\{c_i=c\}} \right) \log \left(\frac{p^{(c)}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \\
&\quad - \ldots - v_{t_{\tau-1}}^{(0)} \sum_{c=1}^{M} \left(\left(\sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \mathbb{1}_{\{c_i=c\}} \right) \log \left(\frac{p^{(c)}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \\
&\quad - \sum_{c=1}^{M} \left(\left(\sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \right) \log \left(\frac{p^{(c)}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) + const. \qquad (5.47)
\end{aligned}
$$

Zur Bestimmung des ML-Schätzers für den Parameter $p^{(c)}$ wird die gewichtete negative log-Likelihood (5.47) nach diesem abgeleitet:

$$
\begin{aligned}
&\frac{\partial}{\partial p^{(c)}} \mathcal{L}^{(\boldsymbol{\lambda})}(p^{(1)}, \ldots, p^{(M)}; c_{(1)}, \ldots, c_{(t_1)}, c_{(t_1+1)}, \ldots, c_{(t_2)}, \ldots, c_{(t_{\tau-1}+1)}, \ldots, c_{(t_\tau)}) \\
&= -\frac{\partial}{\partial p^{(c)}} \left(v_{t_1}^{(0)} \sum_{c'=1}^{M} \left(\left(\sum_{i=t_0+1}^{t_1} \mathbb{1}_{\{c_i=c'\}} \right) \log \left(\frac{p^{(c')}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \right) \\
&\quad - \ldots - \frac{\partial}{\partial p^{(c)}} \left(v_{t_{\tau-1}}^{(0)} \sum_{c'=1}^{M} \left(\left(\sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \mathbb{1}_{\{c_i=c'\}} \right) \log \left(\frac{p^{(c')}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \right) \\
&\quad - \frac{\partial}{\partial p^{(c)}} \left(\sum_{c'=1}^{M} \left(\left(\sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c'\}} \right) \log \left(\frac{p^{(c')}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \right) + 0 \\
&= -v_{t_1}^{(0)} \left(\left(\sum_{i=t_0+1}^{t_1} \mathbb{1}_{\{c_i=c\}} \right) \left(\frac{1}{p^{(c)}} - \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) \right. \\
&\qquad \left. - \sum_{\substack{c'=1, \\ c' \neq c}}^{M} \left(\left(\sum_{i=t_0+1}^{t_1} \mathbb{1}_{\{c_i=c'\}} \right) \cdot \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \\
&\quad - \ldots - v_{t_{\tau-1}}^{(0)} \left(\left(\sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \mathbb{1}_{\{c_i=c\}} \right) \left(\frac{1}{p^{(c)}} - \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) \right. \\
&\qquad \left. - \sum_{\substack{c'=1, \\ c' \neq c}}^{M} \left(\left(\sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \mathbb{1}_{\{c_i=c'\}} \right) \cdot \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \\
&\quad - \left(\sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \right) \left(\frac{1}{p^{(c)}} - \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) + \sum_{\substack{c'=1, \\ c' \neq c}}^{M} \left(\left(\sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c'\}} \right) \cdot \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= -v_{t_1}^{(0)} \left(\left(\sum_{i=t_0+1}^{t_1} \mathbb{1}_{\{c_i=c\}} \right) \cdot \frac{1}{p^{(c)}} - \sum_{c'=1}^{M} \left(\sum_{i=t_0+1}^{t_1} \mathbb{1}_{\{c_i=c'\}} \right) \right) \\
&\quad - \ldots - v_{t_{\tau-1}}^{(0)} \left(\left(\sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \mathbb{1}_{\{c_i=c\}} \right) \cdot \frac{1}{p^{(c)}} - \sum_{c'=1}^{M} \left(\sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \mathbb{1}_{\{c_i=c'\}} \right) \right) \\
&\quad - \left(\left(\sum_{i=t_{\tau-1}+1}^{t_{\tau}} \mathbb{1}_{\{c_i=c\}} \right) \cdot \frac{1}{p^{(c)}} - \sum_{c'=1}^{M} \left(\sum_{i=t_{\tau-1}+1}^{t_{\tau}} \mathbb{1}_{\{c_i=c'\}} \right) \right) \\
&= -v_{t_1}^{(0)} \left(n_{t_0:t_1}^{(c)} \cdot \frac{1}{p^{(c)}} - n_{t_0:t_1} \right) - \ldots - v_{t_{\tau-1}}^{(0)} \left(n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \cdot \frac{1}{p^{(c)}} - n_{t_{\tau-2}:t_{\tau-1}} \right) \\
&\quad - \left(n_{t_{\tau-1}:t_{\tau}}^{(c)} \cdot \frac{1}{p^{(c)}} - n_{t_{\tau-1}:t_{\tau}} \right).
\end{aligned}
$$

Gleichsetzen mit 0 und Auflösen nach $p^{(c)}$ ergibt den aktuellen ML-Schätzer für $p^{(c)}$ zum Zeitpunkt t_τ:

$$
\begin{aligned}
0 &\stackrel{!}{=} -v_{t_1}^{(0)} \left(n_{t_0:t_1}^{(c)} \cdot \frac{1}{p^{(c)}} - n_{t_0:t_1} \right) - \ldots - v_{t_{\tau-1}}^{(0)} \left(n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \cdot \frac{1}{p^{(c)}} - n_{t_{\tau-2}:t_{\tau-1}} \right) \\
&\quad - \left(n_{t_{\tau-1}:t_{\tau}}^{(c)} \cdot \frac{1}{p^{(c)}} - n_{t_{\tau-1}:t_{\tau}} \right) \\
&\Leftrightarrow \\
v_{t_1}^{(0)} n_{t_0:t_1} + \ldots + v_{t_{\tau-1}}^{(0)} n_{t_{\tau-2}:t_{\tau-1}} + n_{t_{\tau-1}:t_{\tau}} &= v_{t_1}^{(0)} n_{t_0:t_1}^{(c)} \cdot \frac{1}{p^{(c)}} + \ldots + v_{t_{\tau-1}}^{(0)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \cdot \frac{1}{p^{(c)}} + n_{t_{\tau-1}:t_{\tau}}^{(c)} \cdot \frac{1}{p^{(c)}} \\
&\Leftrightarrow \\
p^{(c)} &= \frac{v_{t_1}^{(0)} n_{t_0:t_1}^{(c)} + \ldots + v_{t_{\tau-1}}^{(0)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} + n_{t_{\tau-1}:t_{\tau}}^{(c)}}{v_{t_1}^{(0)} n_{t_0:t_1} + \ldots + v_{t_{\tau-1}}^{(0)} n_{t_{\tau-2}:t_{\tau-1}} + n_{t_{\tau-1}:t_{\tau}}} = \sum_{j=1}^{\tau} \frac{v_{t_j}^{(0)} n_{t_{j-1}:t_j}^{(c)}}{\sum_{k=1}^{\tau} v_{t_k}^{(0)} n_{t_{k-1}:t_k}} \\
\Rightarrow \tilde{P}_{t_\tau}^{(c)} &= \sum_{j=1}^{\tau} \frac{v_{t_j}^{(0)} n_{t_{j-1}:t_j}^{(c)}}{\sum_{k=1}^{\tau} v_{t_k}^{(0)} n_{t_{k-1}:t_k}} = \sum_{j=1}^{\tau} \frac{v_{t_j}^{(0)} \sum_{i=t_{j-1}+1}^{t_j} \mathbb{1}_{\{c_i=c\}}}{N_{t_\tau}^{(0)}}, \quad c = 1, \ldots, M. \qquad (5.48)
\end{aligned}
$$

Ein analoger rekursiver Schätzer durch Aktualisierung auf Basis der Klassenausprägungen des neuen Chunks lässt sich aus (5.48) herleiten (vgl. (B.21) in Anhang B.3):

$$
\tilde{P}_{t_\tau}^{(c)} = \left(1 - \frac{n_{t_{\tau-1}:t_\tau}}{N_{t_\tau}^{(0)}} \right) \tilde{P}_{t_{\tau-1}}^{(c)} + \frac{1}{N_{t_\tau}^{(0)}} \cdot n_{t_{\tau-1}:t_\tau}^{(c)}, \quad \tilde{P}_{t_0}^{(c)} := \frac{1}{M}, \qquad (5.49)
$$

wobei der Normierungsfaktor folgendermaßen aktualisiert werden kann (vgl. (B.22)):

$$
N_{t_\tau}^{(0)} = \lambda_{t_{\tau-1}}^{(0)} N_{t_{\tau-1}}^{(0)} + n_{t_{\tau-1}:t_\tau}, \quad N_{t_0}^{(0)} := 0 \text{ oder } N_{t_0}^{(0)} := 1. \qquad (5.50)
$$

Auch hier ist die Initialisierung für $\tilde{P}_{t_0}^{(c)}$ im Falle von $N_{t_0}^{(0)} := 0$ irrelevant (vgl. Seite 84).

Die negative log-Likelihood $J^{(0)}_{t_{\tau+1}}$ für die Klassenausprägungen $c_{t_\tau+1}, \ldots, c_{t_{\tau+1}}$ des kommenden Chunks $\tau+1$ basierend auf den aktuellen Schätzern des Zeitpunktes t_τ

$$\begin{aligned} J^{(0)}_{t_{\tau+1}} &= \mathcal{L}(\tilde{P}^{(1)}_{t_\tau}, \ldots, \tilde{P}^{(M)}_{t_\tau}; c_{t_\tau+1}, \ldots, c_{t_{\tau+1}}) \\ &\overset{(5.47)}{=} -\sum_{c=1}^{M} \left(\left(\sum_{i=t_\tau+1}^{t_{\tau+1}} \mathbb{1}_{\{c_i=c\}} \right) \left(\log \tilde{P}^{(c)}_{t_\tau} - \log \left(\sum_{k=1}^{M} \tilde{P}^{(k)}_{t_\tau} \right) \right) \right) \\ &= -\sum_{c=1}^{M} \left(n^{(c)}_{t_\tau:t_{\tau+1}} \left(\log \tilde{P}^{(c)}_{t_\tau} - \log \left(\sum_{k=1}^{M} \tilde{P}^{(k)}_{t_\tau} \right) \right) \right) \end{aligned}$$

wird herangezogen, um die Prognosegüte für die Klassenausprägungen dieses folgenden Chunks auf Basis der aktuellen Schätzer $\tilde{P}^{(c)}_{t_\tau}$, $c = 1, \ldots, M$, zum Zeitpunkt t_τ zu bewerten.

Auf Basis des Gradienten wird der Faktor $\lambda^{(0)}$ mithilfe eines Gradientenabstiegs im Datenstrom auf Basis jeden neuen Chunks neu angepasst. Dieser Gradient lässt sich folgendermaßen herleiten:

$$\begin{aligned} \left(J^{(0)}_{t_{\tau+1}} \right)' &= \frac{\partial \mathcal{L}(\tilde{P}^{(1)}_{t_\tau}, \ldots, \tilde{P}^{(M)}_{t_\tau}; c_{t_\tau+1}, \ldots, c_{t_{\tau+1}})}{\partial \lambda^{(0)}} \\ &= \frac{\partial}{\partial \lambda^{(0)}} \left(-\sum_{c=1}^{M} \left(n^{(c)}_{t_\tau:t_{\tau+1}} \left(\log \tilde{P}^{(c)}_{t_\tau} - \log \left(\sum_{k=1}^{M} \tilde{P}^{(k)}_{t_\tau} \right) \right) \right) \right) \\ &= -\sum_{c=1}^{M} \left(n^{(c)}_{t_\tau:t_{\tau+1}} \cdot \frac{\partial}{\partial \lambda^{(0)}} \left(\log \tilde{P}^{(c)}_{t_\tau} \right) \right) + \frac{\partial}{\partial \lambda^{(0)}} \left(\log \left(\sum_{k=1}^{M} \tilde{P}^{(k)}_{t_\tau} \right) \sum_{c=1}^{M} n^{(c)}_{t_\tau:t_{\tau+1}} \right) \\ &= -\sum_{c=1}^{M} \left(n^{(c)}_{t_\tau:t_{\tau+1}} \cdot \frac{\left(\tilde{P}^{(c)}_{t_\tau} \right)'}{\tilde{P}^{(c)}_{t_\tau}} \right) + \frac{\sum_{k=1}^{M} \left(\tilde{P}^{(k)}_{t_\tau} \right)'}{\sum_{k=1}^{M} \tilde{P}^{(k)}_{t_\tau}} \cdot n_{t_\tau:t_{\tau+1}} \\ &= -\sum_{c=1}^{M} \left(n^{(c)}_{t_\tau:t_{\tau+1}} \cdot \frac{\left(\tilde{P}^{(c)}_{t_\tau} \right)'}{\tilde{P}^{(c)}_{t_\tau}} \right) + n_{t_\tau:t_{\tau+1}} \sum_{k=1}^{M} \left(\tilde{P}^{(k)}_{t_\tau} \right)' \\ &= -\sum_{c=1}^{M} \left(n^{(c)}_{t_\tau:t_{\tau+1}} \cdot \frac{\left(\tilde{P}^{(c)}_{t_\tau} \right)'}{\tilde{P}^{(c)}_{t_\tau}} - n_{t_\tau:t_{\tau+1}} \left(\tilde{P}^{(c)}_{t_\tau} \right)' \right) \\ &= -\sum_{c=1}^{M} \left(\left(n^{(c)}_{t_\tau:t_{\tau+1}} - n_{t_\tau:t_{\tau+1}} \tilde{P}^{(c)}_{t_\tau} \right) \frac{\left(\tilde{P}^{(c)}_{t_\tau} \right)'}{\tilde{P}^{(c)}_{t_\tau}} \right), \end{aligned}$$

wobei

$$\begin{aligned} \left(N^{(0)}_{t_\tau} \right)' &= \frac{\partial N^{(0)}_{t_\tau}}{\partial \lambda^{(0)}} = \lambda^{(0)}_{t_{\tau-1}} \left(N^{(0)}_{t_{\tau-1}} \right)' + N^{(0)}_{t_{\tau-1}}, \\ \left(\tilde{P}^{(c)}_{t_\tau} \right)' &= \frac{\partial \tilde{P}^{(c)}_{t_\tau}}{\partial \lambda^{(0)}} = \frac{\partial}{\partial \lambda^{(0)}} \left(\tilde{P}^{(c)}_{t_{\tau-1}} \right) - \frac{\partial}{\partial \lambda^{(0)}} \left(\frac{n_{t_{\tau-1}:t_\tau} \tilde{P}^{(c)}_{t_{\tau-1}}}{N^{(0)}_{t_\tau}} \right) + \frac{\partial}{\partial \lambda^{(0)}} \left(\frac{n^{(c)}_{t_{\tau-1}:t_\tau}}{N^{(0)}_{t_\tau}} \right) \end{aligned}$$

$$
\begin{aligned}
&= \left(\tilde{P}^{(c)}_{t_{\tau-1}}\right)' - \frac{n_{t_{\tau-1}:t_\tau}\left(\tilde{P}^{(c)}_{t_{\tau-1}}\right)' N^{(0)}_{t_\tau} - \left(N^{(0)}_{t_\tau}\right)' n_{t_{\tau-1}:t_\tau}\tilde{P}^{(c)}_{t_{\tau-1}}}{\left(N^{(0)}_{t_\tau}\right)^2} \\
&\quad - \frac{\left(N^{(0)}_{t_\tau}\right)' n^{(c)}_{t_{\tau-1}:t_\tau}}{\left(N^{(0)}_{t_\tau}\right)^2} \\
&= \left(\tilde{P}^{(c)}_{t_{\tau-1}}\right)' - \frac{n_{t_{\tau-1}:t_\tau}\left(\tilde{P}^{(c)}_{t_{\tau-1}}\right)'}{N^{(0)}_{t_\tau}} - \frac{\left(N^{(0)}_{t_\tau}\right)' n^{(c)}_{t_{\tau-1}:t_\tau} - \left(N^{(0)}_{t_\tau}\right)' n_{t_{\tau-1}:t_\tau}\tilde{P}^{(c)}_{t_{\tau-1}}}{\left(N^{(0)}_{t_\tau}\right)^2} \\
&= \left(1 - \frac{n_{t_{\tau-1}:t_\tau}}{N^{(0)}_{t_\tau}}\right)\left(\tilde{P}^{(c)}_{t_{\tau-1}}\right)' - \frac{\left(N^{(0)}_{t_\tau}\right)'}{\left(N^{(0)}_{t_\tau}\right)^2}\left(n^{(c)}_{t_{\tau-1}:t_\tau} - n_{t_{\tau-1}:t_\tau}\tilde{P}^{(c)}_{t_{\tau-1}}\right)
\end{aligned}
$$

die Ableitungen von (5.49) und (5.50) sind. Somit können auch die Gradienten mit jedem Chunk schrittweise aktualisiert werden.

Der Gradientenabstieg inklusive Aktualisierung der Schrittweite folgt analog zu (5.36) und (5.37), wobei in den Formeln die Exponenten (c) durch (0) zu ersetzen sind.

Mit den hergeleiteten Formeln liegen nun auch alle nötigen Größen für den Teilalgorithmus *M-AF* (multinomial adaptive forgetting, s. Algorithmus 2 auf Seite 83) im Falle der Betrachtung von Chunks zur Aktualisierung vor, wobei auch hier die Indizes der Größen in Algorithmus 2 auf Chunks anzupassen sind.

Insgesamt können somit die Algorithmen *QDA-AF* (Algorithmus 3 auf Seite 87) und *LDA-AF* (Algorithmus 4 auf Seite 88) auf die Aktualisierung durch einen Chunk an neuen Beobachtungen erweitert werden.

6 Untersuchung der Erwartungstreue der Schätzfunktionen für die Erwartungswertvektoren unter verschiedenen Voraussetzungen

In diesem Kapitel werden die Schätzfunktionen für die Erwartungswertvektoren der Verteilungen der Klassen der verschiedenen Methoden für Online Diskriminanzanalyse, welche jeweils bei der Prognose in der Diskriminanzanalyse herangezogen werden, unter verschiedenen Datensituationen und Annahmen miteinander verglichen. Der Fokus der Analyse liegt darauf zu untersuchen, ob die jeweiligen Schätzfunktionen zum Zeitpunkt t in den verschiedenen Datensituationen erwartungstreu für den Parameter $\boldsymbol{\mu}_{t+1}^{(c)}$, also den Erwartungswertvektor der Klasse c zum Zeitpunkt $t+1$, sind.

In den folgenden Abschnitten werden die bisherigen Schätzer für die Erwartungswertvektoren der Klassen dahingehend untersucht und verglichen. Zunächst werden in Abschnitt 6.1 die beiden betrachteten und unterstellten Situationen erläutert. Abschnitte 6.2 bis 6.4 befassen sich jeweils mit einer der in Kapitel 4 vorgestellten Methoden für Online Diskriminanzanalyse. Zum einen werden jeweils die theoretischen Eigenschaften (Erwartungstreue) der Schätzfunktionen unter einer stabilen Verteilung über die Zeit untersucht. Zum anderen wird herausgestellt, dass die Schätzfunktionen im Falle einer nicht-stabilen Verteilung, insbesondere eines concept drifts in Form der Annahme eines linearen Trends der Erwartungswertvektoren der Klassen über die Zeit, nicht mehr ideal sind, da die Erwartungstreue nicht mehr gewährleistet ist und sich demnach die Prognosegüte der Klassifikationsregel verschlechtern kann.

6.1 Situationen

Es wird jeweils die Erwartungstreue der Schätzfunktionen für die Erwartungswertvektoren für die beiden folgenden Situationen untersucht:

Stabile Verteilung Wenn kein concept drift unterstellt wird, ist die Verteilung der Beobachtungen jeder Klasse im Datenstrom zu jedem Zeitpunkt gleich. Betrachtet wird ein Datenstrom von Beobachtungen $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$. Die Beobachtungen sind Realisierungen der Zufallsvariablen $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$.

Da die geschätzten Erwartungswertvektoren der LDA für jede Klasse einzeln betrachtet und modelliert werden, werden die Beobachtungen in Gruppen derselben Klasse eingeteilt: $\{\boldsymbol{x}_i\}_{i:\, g(\boldsymbol{x}_i)=c;\ i\leq t}$. Interessant ist demnach der Erwartungswertvektor $\boldsymbol{\mu}^{(c)}$ für jede Klasse c. In der Situation einer über die Zeit stabilen Verteilung bzw. keines vorliegenden concept drifts gilt dabei: $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_1^{(c)} = \boldsymbol{\mu}_2^{(c)} = \ldots$. Genauer beschreibt der Zufallsvektor $\boldsymbol{X}^{(c)} := \boldsymbol{X}_i^{(c)}$, $i = 1, \ldots$, die Verteilung in Klasse c.

Linearer Trend der Erwartungswertvektoren In dieser Situation wird der Fall betrachtet, dass die Erwartungswertvektoren der einzelnen Klassen $c \in \{1, \ldots, M\}$ des Datenstroms einem linearen Trend in Abhängigkeit der Zeit unterliegen:

$$\boldsymbol{\mu}_i^{(c)} = \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} i, \quad i = 1, \ldots. \tag{6.1}$$

Demnach können sich für zwei unterschiedliche Zeitpunkte $t \neq s$ die Erwartungswertvektoren der Klasse c für $\boldsymbol{\beta}_1^{(c)} \neq \mathbf{0}$ unterscheiden: $\boldsymbol{\mu}_t^{(c)} \neq \boldsymbol{\mu}_s^{(c)}$.

Die Kovarianzmatrix ist hingegen zu jedem Zeitpunkt i und in allen M Klassen identisch:

$$\boldsymbol{\Sigma} := \boldsymbol{\Sigma}_i^{(1)} = \ldots = \boldsymbol{\Sigma}_i^{(M)}, \quad i = 1, \ldots.$$

Die drei in dieser Arbeit betrachteten Methoden für Online Diskriminanzanalyse werden im Folgenden getrennt voneinander unter diesen zwei Situationen und weiteren Annahmen bzw. Voraussetzungen analysiert. Eine davon ist der folgende Spezialfall:

> **Voraussetzung 4.** Bis zum Zeitpunkt t werden im Datenstrom nur Beobachtungen in Klasse c realisiert: $g(\boldsymbol{x}_i) = c$ für alle $i = 1, \ldots, t$. In diesem Spezialfall gilt
>
> $$n_t^{(c)} = \sum_{i=1}^{t} \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} = t \quad \text{und} \quad n_i^{(c)} = i,\ i = 1, \ldots, t,$$
>
> für die Anzahl der Beobachtungen in Klasse c zum Zeitpunkt i. Die Anzahl ist demnach fest und nicht zufällig.
>
> Zudem besteht die Menge $\{\boldsymbol{x}_i\}_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t}} = \{\boldsymbol{x}_i : g(\boldsymbol{x}_i) = c,\ i \leq t\}$ aus allen Beobachtungen von Zeitpunkt 1 bis t, also
>
> $$\{\boldsymbol{x}_i\}_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t}} = \{\boldsymbol{x}_i : g(\boldsymbol{x}_i) = c,\ i \leq t\} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t\}.$$

6.2 Sequential Incremental LDA (Sequential ILDA)

Zunächst wird die Online LDA ohne Vergessen bzw. Gewichtung (alle Beobachtungen haben ein identisches Gewicht), also die Methode von Pang et al. (2005b) betrachtet, welche in Abschnitt 4.2 vorgestellt wurde.

6.2.1 Situation: Stabile Verteilung

Als Erstes wird die Situation betrachtet, dass kein concept drift, sondern eine stabile Verteilung über die Zeit vorliegt. Im Folgenden wird gezeigt, dass in diesem Fall die von Pang et al. (2005b) herangezogene Schätzfunktion für den Erwartungswertvektor zum Zeitpunkt t erwartungstreu für $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$, d. h. den Erwartungswertvektor von Klasse c zum Zeitpunkt $t+1$, ist.

Es wird der standardmäßige Mittelwertvektor (4.3) der Beobachtungen von Zeitpunkt 1 bis t aus Klasse c als Schätzwert für den Erwartungswertvektor betrachtet. Dieser ist die iterative Variante der Aktualisierungsformel (4.8), welche von Pang et al. (2005b) vorgestellt wurde:

$$\boldsymbol{m}_{n_t^{(c)}}^{(c)} = \frac{1}{n_t^{(c)}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t}} \boldsymbol{x}_i = \frac{1}{\sum_{j=1}^{t} \mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}}} \sum_{i=1}^{t} \left(\boldsymbol{x}_i \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right). \tag{6.2}$$

Die entsprechende Schätzfunktion für $\boldsymbol{\mu}^{(c)}$ sieht dann folgendermaßen aus:

$$T_1^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) = \bar{\boldsymbol{X}}_t^{(c)} = \frac{1}{\sum_{j=1}^{t} \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}}} \sum_{i=1}^{t} \left(\boldsymbol{X}_i \cdot \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}}\right). \tag{6.3}$$

Die Bezeichnung $\boldsymbol{X}_j \sim F_c$ wurde auf Seite 32 eingeführt.
Die Schätzfunktion (6.3) ist erwartungstreu für den Erwartungswertvektor der Prognose $\boldsymbol{\mu}_{t+1}^{(c)}$. Um dies zu zeigen, wird auf Rechenregeln von bedingten Erwartungswerten zurückgegriffen. Zunächst wird dazu eine Voraussetzung für alle folgenden Schätzfunktionen eingeführt, die durch die praktische Anwendbarkeit der Schätzwerte begründet ist:

Voraussetzung 5. Zum Zeitpunkt t sei mindestens eine Beobachtung aus Klasse c realisiert, d. h. $\exists\ i \in \{1, \ldots, t\} : g(\boldsymbol{x}_i) = c$. Falls dies nicht der Fall ist, wird der Schätzwert nicht berechnet, die Schätzfunktion ist nicht definiert. Der Fall $\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 0$ für alle $i = 1, \ldots, t$ wird also ausgeschlossen. Die Wahrscheinlichkeit im Zeitraum 1 bis t beträgt demnach 0, also

$$\mathrm{P}\left(\bigcap_{i=1}^{t} \left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 0\right)\right) = 0 \Longleftrightarrow \left(\prod_{i=1}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 0\right)\right) = 0. \qquad (*)$$

Da die Summe über die Wahrscheinlichkeiten aller Möglichkeiten der Zusammensetzung der Zufallsvektoren von Zeitpunkt 1 bis t Eins ist, gilt:

$$\sum_{j_1=0}^{1} \cdots \sum_{j_{t-1}=0}^{1} \sum_{j_t=\mathbb{1}_{\{j_1=0\}} \cdots \mathbb{1}_{\{j_{t-1}=0\}}}^{1} \left(\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = j_1\right) \cdots \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_t \sim F_c\}} = j_t\right)\right)$$
$$+ \underbrace{\prod_{i=1}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 0\right)}_{=0 \text{ nach } (*)} = 1$$
$$\Rightarrow \sum_{j_1=0}^{1} \cdots \sum_{j_{t-1}=0}^{1} \sum_{j_t=\mathbb{1}_{\{j_1=0\}} \cdots \mathbb{1}_{\{j_{t-1}=0\}}}^{1} \left(\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = j_1\right) \cdots \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_t \sim F_c\}} = j_t\right)\right) = 1.$$

Die Erwartungstreue der Schätzfunktion wird durch den folgenden Satz beschrieben:

Satz 4. Unter der Annahme einer stabilen Verteilung und der Voraussetzung 5 ist die Schätzfunktion $T_1^{(c),P}$ aus (6.3), welche von Pang et al. (2005b) herangezogen wird, zum Zeitpunkt t erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c:

$$\mathrm{E}\left(T_1^{(c),P}\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t\right)\right) = \boldsymbol{\mu}^{(c)}.$$

Beweis. Für den Erwartungswert der Schätzfunktion gilt:

$$\mathrm{E}\left(T_1^{(c),P}\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t\right)\right) = \mathrm{E}\left(\bar{\boldsymbol{X}}_t^{(c)}\right) = \mathrm{E}\left(\frac{1}{\sum_{j=1}^{t} \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}}} \sum_{i=1}^{t} \left(\boldsymbol{X}_i \cdot \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}}\right)\right)$$
$$= \mathrm{E}\left(\mathrm{E}\left(\frac{1}{\sum_{j=1}^{t} \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}}} \sum_{i=1}^{t} \left(\boldsymbol{X}_i \cdot \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}}\right) \middle| \mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}}, \ldots, \mathbb{1}_{\{\boldsymbol{X}_t \sim F_c\}}\right)\right). \tag{6.4}$$

Aus Darstellungsgründen bezeichne im Folgenden: $B_i := \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}}$, $i = 1, \ldots, t$. Der innere Erwartungswert von (6.4) sieht folgendermaßen aus:

$$\mathrm{E}\left(\frac{1}{\sum_{j=1}^{t} B_j} \sum_{i=1}^{t} (\boldsymbol{X}_i B_i) \middle| B_1,\ \ldots,\ B_t\right)$$
$$= \begin{cases} \mathrm{E}\left(\boldsymbol{X}_1 \middle| B_1 = 1,\ B_i = 0\ (i \neq 1)\right), & B_1 = 1,\ B_i = 0\ (i \neq 1), \\ \vdots & \\ \mathrm{E}\left(\boldsymbol{X}_t \middle| B_t = 1,\ B_i = 0\ (i \neq t)\right), & B_t = 1,\ B_i = 0\ (i \neq t), \\ \mathrm{E}\left(\frac{1}{2}\left(\boldsymbol{X}_1 + \boldsymbol{X}_2\right) \middle| B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i)\right), & B_i = 1\ (i = 1,2), \\ & B_j = 0\ (j \neq i), \\ \vdots & \\ \mathrm{E}\left(\frac{1}{2}\left(\boldsymbol{X}_{t-1} + \boldsymbol{X}_t\right) \middle| B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i)\right), & B_i = 1\ (i = t-1,t), \\ & B_j = 0\ (j \neq i), \\ \vdots & \\ \mathrm{E}\left(\frac{1}{t}\sum_{i=1}^{t} \boldsymbol{X}_i \middle| B_i = 1\ (i = 1, \ldots, t)\right), & B_i = 1\ (i = 1, \ldots, t), \end{cases}$$

$$
= \begin{cases}
\mathrm{E}\left(\boldsymbol{X}_1^{(c)} \middle| B_1 = 1,\ B_i = 0\ (i \neq 1)\right), & B_1 = 1,\ B_i = 0\ (i \neq 1), \\
\vdots & \\
\mathrm{E}\left(\boldsymbol{X}_t^{(c)} \middle| B_t = 1,\ B_i = 0\ (i \neq t)\right), & B_t = 1,\ B_i = 0\ (i \neq t), \\
\mathrm{E}\left(\frac{1}{2}\left(\boldsymbol{X}_1^{(c)} + \boldsymbol{X}_2^{(c)}\right) \middle| B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i)\right), & B_i = 1\ (i = 1,2), \\
& B_j = 0\ (j \neq i), \\
\vdots & \\
\mathrm{E}\left(\frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)} + \boldsymbol{X}_t^{(c)}\right) \middle| B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i)\right), & B_i = 1\ (i = t-1,t), \\
& B_j = 0\ (j \neq i), \\
\vdots & \\
\mathrm{E}\left(\frac{1}{t}\sum_{i=1}^{t} \boldsymbol{X}_i^{(c)} \middle| B_i = 1\ (i = 1,\ldots,t)\right), & B_i = 1\ (i = 1,\ldots,t),
\end{cases}
$$

$$
= \begin{cases}
\mathrm{E}\left(\boldsymbol{X}^{(c)} \middle| B_1 = 1,\ B_i = 0\ (i \neq 1)\right), & B_1 = 1,\ B_i = 0\ (i \neq 1), \\
\vdots & \\
\mathrm{E}\left(\boldsymbol{X}^{(c)} \middle| B_t = 1,\ B_i = 0\ (i \neq t)\right), & B_t = 1,\ B_i = 0\ (i \neq t), \\
\mathrm{E}\left(\boldsymbol{X}^{(c)} \middle| B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i)\right), & B_i = 1\ (i = 1,2), \\
& B_j = 0\ (j \neq i), \\
\vdots & \\
\mathrm{E}\left(\boldsymbol{X}^{(c)} \middle| B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i)\right), & B_i = 1\ (i = t-1,t), \\
& B_j = 0\ (j \neq i), \\
\vdots & \\
\mathrm{E}\left(\boldsymbol{X}^{(c)} \middle| B_i = 1\ (i = 1,\ldots,t)\right), & B_i = 1\ (i = 1,\ldots,t),
\end{cases}
$$

$$
= \begin{cases}
\int_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_1 = 1,\ B_i = 0\ (i \neq 1))\,\mathrm{d}\boldsymbol{x}, & B_1 = 1,\ B_i = 0\ (i \neq 1), \\
\vdots & \\
\int_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_t = 1,\ B_i = 0\ (i \neq t))\,\mathrm{d}\boldsymbol{x}, & B_t = 1,\ B_i = 0\ (i \neq t), \\
\int_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i))\,\mathrm{d}\boldsymbol{x}, & B_i = 1\ (i = 1,2), \\
& B_j = 0\ (j \neq i), \\
\vdots & \\
\int_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i))\,\mathrm{d}\boldsymbol{x}, & B_i = 1\ (i = t-1,t), \\
& B_j = 0\ (j \neq i), \\
\vdots & \\
\int_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_i = 1\ (i = 1,\ldots,t))\,\mathrm{d}\boldsymbol{x}, & B_i = 1\ (i = 1,\ldots,t),
\end{cases}
$$

$$= \begin{cases} \int_{-\infty}^{\infty} \boldsymbol{x} f^{(c)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, & B_1 = 1,\ B_i = 0\ (i \neq 1), \\ \vdots & \\ \int_{-\infty}^{\infty} \boldsymbol{x} f^{(c)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, & B_t = 1,\ B_i = 0\ (i \neq t), \\ \int_{-\infty}^{\infty} \boldsymbol{x} f^{(c)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, & B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i), \\ \vdots & \\ \int_{-\infty}^{\infty} \boldsymbol{x} f^{(c)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, & B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i), \\ \vdots & \\ \int_{-\infty}^{\infty} \boldsymbol{x} f^{(c)}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}, & B_i = 1\ (i = 1,\ldots,t), \end{cases}$$

$$= \begin{cases} \boldsymbol{\mu}^{(c)}, & B_1 = 1,\ B_i = 0\ (i \neq 1), \\ \vdots & \\ \boldsymbol{\mu}^{(c)}, & B_t = 1,\ B_i = 0\ (i \neq t), \\ \boldsymbol{\mu}^{(c)}, & B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i), \\ \vdots & \\ \boldsymbol{\mu}^{(c)}, & B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i), \\ \vdots & \\ \boldsymbol{\mu}^{(c)}, & B_i = 1\ (i = 1,\ldots,t). \end{cases}$$

Dieser Erwartungswert ist eine diskrete Zufallsvariable mit $2^t - 1$ Ausprägungen. Insgesamt:

$$\begin{aligned} &\mathrm{E}\left(T_1^{(c),P}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t)\right) \qquad (6.5)\\ &= \mathrm{E}\left(\bar{\boldsymbol{X}}_t^{(c)}\right) = \mathrm{E}\left(\mathrm{E}\left(\frac{1}{\sum_{j=1}^t B_j}\sum_{i=1}^t (\boldsymbol{X}_i B_i)\middle| B_1,\ldots,B_t\right)\right)\\ &= \boldsymbol{\mu}^{(c)}\,\mathrm{P}\left((B_1=1)\cap\bigcap_{i=2}^t (B_i=0)\right) + \ldots + \boldsymbol{\mu}^{(c)}\,\mathrm{P}\left(\bigcap_{i=1}^t (B_i=1)\right)\\ &= \boldsymbol{\mu}^{(c)}\,\mathrm{P}(B_1=1)\prod_{i=2}^t \mathrm{P}(B_i=0) + \ldots + \boldsymbol{\mu}^{(c)}\prod_{i=1}^t \mathrm{P}(B_i=1)\\ &= \boldsymbol{\mu}^{(c)}\left(\sum_{j_1=0}^1\cdots\sum_{j_{t-1}=0}^1 \sum_{j_t=\prod_{k=1}^{t-1}\mathbb{1}_{\{j_k=0\}}}^1 \left(\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1\sim F_c\}}=j_1\right)\cdots\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_t\sim F_c\}}=j_t\right)\right)\right)\\ &= \boldsymbol{\mu}^{(c)}\underbrace{\left(\sum_{j_1=0}^1\cdots\sum_{j_{t-1}=0}^1 \sum_{j_t=\prod_{k=1}^{t-1}\mathbb{1}_{\{j_k=0\}}}^1 \prod_{i=1}^t \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}=j_i\right)\right)}_{=1 \text{ nach Voraussetzung 5}} = \boldsymbol{\mu}^{(c)}. \end{aligned}$$

□

6.2.2 Situation: Linearer Trend der Erwartungswertvektoren

Als Nächstes wird die Datensituation unter concept drift, bzw. genauer einem linearen Trend der Erwartungswertvektoren betrachtet (vgl. Seite 138). In dieser Situation ist die Schätzfunktion (6.3) von Zeitpunkt t nicht mehr erwartungstreu für den Erwartungswertvektor aus Klasse c des kommenden Zeitpunktes $t+1$:

Satz 5. Unter der Annahme eines linearen Trends (6.1) der Erwartungswertvektoren der Klassen ist die Schätzfunktion $T_1^{(c),P}$ von Zeitpunkt t aus (6.3), welche von Pang et al. (2005b) herangezogen wird, nicht mehr erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c des kommenden Zeitpunktes $t+1$:

$$\mathrm{E}\left(T_1^{(c),P}\left(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t\right)\right) \neq \boldsymbol{\mu}_{t+1}^{(c)}.$$

Beweis. Da der Erwartungswertvektor $\boldsymbol{\mu}_i^{(c)}$ der Klasse c nun nicht mehr über alle Zeitpunkte i identisch bleibt, ergibt sich der innere Erwartungswert von (6.4) durch

$$\mathrm{E}\left(\frac{1}{\sum_{j=1}^{t} B_j}\sum_{i=1}^{t}\left(\boldsymbol{X}_i B_i\right)\middle| B_1,\ldots,B_t\right)$$

$$= \begin{cases} \mathrm{E}\left(\boldsymbol{X}_1 \middle| B_1=1,\ B_i=0\ (i\neq 1)\right), & B_1=1,\ B_i=0\ (i\neq 1),\\ \vdots & \\ \mathrm{E}\left(\boldsymbol{X}_t \middle| B_t=1,\ B_i=0\ (i\neq t)\right), & B_t=1,\ B_i=0\ (i\neq t),\\ \mathrm{E}\left(\frac{1}{2}\left(\boldsymbol{X}_1+\boldsymbol{X}_2\right)\middle| B_i=1\ (i=1,2),\ B_j=0\ (j\neq i)\right), & B_i=1\ (i=1,2),\\ & B_j=0\ (j\neq i),\\ \vdots & \\ \mathrm{E}\left(\frac{1}{2}\left(\boldsymbol{X}_{t-1}+\boldsymbol{X}_t\right)\middle| B_i=1\ (i=t-1,t),\ B_j=0\ (j\neq i)\right), & B_i=1\ (i=t-1,t),\\ & B_j=0\ (j\neq i),\\ \vdots & \\ \mathrm{E}\left(\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{X}_i\middle| B_i=1\ (i=1,\ldots,t)\right), & B_i=1\ (i=1,\ldots,t), \end{cases}$$

$$= \begin{cases} \mathrm{E}\left(\boldsymbol{X}_1^{(c)} \middle| B_1=1,\ B_i=0\ (i\neq 1)\right), & B_1=1,\ B_i=0\ (i\neq 1),\\ \vdots & \\ \mathrm{E}\left(\boldsymbol{X}_t^{(c)} \middle| B_t=1,\ B_i=0\ (i\neq t)\right), & B_t=1,\ B_i=0\ (i\neq t),\\ \mathrm{E}\left(\frac{1}{2}\left(\boldsymbol{X}_1^{(c)}+\boldsymbol{X}_2^{(c)}\right)\middle| B_i=1\ (i=1,2),\ B_j=0\ (j\neq i)\right), & B_i=1\ (i=1,2),\\ & B_j=0\ (j\neq i),\\ \vdots & \\ \mathrm{E}\left(\frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)}+\boldsymbol{X}_t^{(c)}\right)\middle| B_i=1\ (i=t-1,t),\ B_j=0\ (j\neq i)\right), & B_i=1\ (i=t-1,t),\\ & B_j=0\ (j\neq i),\\ \vdots & \\ \mathrm{E}\left(\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{X}_i^{(c)}\middle| B_i=1\ (i=1,\ldots,t)\right), & B_i=1\ (i=1,\ldots,t), \end{cases}$$

$$
= \begin{cases}
\mathrm{E}\left(\boldsymbol{X}_1^{(c)} \middle| B_1 = 1,\ B_i = 0\ (i \neq 1)\right), & B_1 = 1,\ B_i = 0\ (i \neq 1),\\
\vdots & \\
\mathrm{E}\left(\boldsymbol{X}_t^{(c)} \middle| B_t = 1,\ B_i = 0\ (i \neq t)\right), & B_t = 1,\ B_i = 0\ (i \neq t),\\
\mathrm{E}\left(\frac{1}{2}\boldsymbol{X}_1^{(c)} \middle| B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i)\right), & B_i = 1\ (i = 1,2),\\
+\,\mathrm{E}\left(\frac{1}{2}\boldsymbol{X}_2^{(c)} \middle| B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i)\right), & B_j = 0\ (j \neq i),\\
\vdots & \\
\mathrm{E}\left(\frac{1}{2}\boldsymbol{X}_{t-1}^{(c)} \middle| B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i)\right), & B_i = 1\ (i = t-1,t),\\
+\,\mathrm{E}\left(\frac{1}{2}\boldsymbol{X}_t^{(c)} \middle| B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i)\right), & B_j = 0\ (j \neq i),\\
\vdots & \\
\sum_{i=1}^{t} \mathrm{E}\left(\frac{1}{t}\boldsymbol{X}_i^{(c)} \middle| B_i = 1\ (i = 1,\ldots,t)\right), & B_i = 1\ (i = 1,\ldots,t),
\end{cases}
$$

$$
= \begin{cases}
\int\limits_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}_1^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_1 = 1,\ B_i = 0\ (i \neq 1))\,\mathrm{d}\boldsymbol{x}, & B_1 = 1,\ B_i = 0\ (i \neq 1),\\
\vdots & \\
\int\limits_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}_t^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_t = 1,\ B_i = 0\ (i \neq t))\,\mathrm{d}\boldsymbol{x}, & B_t = 1,\ B_i = 0\ (i \neq t),\\
\frac{1}{2}\int\limits_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}_1^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i))\,\mathrm{d}\boldsymbol{x} & B_i = 1\ (i = 1,2),\\
+\,\frac{1}{2}\int\limits_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}_2^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_i = 1\ (i = 1,2),\ B_j = 0\ (j \neq i))\,\mathrm{d}\boldsymbol{x}, & B_j = 0\ (j \neq i),\\
\vdots & \\
\frac{1}{2}\int\limits_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}_{t-1}^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i))\,\mathrm{d}\boldsymbol{x} & B_i = 1\ (i = t-1,t),\\
+\,\frac{1}{2}\int\limits_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}_t^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_i = 1\ (i = t-1,t),\ B_j = 0\ (j \neq i))\,\mathrm{d}\boldsymbol{x}, & B_j = 0\ (j \neq i),\\
\vdots & \\
\sum_{i=1}^{t} \frac{1}{t}\int\limits_{-\infty}^{\infty} \boldsymbol{x} f_{\boldsymbol{X}_i^{(c)}|B_1,\ldots,B_t}(\boldsymbol{x} \,|\, B_i = 1\ (i = 1,\ldots,t))\,\mathrm{d}\boldsymbol{x}, & B_i = 1\ (i = 1,\ldots,t),
\end{cases}
$$

$$
= \begin{cases}
\boldsymbol{\mu}_1^{(c)}, & B_1 = 1,\ B_i = 0\ (i \neq 1)\\
\vdots & \\
\boldsymbol{\mu}_t^{(c)}, & B_t = 1,\ B_i = 0\ (i \neq t),\\
\frac{1}{2}\left(\boldsymbol{\mu}_1^{(c)} + \boldsymbol{\mu}_2^{(c)}\right), & B_i = 1\ (i = 1,2),\\
 & B_j = 0\ (j \neq i),\\
\vdots & \\
\frac{1}{2}\left(\boldsymbol{\mu}_{t-1}^{(c)} + \boldsymbol{\mu}_t^{(c)}\right), & B_i = 1\ (i = t-1,t),\\
 & B_j = 0\ (j \neq i),\\
\vdots & \\
\frac{1}{t}\sum_{i=1}^{t} \boldsymbol{\mu}_i^{(c)}, & B_i = 1\ (i = 1,\ldots,t).
\end{cases}
$$

Insgesamt gilt dann:

$$\begin{aligned}
&\mathrm{E}\left(T_1^{(c),P}\left(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t\right)\right) = \mathrm{E}\left(\bar{\boldsymbol{X}}_t^{(c)}\right) = \mathrm{E}\left(\frac{1}{\sum_{j=1}^{t}\mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}}}\sum_{i=1}^{t}\left(\boldsymbol{X}_i\cdot\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}\right)\right)\\
&= \mathrm{E}\left(\mathrm{E}\left(\frac{1}{\sum_{j=1}^{t}\mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}}}\sum_{i=1}^{t}\left(\boldsymbol{X}_i\cdot\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}\right)\middle|\mathbb{1}_{\{\boldsymbol{X}_1\sim F_c\}},\ldots,\mathbb{1}_{\{\boldsymbol{X}_t\sim F_c\}}\right)\right)\\
&= \boldsymbol{\mu}_1^{(c)}\,\mathrm{P}\left(B_1=1\right)\prod_{i=2}^{t}\mathrm{P}\left(B_i=0\right)+\ldots+\left(\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)\prod_{i=1}^{t}\mathrm{P}\left(B_i=1\right)\\
&= \boldsymbol{\mu}_1^{(c)}\,\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1\sim F_c\}}=1\right)\prod_{i=2}^{t}\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}=0\right)+\ldots+\left(\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)\prod_{i=1}^{t}\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}=1\right).
\end{aligned} \tag{6.6}$$

Zur Widerlegung der Erwartungstreue der Schätzfunktion wird im Folgenden der Spezialfall identischer a-priori Klassenwahrscheinlichkeiten über die Zeit betrachtet.

> **Voraussetzung 6. Identische a-priori Wahrscheinlichkeiten über die Zeit**
> Es sei für $c=1,\ldots,M$: $p^{(c)} := \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1\sim F_c\}}=1\right)=\ldots=\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_t\sim F_c\}}=1\right)$.

Unter Voraussetzung 5 (s. Seite 139 f.), Voraussetzung 6 und mithilfe von Binomialkoeffizienten bzw. der Dichte der Binomialverteilung lässt sich zeigen, dass $T_1^{(c),P}$ aus (6.3) nun nicht mehr erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Klasse c des folgenden Zeitpunktes $t+1$ ist:

$$\begin{aligned}
&\mathrm{E}\left(T_1^{(c),P}\left(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t\right)\right) = \mathrm{E}\left(\bar{\boldsymbol{X}}_t^{(c)}\right) = \mathrm{E}\left(\frac{1}{\sum_{j=1}^{t}\mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}}}\sum_{i=1}^{t}\left(\boldsymbol{X}_i\cdot\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}\right)\right)\\
&= \mathrm{E}\left(\mathrm{E}\left(\frac{1}{\sum_{j=1}^{t}\mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}}}\sum_{i=1}^{t}\left(\boldsymbol{X}_i\cdot\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}\right)\middle|\mathbb{1}_{\{\boldsymbol{X}_1\sim F_c\}},\ldots,\mathbb{1}_{\{\boldsymbol{X}_t\sim F_c\}}\right)\right)\\
&= \boldsymbol{\mu}_1^{(c)}\,\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1\sim F_c\}}=1\right)\prod_{i=2}^{t}\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}=0\right)+\ldots+\\
&\quad+\left(\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)\prod_{i=1}^{t}\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}=1\right)\\
&\overset{\text{Vor. 6}}{=} \boldsymbol{\mu}_1^{(c)}\,p^{(c)}\left(1-p^{(c)}\right)^{t-1}+\ldots+\left(\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)\left(p^{(c)}\right)^{t}\\
&= p^{(c)}\left(1-p^{(c)}\right)^{t-1}\left(\boldsymbol{\mu}_1^{(c)}+\ldots+\boldsymbol{\mu}_t^{(c)}\right)\\
&\quad+\left(p^{(c)}\right)^{2}\left(1-p^{(c)}\right)^{t-2}\left(\frac{1}{2}\left(\boldsymbol{\mu}_1^{(c)}+\boldsymbol{\mu}_2^{(c)}\right)+\ldots+\frac{1}{2}\left(\boldsymbol{\mu}_{t-1}^{(c)}+\boldsymbol{\mu}_t^{(c)}\right)\right)+\ldots+\\
&\quad+\left(p^{(c)}\right)^{t}\left(\frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= p^{(c)}\left(1-p^{(c)}\right)^{t-1}\cdot\frac{1}{t}\cdot\binom{t}{1}\left(\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)\\
&\quad+\left(p^{(c)}\right)^2\left(1-p^{(c)}\right)^{t-2}\cdot\frac{1}{t}\cdot\binom{t}{2}\left(\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)+\ldots+\left(p^{(c)}\right)^{t}\cdot\frac{1}{t}\cdot\binom{t}{t}\left(\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)\\
&= \frac{1}{t}\left(\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)\left(\sum_{i=1}^{t}\binom{t}{i}\left(p^{(c)}\right)^{i}\left(1-p^{(c)}\right)^{t-i}\right)\\
&\overset{\text{Vor. 5}}{=} \frac{1}{t}\left(\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)\left(\sum_{i=1}^{t}\binom{t}{i}\left(p^{(c)}\right)^{i}\left(1-p^{(c)}\right)^{t-i}+\binom{t}{0}\left(p^{(c)}\right)^{0}\left(1-p^{(c)}\right)^{t}\right)\\
&= \frac{1}{t}\left(\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\right)\Bigg(\underbrace{\sum_{i=0}^{t}\binom{t}{i}\left(p^{(c)}\right)^{i}\left(1-p^{(c)}\right)^{t-i}}_{=1}\Bigg)\\
&= \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\mu}_i^{(c)}\\
&\overset{(6.1)}{=} \frac{1}{t}\sum_{i=1}^{t}\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}i\right)=\frac{1}{t}\left(t\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}\sum_{i=1}^{t}i\right)\\
&= \boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}\cdot\frac{t(t+1)}{2t}=\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}\cdot\frac{t+1}{2}\\
&\overset{(6.1)}{=} \boldsymbol{\mu}_{\frac{t+1}{2}}^{(c)}\neq\boldsymbol{\mu}_{t+1}^{(c)}.
\end{aligned}
$$

Da $T_1^{(c),P}$ bereits im Spezialfall identischer a-priori Klassenwahrscheinlichkeiten über die Zeit nicht erwartungstreu für $\boldsymbol{\mu}_{t+1}^{(c)}$ ist, kann die Erwartungstreue auch im allgemeinen Fall ausgeschlossen werden. □

6.3 Online Linear Discriminant Classifier (OLDC)

Die Methode von Kuncheva und Plumpton (2008) wurde in Abschnitt 4.3 vorgestellt. Die Autorinnen betrachten eine gewichtete Variante des Mittelwertvektors als Schätzwert für den Erwartungswertvektor der Klasse c, um eine Adaption an die aktuelle Verteilung zu realisieren (vgl. Formel (4.27)):

$$
\boldsymbol{m}_{n_t^{(c)}}^{(c)}=\begin{cases}\boldsymbol{m}_{n_{t-1}^{(c)}}^{(c)}, & \text{falls } g(\boldsymbol{x}_t)\neq c,\\ \dfrac{(1-\lambda)n_{t-1}^{(c)}\boldsymbol{m}_{n_{t-1}^{(c)}}^{(c)}+\lambda\boldsymbol{x}_t}{(1-\lambda)n_{t-1}^{(c)}+\lambda}, & \text{falls } g(\boldsymbol{x}_t)=c\in\{1,\ldots,M\},\\ \boldsymbol{x}_t, & \text{falls } g(\boldsymbol{x}_t)=c=M+1.\end{cases}
\tag{6.7}
$$

Diese rekursive Formel lässt sich auch analog durch eine iterative Variante auf Basis aller Beobachtungen $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_t$ formulieren. Aufgrund der Gewichtungen bzw. der Lernrate λ und weil nicht zu jedem Zeitpunkt eine Beobachtung in Klasse c auftreten muss und

demnach die drei Fälle aus (6.7) zusammengefasst werden müssen, ist die allgemeine Darstellung jedoch recht komplex:

$$\begin{aligned}\boldsymbol{m}^{(c)}_{n_t^{(c)}} = \frac{1}{(1-\lambda)(n_t^{(c)}-1)+\lambda}\cdot \\ \Bigg(\frac{(1-\lambda)^{n_t^{(c)}-1}(n_t^{(c)}-1)!}{\prod\limits_{\substack{j:\, g(\boldsymbol{x}_j)=c \cap n_j^{(c)}>2 \\ 3\le j\le t}} ((1-\lambda)(n_j^{(c)}-2)+\lambda)} \cdot \boldsymbol{x}_{\substack{\operatorname{argmin} \boldsymbol{x}_i \\ i:\, g(\boldsymbol{x}_i)=c \\ i<t}} \\ + \sum_{\substack{j=\operatorname{argmin} \boldsymbol{x}_i+1 \\ i:\, g(\boldsymbol{x}_i)=c \\ i<t}}^{\substack{\operatorname{argmax} \boldsymbol{x}_i-1 \\ i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} \frac{\frac{(n_t^{(c)}-1)!}{(n_j^{(c)}-1)!}\cdot(1-\lambda)^{n_t^{(c)}-n_j^{(c)}}\lambda}{\prod\limits_{\substack{k:\, g(\boldsymbol{x}_k)=c \cap n_k^{(c)}>2 \\ j+1\le k\le t}} ((1-\lambda)(n_k^{(c)}-2)+\lambda)} \cdot \boldsymbol{x}_j \cdot \mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}} \\ + \lambda\, \boldsymbol{x}_{\substack{\operatorname{argmax} \boldsymbol{x}_i \\ i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} \Bigg) \cdot \mathbb{1}_{\{g(\boldsymbol{x}_t)\in\{1,\dots,M\}\}} + \boldsymbol{x}_t \cdot \mathbb{1}_{\{g(\boldsymbol{x}_t)=M+1\}} + \boldsymbol{m}^{(c)}_{n_{t-1}^{(c)}} \cdot \mathbb{1}_{\{g(\boldsymbol{x}_t)\neq c\}}.\end{aligned} \tag{6.8}$$

Im Folgenden sei in Abschnitt 6.3.1 bzw. 6.3.2 jedoch der Spezialfall $g(\boldsymbol{x}_i) = c$, $i = 1, \dots, t$, betrachtet bzw. die Voraussetzung 4 (vgl. Seite 138) unterstellt.

Die iterative Variante (6.8) des Schätzwertes $\boldsymbol{m}^{(c)}_{n_t^{(c)}}$ lässt sich unter dieser Voraussetzung 4 vereinfachen zu

$$\begin{aligned}\boldsymbol{m}^{(c)}_{n_t^{(c)}} = \frac{1}{(1-\lambda)(t-1)+\lambda}\cdot \\ \left(\frac{(1-\lambda)^{t-1}\prod\limits_{j=1}^{t-1} j}{\prod\limits_{j=3}^{t} ((1-\lambda)(j-2)+\lambda)} \cdot \boldsymbol{x}_1 + \sum_{j=2}^{t-1} \frac{(1-\lambda)^{t-j}\lambda \prod\limits_{k=j}^{t-1} k}{\prod\limits_{k=j+1}^{t} ((1-\lambda)(k-2)+\lambda)} \cdot \boldsymbol{x}_j + \lambda \boldsymbol{x}_t \right).\end{aligned} \tag{6.9}$$

Die entsprechende Schätzfunktion für $\boldsymbol{\mu}^{(c)}$ sieht dann folgendermaßen aus:

$$T_1^{(c),K}(\boldsymbol{X}_1, \dots, \boldsymbol{X}_t) := T_1^{(c),K}(\boldsymbol{X}_1^{(c)}, \dots, \boldsymbol{X}_t^{(c)}) \tag{6.10}$$

$$\begin{aligned}= \frac{1}{(1-\lambda)(t-1)+\lambda}\cdot \\ \left(\frac{(1-\lambda)^{t-1}\prod\limits_{j=1}^{t-1} j}{\prod\limits_{j=3}^{t} ((1-\lambda)(j-2)+\lambda)} \cdot \boldsymbol{X}_1^{(c)} + \sum_{j=2}^{t-1} \frac{(1-\lambda)^{t-j}\lambda \prod\limits_{k=j}^{t-1} k}{\prod\limits_{k=j+1}^{t} ((1-\lambda)(k-2)+\lambda)} \cdot \boldsymbol{X}_j^{(c)} + \lambda \boldsymbol{X}_t^{(c)} \right).\end{aligned}$$

6.3.1 Situation: Stabile Verteilung

Im Falle einer stabilen Verteilung in Klasse c über die Zeit gilt (vgl. Seite 138): $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_1^{(c)} = \boldsymbol{\mu}_2^{(c)} = \ldots$.

Satz 6. Unter der Annahme einer stabilen Verteilung und dem Spezialfall aus Voraussetzung 4 (Seite 138) ist die Schätzfunktion $T_1^{(c),K}$ aus (6.10), welche von Kuncheva und Plumpton (2008) vorgestellt wurde, zum Zeitpunkt t (für alle Zeitpunkte $t \geq 3$ im Datenstrom) erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c:

$$\mathrm{E}\left(T_1^{(c),K}\left(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t\right)\right) = \mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right) = \boldsymbol{\mu}^{(c)}.$$

Beweis. Für den Erwartungswert der aus Voraussetzung 4 resultierenden Schätzfunktion (6.10) gilt:

$$\begin{aligned}
&\mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t)\right) = \mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right)\\
&= \frac{1}{(1-\lambda)(t-1)+\lambda}\left(\frac{(1-\lambda)^{t-1}\prod_{j=1}^{t-1} j}{\prod_{j=3}^{t}((1-\lambda)(j-2)+\lambda)}\cdot \mathrm{E}\left(\boldsymbol{X}_1^{(c)}\right)\right.\\
&\qquad\left. + \sum_{j=2}^{t-1}\frac{(1-\lambda)^{t-j}\lambda\prod_{k=j}^{t-1} k}{\prod_{k=j+1}^{t}((1-\lambda)(k-2)+\lambda)}\cdot \mathrm{E}\left(\boldsymbol{X}_j^{(c)}\right) + \lambda\,\mathrm{E}\left(\boldsymbol{X}_t^{(c)}\right)\right)\\
&= \frac{1}{(1-\lambda)(t-1)+\lambda}\cdot\\
&\quad\left(\frac{(1-\lambda)^{t-1}(t-1)!}{\prod_{j=1}^{t-2}((1-\lambda)j+\lambda)}\cdot\boldsymbol{\mu}^{(c)} + \sum_{j=2}^{t-1}\frac{\frac{(t-1)!}{(j-1)!}\cdot(1-\lambda)^{t-j}\lambda}{\prod_{k=j-1}^{t-2}((1-\lambda)k+\lambda)}\cdot\boldsymbol{\mu}^{(c)} + \lambda\boldsymbol{\mu}^{(c)}\right)\\
&= \boldsymbol{\mu}^{(c)}\left(\frac{(1-\lambda)^{t-1}(t-1)!}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{t-1}\frac{\frac{(t-1)!}{(j-1)!}\cdot(1-\lambda)^{t-j}\lambda}{\prod_{k=j-1}^{t-1}((1-\lambda)k+\lambda)} + \frac{\lambda}{(1-\lambda)(t-1)+\lambda}\right) \qquad (6.11)\\
&= \boldsymbol{\mu}^{(c)}\left(\frac{(1-\lambda)^{t-1}(t-1)!}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{t-1}\frac{\frac{(t-1)!}{(j-1)!}\cdot(1-\lambda)^{t-j}\lambda}{\prod_{k=j-1}^{t-1}((1-\lambda)k+\lambda)} + \frac{\lambda\prod_{j=1}^{t-2}((1-\lambda)j+\lambda)}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)}\right)\\
&= \boldsymbol{\mu}^{(c)}\left(\frac{(1-\lambda)^{t-1}(t-1)!+\lambda\prod_{j=1}^{t-2}((1-\lambda)j+\lambda)}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{t-1}\frac{\frac{(t-1)!}{(j-1)!}\cdot(1-\lambda)^{t-j}\lambda}{\prod_{k=j-1}^{t-1}((1-\lambda)k+\lambda)}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\mu}^{(c)} \left(\frac{(1-\lambda)^{t-1}(t-1)! + \lambda \prod_{j=1}^{t-2}((1-\lambda)j+\lambda)}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} + \sum_{j=3}^{t-1} \frac{\frac{(t-1)!}{(j-1)!} \cdot (1-\lambda)^{t-j}\lambda}{\prod_{k=j-1}^{t-1}((1-\lambda)k+\lambda)} \right. \\
&\qquad \left. + \frac{(t-1)!\,(1-\lambda)^{t-2}\lambda}{\prod_{k=1}^{t-1}((1-\lambda)k+\lambda)} \right) \\
&= \left((1-\lambda)^{t-1}(t-1)! + \lambda \prod_{j=1}^{t-2}((1-\lambda)j+\lambda) + (1-\lambda)^{t-2}(t-1)!\,\lambda \right. \\
&\qquad \left. + \sum_{j=3}^{t-1} \left((1-\lambda)^{t-j} \cdot \frac{(t-1)!}{(j-1)!} \cdot \lambda \prod_{k=1}^{j-2}((1-\lambda)k+\lambda) \right) \right) \cdot \frac{\boldsymbol{\mu}^{(c)}}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} \\
&= \frac{\boldsymbol{\mu}^{(c)}}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} \left((1-\lambda)^{t-2}(t-1)! + \lambda \prod_{j=1}^{t-2}((1-\lambda)j+\lambda) \right. \\
&\qquad \left. + \sum_{j=3}^{t-1} \left((1-\lambda)^{t-j} \cdot \frac{(t-1)!}{(j-1)!} \cdot \lambda \prod_{k=1}^{j-2}((1-\lambda)k+\lambda) \right) \right) \\
&= \frac{\boldsymbol{\mu}^{(c)}}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} \left((1-\lambda)^{t-2}(t-1)! \right. \\
&\qquad \left. + \sum_{j=3}^{t} \left((1-\lambda)^{t-j} \cdot \frac{(t-1)!}{(j-1)!} \cdot \lambda \prod_{k=1}^{j-2}((1-\lambda)k+\lambda) \right) \right) \\
&= \frac{\boldsymbol{\mu}^{(c)}}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} \cdot \\
&\quad \left((1-\lambda)^{t-2}(t-1)! + \lambda(t-1)! \sum_{j=3}^{t} \left(\frac{(1-\lambda)^{t-j}}{(j-1)!} \prod_{k=1}^{j-2}((1-\lambda)k+\lambda) \right) \right).
\end{aligned}
$$

Um zu zeigen, dass $T_1^{(c),K}$ erwartungstreu für $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ ist, müssen der Nenner des Vorfaktors und der in Klammern hintere Teil der Formel identisch sein. Dies lässt sich mithilfe vollständiger Induktion zeigen. Die Induktionsvoraussetzung sei:

$$
\begin{aligned}
&\prod_{j=1}^{t-1}((1-\lambda)j+\lambda) \qquad \text{(IV)} \\
&= \left((1-\lambda)^{t-2}(t-1)! + \lambda(t-1)! \sum_{j=3}^{t} \left(\frac{(1-\lambda)^{t-j}}{(j-1)!} \prod_{k=1}^{j-2}((1-\lambda)k+\lambda) \right) \right).
\end{aligned}
$$

Die Gleichheit gilt lediglich für $t \geq 3$, als Induktionsanfang wird daher der Fall $t = 3$ betrachtet:

$$\begin{aligned}\prod_{j=1}^{2}((1-\lambda)j+\lambda) &= ((1-\lambda)\cdot 1+\lambda)((1-\lambda)\cdot 2+\lambda)\\ &= (1-\lambda)^2\cdot 1\cdot 2+\lambda(1-\lambda)\cdot 1+\lambda(1-\lambda)\cdot 2+\lambda^2\\ &= (1-\lambda)^2\cdot 2+3\lambda(1-\lambda)+\lambda^2\\ &= 2-4\lambda+2\lambda^2+3\lambda-3\lambda^2+\lambda^2 = 2-\lambda.\end{aligned}$$

$$\begin{aligned}&(1-\lambda)^{3-2}(3-1)!+\lambda(3-1)!\sum_{j=3}^{3}\left(\frac{(1-\lambda)^{3-j}}{(j-1)!}\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\\ &= 2(1-\lambda)+2\lambda\left(\frac{1}{2}((1-\lambda)\cdot 1+\lambda)\right) = 2-2\lambda+\lambda = 2-\lambda.\end{aligned}$$

Für den Induktionsschluss wird die Induktionsvoraussetzung für t herangezogen und aus dieser geschlossen, dass die Gleichheit auch für $t+1$ mit $t \geq 3$ gilt. Zu zeigen ist demnach folgende Gleichung:

$$\prod_{j=1}^{t}((1-\lambda)j+\lambda) \stackrel{!}{=} \left((1-\lambda)^{t-1}\,t!+\lambda\,t!\sum_{j=3}^{t+1}\left(\frac{(1-\lambda)^{t-j+1}}{(j-1)!}\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\right).$$

Mit Beginn auf der rechten Seite der Gleichung gilt:

$$\begin{aligned}&(1-\lambda)^{t-1}\,t!+\lambda\,t!\sum_{j=3}^{t+1}\left(\frac{(1-\lambda)^{t-j+1}}{(j-1)!}\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\\ &= (1-\lambda)^{t-2}(t-1)!\,(1-\lambda)t+\lambda(t-1)!\,t(1-\lambda)\left(\sum_{j=3}^{t}\left(\frac{(1-\lambda)^{t-j}}{(j-1)!}\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\right.\\ &\qquad\left.+\frac{1}{1-\lambda}\cdot\frac{1}{t!}\prod_{k=1}^{t-1}((1-\lambda)k+\lambda)\right)\\ &= (1-\lambda)t\left((1-\lambda)^{t-2}(t-1)!+\lambda(t-1)!\left(\sum_{j=3}^{t}\left(\frac{(1-\lambda)^{t-j}}{(j-1)!}\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\right.\right.\\ &\qquad\left.\left.+\frac{1}{1-\lambda}\cdot\frac{1}{t!}\prod_{k=1}^{t-1}((1-\lambda)k+\lambda)\right)\right)\\ &= (1-\lambda)t\left((1-\lambda)^{t-2}(t-1)!+\lambda(t-1)!\sum_{j=3}^{t}\left(\frac{(1-\lambda)^{t-j}}{(j-1)!}\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\right.\\ &\qquad\left.+\frac{\lambda}{(1-\lambda)t}\prod_{k=1}^{t-1}((1-\lambda)k+\lambda)\right)\end{aligned}$$

$$
\begin{aligned}
&= (1-\lambda)t\left((1-\lambda)^{t-2}(t-1)! + \lambda(t-1)!\sum_{j=3}^{t}\left(\frac{(1-\lambda)^{t-j}}{(j-1)!}\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\right)\\
&\quad + \frac{(1-\lambda)t\lambda}{(1-\lambda)t}\prod_{k=1}^{t-1}((1-\lambda)k+\lambda)\\
&\overset{\text{(IV)}}{=} (1-\lambda)t\prod_{j=1}^{t-1}((1-\lambda)j+\lambda) + \lambda\prod_{k=1}^{t-1}((1-\lambda)k+\lambda)\\
&= ((1-\lambda)t+\lambda)\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)\\
&= \prod_{j=1}^{t}((1-\lambda)j+\lambda).
\end{aligned}
$$

Insgesamt folgt für den Erwartungswert der Schätzfunktion:

$$\mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t)\right) = \mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right) = \boldsymbol{\mu}^{(c)} \text{ für } t \geq 3. \tag{6.12}$$

Die Schätzfunktion ist im Spezialfall $g(\boldsymbol{x}_i) = c$, $i = 1,\ldots,t$, (siehe Voraussetzung 4) also erwartungstreu für den wahren Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ der Klasse c. □

6.3.2 Situation: Linearer Trend der Erwartungswertvektoren

Auch für den Fall eines linearen Trends der Erwartungswertvektoren in den Klassen (6.1) wird der Spezialfall $g(\boldsymbol{x}_i) = c,\ i = 1,\ldots,t$, aus Voraussetzung 4 (Seite 138) betrachtet. Da für diesen Spezialfall die Erwartungstreue der Schätzfunktion $T_1^{(c),K}$ zum Zeitpunkt t für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ aus Klasse c des folgenden Zeitpunktes $t+1$ in der Datensituation unter concept drift, bzw. genauer eines linearen Trends der Erwartungswertvektoren widerlegt werden kann, kann auch im Allgemeinen bei linearem Trend der Erwartungswertvektoren nicht von Erwartungstreue ausgegangen werden.

Satz 7. Unter der Annahme eines linearen Trends (6.1) der Erwartungswertvektoren der Klassen ist die Schätzfunktion $T_1^{(c),K}$ von Zeitpunkt t, welche von Kuncheva und Plumpton (2008) vorgestellt wurde, nicht mehr erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c des kommenden Zeitpunktes $t+1$:

$$\mathrm{E}\left(T_1^{(c),K}\left(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t\right)\right) \neq \boldsymbol{\mu}_{t+1}^{(c)}.$$

Beweis. Der Erwartungswert der aus Voraussetzung 4 (Seite 138) resultierenden Schätzfunktion (6.10) berechnet sich unter der Annahme eines linearen Trends (6.1) folgendermaßen:

$$\mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t)\right) = \mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right)$$

$$= \frac{1}{(1-\lambda)(t-1)+\lambda} \left(\frac{(1-\lambda)^{t-1} \prod_{j=1}^{t-1} j}{\prod_{j=3}^{t} ((1-\lambda)(j-2)+\lambda)} \cdot \mathrm{E}\left(\boldsymbol{X}_1^{(c)}\right) \right.$$
$$\left. + \sum_{j=2}^{t-1} \frac{(1-\lambda)^{t-j}\lambda \prod_{k=j}^{t-1} k}{\prod_{k=j+1}^{t} ((1-\lambda)(k-2)+\lambda)} \cdot \mathrm{E}\left(\boldsymbol{X}_j^{(c)}\right) + \lambda \, \mathrm{E}\left(\boldsymbol{X}_t^{(c)}\right) \right)$$

$$= \frac{1}{(1-\lambda)(t-1)+\lambda} \cdot$$
$$\left(\frac{(1-\lambda)^{t-1}(t-1)!}{\prod_{j=1}^{t-2} ((1-\lambda)j+\lambda)} \cdot \boldsymbol{\mu}_1^{(c)} + \sum_{j=2}^{t-1} \frac{\frac{(t-1)!}{(j-1)!} \cdot (1-\lambda)^{t-j}\lambda}{\prod_{k=j-1}^{t-2} ((1-\lambda)k+\lambda)} \cdot \boldsymbol{\mu}_j^{(c)} + \lambda \boldsymbol{\mu}_t^{(c)} \right)$$

$$\overset{(6.1)}{=} \frac{1}{(1-\lambda)(t-1)+\lambda} \left(\frac{(1-\lambda)^{t-1}(t-1)!}{\prod_{j=1}^{t-2} ((1-\lambda)j+\lambda)} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}\right) \right.$$
$$\left. + \sum_{j=2}^{t-1} \frac{\frac{(t-1)!}{(j-1)!} \cdot (1-\lambda)^{t-j}\lambda}{\prod_{k=j-1}^{t-2} ((1-\lambda)k+\lambda)} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j\right) + \lambda \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} t\right) \right)$$

$$= \underbrace{\boldsymbol{\beta}_0^{(c)} \left(\frac{(1-\lambda)^{t-1}(t-1)!}{\prod_{j=1}^{t-1} ((1-\lambda)j+\lambda)} + \sum_{j=2}^{t-1} \frac{\frac{(t-1)!}{(j-1)!} \cdot (1-\lambda)^{t-j}\lambda}{\prod_{k=j-1}^{t-1} ((1-\lambda)k+\lambda)} + \frac{\lambda}{(1-\lambda)(t-1)+\lambda} \right)}_{(*)}$$
$$+ \boldsymbol{\beta}_1^{(c)} \left(\frac{(1-\lambda)^{t-1}(t-1)!}{\prod_{j=1}^{t-1} ((1-\lambda)j+\lambda)} + \sum_{j=2}^{t-1} \frac{\frac{(t-1)!}{(j-1)!} \cdot (1-\lambda)^{t-j}\lambda j}{\prod_{k=j-1}^{t-1} ((1-\lambda)k+\lambda)} + \frac{\lambda t}{(1-\lambda)(t-1)+\lambda} \right).$$

Der erste Summand $(*)$ ergibt $\boldsymbol{\beta}_0^{(c)}$, da der Ausdruck in Klammern sich zu Eins vereinfachen lässt. Der Beweis kann analog wie bei der Berechnung des Erwartungswertes der Schätzfunktion im Falle einer stabilen Verteilung ab Seite 148 mittels vollständiger Induktion erfolgen (vgl. dazu die Struktur von (6.11)).

Es gilt also weiter:

$$\mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)\right) = \mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})\right) \tag{6.13}$$
$$= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} \left(\frac{(1-\lambda)^{t-1}(t-1)!}{\prod_{j=1}^{t-1} ((1-\lambda)j+\lambda)} + \sum_{j=2}^{t-1} \frac{\frac{(t-1)!}{(j-1)!} \cdot (1-\lambda)^{t-j}\lambda j}{\prod_{k=j-1}^{t-1} ((1-\lambda)k+\lambda)} + \frac{\lambda t}{(1-\lambda)(t-1)+\lambda} \right).$$

Damit die Schätzfunktion erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose bzw. des kommenden Zeitpunktes $t+1$ wäre, müsste der hintere Teil in Klammern $t+1$ sein, da somit $\mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right) = \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(t+1) \overset{(6.1)}{=} \boldsymbol{\mu}_{t+1}^{(c)}$ wäre.

Im Folgenden wird gezeigt, dass dies nicht im Allgemeinen gilt. Sei dazu nur der in Klammern hintere Teil des Erwartungswertes (6.13) betrachtet. Dieser lässt sich umformen zu:

$$
\begin{aligned}
&\frac{(1-\lambda)^{t-1}(t-1)!}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{t-1} \frac{\frac{(t-1)!}{(j-1)!}\cdot(1-\lambda)^{t-j}\lambda j}{\prod_{k=j-1}^{t-1}((1-\lambda)k+\lambda)} + \frac{\lambda t}{(1-\lambda)(t-1)+\lambda} \\
&= \frac{(1-\lambda)^{t-1}(t-1)! + \lambda t\prod_{j=1}^{t-2}((1-\lambda)j+\lambda)}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{t-1} \frac{\frac{(t-1)!}{(j-1)!}\cdot(1-\lambda)^{t-j}\lambda j}{\prod_{k=j-1}^{t-1}((1-\lambda)k+\lambda)} \\
&= \frac{(1-\lambda)^{t-1}(t-1)! + \lambda t\prod_{j=1}^{t-2}((1-\lambda)j+\lambda)}{\prod_{j=1}^{t-1}((1-\lambda)j+\lambda)} + \sum_{j=3}^{t-1} \frac{\frac{(t-1)!}{(j-1)!}\cdot(1-\lambda)^{t-j}\lambda j}{\prod_{k=j-1}^{t-1}((1-\lambda)k+\lambda)} \\
&\quad + \frac{2(t-1)!\,(1-\lambda)^{t-2}\lambda}{\prod_{k=1}^{t-1}((1-\lambda)k+\lambda)} \\
&= \frac{1}{\prod_{i=1}^{t-1}((1-\lambda)i+\lambda)} \Bigg((1-\lambda)^{t-2}(t-1)!\,(1+\lambda) + \lambda t\prod_{j=1}^{t-2}((1-\lambda)j+\lambda) \\
&\qquad + \sum_{j=3}^{t-1}\left((1-\lambda)^{t-j}\cdot\frac{(t-1)!}{(j-1)!}\cdot\lambda j\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\Bigg) \\
&= \frac{1}{\prod_{i=1}^{t-1}((1-\lambda)i+\lambda)} \Bigg((1-\lambda)^{t-2}(t-1)!\,(1+\lambda) \\
&\qquad + \sum_{j=3}^{t}\left((1-\lambda)^{t-j}\cdot\frac{(t-1)!}{(j-1)!}\cdot\lambda j\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\Bigg) \\
&= \frac{1}{\prod_{i=1}^{t-1}((1-\lambda)i+\lambda)} \Bigg((1-\lambda)^{t-2}(t-1)!\,(1+\lambda) \\
&\qquad + \lambda(t-1)!\sum_{j=3}^{t}\left(\frac{j(1-\lambda)^{t-j}}{(j-1)!}\prod_{k=1}^{j-2}((1-\lambda)k+\lambda)\right)\Bigg).
\end{aligned}
$$

Sei nun der minimal mögliche Zeitpunkt $t=3$ betrachtet. In diesem Fall beträgt der hintere Teil des Erwartungswertes

$$
\begin{aligned}
&\frac{1}{((1-\lambda)\cdot 1+\lambda)((1-\lambda)\cdot 2+\lambda)}\left((1-\lambda)\cdot 2!\,(1+\lambda) + \lambda\cdot 2!\,\left(\frac{3}{2!}((1-\lambda)\cdot 1+\lambda)\right)\right) \\
&= \frac{2(1-\lambda)(1+\lambda)+3\lambda}{2-\lambda} = \frac{2-2\lambda^2+3\lambda}{2-\lambda} = \frac{(2-\lambda)(2\lambda+1)}{2-\lambda} = 1+2\lambda.
\end{aligned}
$$

Dies ist eine lineare Funktion in λ. Da $\lambda \in (0,1)$, beträgt der Wertebereich für $t = 3$: $1 + 2\lambda \in (1,3)$. Der Wert $t + 1 = 4$ kann demnach nicht angenommen werden.

Für $t = 4$ gilt:

$$\begin{aligned}
&\frac{1}{((1-\lambda)+\lambda)(2(1-\lambda)+\lambda)(3(1-\lambda)+\lambda)} \cdot \\
&\left((1-\lambda)^2 \cdot 3!\,(1+\lambda) + \lambda \cdot 3!\,\left(\frac{3(1-\lambda)}{2!}((1-\lambda)+\lambda) + \frac{4}{3!}((1-\lambda)+\lambda)(2(1-\lambda)+\lambda)\right)\right) \\
&= \frac{1}{(2-\lambda)(3-2\lambda)}\left(6(1-\lambda)^2(1+\lambda) + 6\lambda\left(\frac{3(1-\lambda)}{2} + \frac{2(2-\lambda)}{3}\right)\right) \\
&= \frac{6 - 12\lambda + 6\lambda^2 + 6\lambda - 12\lambda^2 + 6\lambda^3 + 9\lambda - 9\lambda^2 + 8\lambda - 4\lambda^2}{(2-\lambda)(3-2\lambda)} \\
&= \frac{6 + 11\lambda - 19\lambda^2 + 6\lambda^3}{(2-\lambda)(3-2\lambda)} = \frac{(2-\lambda)(3-2\lambda)(1+3\lambda)}{(2-\lambda)(3-2\lambda)} = 1 + 3\lambda \in (1,4).
\end{aligned}$$

Analog kann hier der Wert $t + 1 = 5$ nicht angenommen werden.

Es gilt somit für alle möglichen Gewichte $\lambda \in (0,1)$:

$$\begin{aligned}
&\mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)}, \boldsymbol{X}_2^{(c)}, \boldsymbol{X}_3^{(c)})\right) = \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(1+2\lambda) \neq \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(3+1) \overset{(6.1)}{=} \boldsymbol{\mu}_4^{(c)}, \\
&\mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)}, \boldsymbol{X}_2^{(c)}, \boldsymbol{X}_3^{(c)}, \boldsymbol{X}_4^{(c)})\right) = \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(1+3\lambda) \neq \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(4+1) \overset{(6.1)}{=} \boldsymbol{\mu}_5^{(c)}.
\end{aligned}$$

Da sich also mindestens zwei Gegenbeispiele durch die Zeitpunkte $t = 3$ und $t = 4$ finden lassen, kann die Erwartungstreue nicht allgemein für beliebige Zeitpunkte gelten. Da dies zudem bereits für den Spezialfall von ausschließlich auftretenden Beobachtungen aus Klasse c im Datenstrom bis zum Zeitpunkt t (Voraussetzung 4) gilt, ist die Schätzfunktion $T_1^{(c),K}$ im Falle eines linearen Trends (6.1) der Erwartungswertvektoren nicht mehr erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose der Klasse c zum Zeitpunkt $t + 1$. □

6.4 Online Diskriminanzanalyse mit adaptivem Vergessen

Bei der Online Diskriminanzanalyse mit adaptivem Vergessen von Anagnostopoulos et al. (2012) (Abschnitt 4.4) lässt sich der Schätzwert für den Erwartungswertvektor von Klasse c im Falle der Betrachtung von exponentiellem Vergessen mithilfe von (4.33), (4.34) und (4.37) aus Abschnitt 4.4 ausformuliert iterativ definieren:

$$\begin{aligned}
\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} &\overset{(4.34)}{=} \frac{1}{N_t^{(c)}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} v_i^{(c)} \boldsymbol{x}_i \overset{(4.37)}{=} \frac{1}{\sum\limits_{\substack{k:\, g(\boldsymbol{x}_k)=c \\ k \leq t-1}} v_k^{(c)} + 1} \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} v_i^{(c)} \boldsymbol{x}_i + 1 \cdot \boldsymbol{x}_t \right) \\
&\overset{(4.33)}{=} \frac{1}{\sum\limits_{\substack{k:\, g(\boldsymbol{x}_k)=c \\ k \leq t-1}} \left(\prod\limits_{j=\sum_{l=1}^{k} \mathbb{1}_{\{g(\boldsymbol{x}_l)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) + 1} \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} \left(\prod_{j=\sum_{l=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_l)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) \boldsymbol{x}_i + \boldsymbol{x}_t \right).
\end{aligned}$$

In beiden folgenden Situationen in Abschnitt 6.4.1 und 6.4.2 wird wie in Abschnitt 6.3 der Spezialfall betrachtet, dass bis zum Zeitpunkt t im Datenstrom alle Beobachtungen in Klasse c realisiert werden: $g(\boldsymbol{x}_i) = c,\ i = 1, \ldots, t$ (vgl. Voraussetzung 4 auf Seite 138).

Unter Voraussetzung 4 lässt sich der Schätzwert vereinfachen zu

$$\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}} = \frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)\boldsymbol{x}_i + \boldsymbol{x}_t\right).$$

Die entsprechende Schätzfunktion für $\boldsymbol{\mu}^{(c)}$ sieht dann folgendermaßen aus:

$$\begin{aligned} T_1^{(c),A}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) &:= T_1^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)}) \\ &= \frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)\boldsymbol{X}_i^{(c)} + \boldsymbol{X}_t^{(c)}\right). \end{aligned} \tag{6.14}$$

6.4.1 Situation: Stabile Verteilung

Im Falle einer stabilen Verteilung in Klasse c über die Zeit gilt für die Erwartungswertvektoren (vgl. Seite 138): $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_1^{(c)} = \boldsymbol{\mu}_2^{(c)} = \ldots$.

Satz 8. Unter der Annahme einer stabilen Verteilung und dem Spezialfall aus Voraussetzung 4 (Seite 138) ist die Schätzfunktion $T_1^{(c),A}$ aus (6.14), welche von Anagnostopoulos et al. (2012) herangezogen wird, zum Zeitpunkt t erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ des kommenden Zeitpunktes $t+1$ von Klasse c:

$$\mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)\right) = \mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})\right) = \boldsymbol{\mu}^{(c)}.$$

Beweis.

$$\begin{aligned} &\mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)\right) = \mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})\right) \\ &= \frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)\mathrm{E}\left(\boldsymbol{X}_i^{(c)}\right) + \mathrm{E}\left(\boldsymbol{X}_t^{(c)}\right)\right) \\ &= \frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)\boldsymbol{\mu}_i^{(c)} + \boldsymbol{\mu}_t^{(c)}\right) \\ &= \frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)\boldsymbol{\mu}^{(c)} + \boldsymbol{\mu}^{(c)}\right) \\ &= \frac{\boldsymbol{\mu}^{(c)}}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)+1\right) = \boldsymbol{\mu}^{(c)}. \end{aligned} \tag{6.15}$$

□

6.4.2 Situation: Linearer Trend der Erwartungswertvektoren

Auch für die Situation unter concept drift bzw. den Fall eines linearen Trends (6.1) der Erwartungswertvektoren in den Klassen wird der Spezialfall aus Voraussetzung 4 (Seite 138) betrachtet, dass bis zum Zeitpunkt t nur Beobachtungen in Klasse c realisiert werden.

Unter dieser Voraussetzung kann gezeigt werden, dass die Schätzfunktion $T_1^{(c),A}$ zum Zeitpunkt t im Allgemeinen nicht mehr erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ aus Klasse c des folgenden Zeitpunktes $t+1$ ist. Da die Erwartungstreue bereits in diesem Spezialfall nicht gilt, kann auch im allgemeinen Fall keine Erwartungstreue der Schätzfunktion im Falle eines linearen Trends der Erwartungswertvektoren in den Klassen angenommen werden. Dies drückt der folgende Satz aus.

Satz 9. Unter der Annahme eines linearen Trends (6.1) der Erwartungswertvektoren der Klassen ist die Schätzfunktion $T_1^{(c),A}$ zum Zeitpunkt t, welche von Anagnostopoulos et al. (2012) vorgestellt wurde, nicht mehr erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c des kommenden Zeitpunktes $t+1$:

$$\mathrm{E}\left(T_1^{(c),A}\left(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t\right)\right) \neq \boldsymbol{\mu}_{t+1}^{(c)}.$$

Beweis. Im Falle eines linearen Trends (6.1) der Erwartungswertvektoren in den Klassen gilt unter der Voraussetzung 4 für den Erwartungswert der Schätzfunktion (6.14):

$$\begin{aligned}
&\mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t)\right) = \mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right)\\
&= \frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)\mathrm{E}\left(\boldsymbol{X}_i^{(c)}\right)+\mathrm{E}\left(\boldsymbol{X}_t^{(c)}\right)\right)\\
&= \frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)\boldsymbol{\mu}_i^{(c)}+\boldsymbol{\mu}_t^{(c)}\right)\\
&\overset{(6.1)}{=} \frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}i\right)+\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}t\right)\right)\\
&= \frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\boldsymbol{\beta}_0^{(c)}\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)+\boldsymbol{\beta}_1^{(c)}\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)i+\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}t\right)\right)\\
&= \boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}\cdot\underbrace{\frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)i+t\right)}_{(*)}. \qquad (6.16)
\end{aligned}$$

Damit die Schätzfunktion erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose bzw. des kommenden Zeitpunktes $t+1$ wäre, müsste der hintere Teil $(*)$ $t+1$ sein, da in diesem Fall $\mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right) = \boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}(t+1) \overset{(6.1)}{=} \boldsymbol{\mu}_{t+1}^{(c)}$.

Der Wertebereich der Faktoren beträgt $\lambda_{(i)}^{(c)} \in [0,1]$, wodurch die Likelihood Terme vergangener Beobachtungen ein zunehmend geringeres (bzw. zumindest kein größeres) Gewicht (exponentielles Vergessen) bekommen. Der Wertebereich der einzelnen Komponenten von $(*)$ sieht dadurch folgendermaßen aus:

$$\underbrace{\frac{1}{\underbrace{\sum_{k=1}^{t-1}\underbrace{\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)}_{\in[0,1]}+1}_{\text{Nenner}\in[1,t],\ \text{Bruch}\in\left[\frac{1}{t},1\right]}}}_{}\underbrace{\left(\sum_{i=1}^{t-1}\underbrace{\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)}_{\in[0,i]}i+t\right)}_{\in\left[0,\frac{(t-1)t}{2}\right]\ \in\left[t,\frac{t(t+1)}{2}\right]}.$$

Das Gewicht ist eine monoton steigende Funktion über die Zeitpunkte $1,\ldots,t-1$, da aufgrund der Produktbildung über die Faktoren für $i,j \in \{1,\ldots,t-1\}$ mit $i \leq j$

$$v_i^{(c)} = \prod_{k=i}^{t-1}\lambda_{(k)}^{(c)} \leq \prod_{k=j}^{t-1}\lambda_{(k)}^{(c)} = v_j^{(c)}$$

gilt. Die Multiplikation mit dem jeweiligen Zeitpunkt ändert nichts an der Monotonie:

$$\left(\prod_{k=i}^{t-1}\lambda_{(k)}^{(c)}\right)i \leq \left(\prod_{k=j}^{t-1}\lambda_{(k)}^{(c)}\right)j.$$

Das Gewicht als Produkt der Faktoren ist sowohl im Vorfaktor als auch in der Summe von $(*)$ identisch enthalten. Die Wertebereiche des Vorfaktors $\left[\frac{1}{t},\,1\right]$ und der Summe $\left[t,\,\frac{t(t+1)}{2}\right]$ überschneiden sich für $t > 1$ nicht und enthalten beide jeweils nur positive Zahlen. Für ein größeres Gewicht $v_i^{(c)} = \prod_{k=i}^{t-1}\lambda_{(k)}^{(c)}$ nimmt die Summe größere Werte aus ihrem Wertebereich an, während der Vorfaktor kleinere Werte annimmt.

Insgesamt lässt sich mit diesen Überlegungen der gesamte Wertebereich von $(*)$ herleiten:

$$\frac{1}{\sum_{k=1}^{t-1}\left(\prod_{j=k}^{t-1}\lambda_{(j)}^{(c)}\right)+1}\left(\sum_{i=1}^{t-1}\left(\prod_{j=i}^{t-1}\lambda_{(j)}^{(c)}\right)i+t\right) \in \left[\frac{1}{t}\cdot\frac{t(t+1)}{2},\,1\cdot t\right] = \left[\frac{t+1}{2},\,t\right].$$

Der Wert $t+1$ kann nicht angenommen werden. Die Schätzung des Erwartungswertvektors $\boldsymbol{\mu}_{t+1}^{(c)}$ durch die Schätzfunktion $T_1^{(c),A}$ hängt im Falle eines linearen Trends der Erwartungswertvektoren immer zeitlich hinterher.

Abhängig von den Faktoren $\lambda_{(j)}^{(c)}$ und damit Gewichten gilt:

$$\mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right) \in \left[\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}\cdot\frac{t+1}{2},\,\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}t\right] \overset{(6.1)}{=} \left[\boldsymbol{\mu}_{\frac{t+1}{2}}^{(c)},\boldsymbol{\mu}_t^{(c)}\right] \not\ni \boldsymbol{\mu}_{t+1}^{(c)}.$$

Da bereits im Spezialfall $g(\boldsymbol{x}_i) = c$, $i = 1,\ldots,t$, (s. Voraussetzung 4 auf Seite 138) keine Erwartungstreue der Schätzfunktion $T_1^{(c),A}$ vorliegt, kann geschlussfolgert werden, dass diese Eigenschaft auch allgemein nicht gilt. □

7 Verbesserung der Prognosegüte bei Concept Drift

Im vorherigen Kapitel wurde deutlich, dass die betrachteten Methoden für Online Diskriminanzanalyse gute theoretische Eigenschaften aufweisen, wenn das concept bzw. die Verteilung über die Zeit stabil ist, das heißt Erwartungswertvektoren und Kovarianzmatrizen der Verteilungen in den einzelnen Klassen mit der Zeit unveränderlich bleiben. Es ist zu vermuten, dass die jeweiligen Schätzfunktionen für die Erwartungswertvektoren der Klassen unter diesen gegebenen Annahmen der Verteilung im allgemeinen Fall erwartungstreu sind, auch für Erwartungswertvektoren zukünftiger Zeitpunkte. Für die Schätzfunktion $T_1^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ aus *Sequential ILDA* konnte dies in Abschnitt 6.2.1 (vgl. Seite 139 ff.) bewiesen werden. Für die Schätzfunktion $T_1^{(c),K}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ aus *OLDC* bzw. $T_1^{(c),A}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ aus der *Diskriminanzanalyse mit adaptivem Vergessen* konnte der Beweis der Erwartungstreue für den Spezialfall, dass bis zum Zeitpunkt t nur Beobachtungen in Klasse c vorkommen ($g(\boldsymbol{x}_i) = c$, $i = 1, \ldots, t$), in Abschnitt 6.3.1 (vgl. Seite 148 ff.) bzw. in Abschnitt 6.4.1 (vgl. Seite 155) erfolgen.

Falls jedoch ein concept drift vorliegt, ist diese Eigenschaft der (vermuteten) Erwartungstreue der Schätzfunktionen nicht mehr gewährleistet (vgl. Abschnitte 6.2.2, 6.3.2 und 6.4.2). Da die resultierenden Schätzer in der Klassifikationsregel der (Linearen) Diskriminanzanalyse herangezogen werden, kann sich die Prognosegüte der Klassifikationsregel stark verschlechtern, wenn die Klassen zukünftiger Zeitpunkte prognostiziert werden sollen.

Im Folgenden wird daher eine Methodik entwickelt, mithilfe derer bestehende Methoden für Online Diskriminanzanalyse erweitert werden können. Diese Erweiterung zielt auf eine Verbesserung der Prognosegüte der stetig aktualisierten Klassifikatoren beim Hinzuziehen neuer Beobachtungen im Datenstrom im Fall des Vorliegens von concept drift. Die Idee wurde bereits in kurzer Form exemplarisch für die Methode *OLDC* mit fester Lernrate $\lambda = 0.5$ (Kuncheva und Plumpton, 2008) in Schnackenberg et al. (2018) vorgestellt.

7.1 Problematik

Es existieren bereits einige Methoden für Online Diskriminanzanalyse, mithilfe derer Klassifikatoren im Datenstrom schrittweise aktualisiert werden können. Insbesondere auch Methoden, welche eine Anpassung an einen eventuellen concept drift beachten und zulassen

(vgl. Kapitel 4). Wird jedoch ein zeitlicher (linearer) Trend der Erwartungswertvektoren der Klassen unterstellt, so verliert der bisherige Schätzer $\boldsymbol{m}^{(c)}_{n_t^{(c)}}$ an Güte hinsichtlich der Prognose zukünftiger Beobachtungen durch die Klassifikationsregel der Diskriminanzanalyse. Dadurch, dass in die Klassifikationsregel nur bisherige Beobachtungen einbezogen werden, fließt in die Prognose ein „veralteter“ Schätzer für die Erwartungswertvektoren der Klassen ein, falls ein (linearer) Trend der Erwartungswerte angenommen wird. Dies ist auch der Fall bei Methoden, welche eine Anpassung des Klassifikators an concept drift berücksichtigen, da trotz allem ein zukünftiger Trend nicht beachtet wird. Vereinfacht gesagt und anschaulich bedeutet dies, dass die Prognose einer neuen Beobachtung zeitlich immer etwas „hinterher hängt“. In diesem Kapitel wird daher eine Idee entwickelt, um die Prognosegüte der Klassifikatoren unter bestimmten Annahmen zu verbessern.

7.2 Modellannahmen

Es wird ein concept drift bezüglich der Erwartungswertvektoren der einzelnen Klassen $\boldsymbol{\mu}_i^{(c)}$, $c = 1, \ldots, M$, $i = 1, \ldots,$ unterstellt. Spezieller wird der Fall betrachtet, dass die Erwartungswertvektoren der Klassen $c = 1, \ldots, M$ einem linearen Trend in Abhängigkeit der Zeit folgen, dass also gilt (vgl. auch Seite 138):

$$\boldsymbol{\mu}_i^{(c)} = \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} i, \quad i = 1, \ldots. \tag{7.1}$$

Dies bedeutet, dass sich die Erwartungswertvektoren in einer Klasse c für zwei verschiedene Zeitpunkte $t \neq s$ unterscheiden können und somit für $\boldsymbol{\beta}_1^{(c)} \neq \mathbf{0}$:

$$\boldsymbol{\mu}_t^{(c)} \neq \boldsymbol{\mu}_s^{(c)}, \quad t \neq s.$$

Im speziellen Fall des kontinuierlichen linearen Trends lässt sich dies genauer formulieren:

$$\exists\, j \in \{1, \ldots, p\} : \left[\boldsymbol{\mu}_t^{(c)}\right]_j = \left[\boldsymbol{\mu}_s^{(c)}\right]_j + (t - s) \left[\boldsymbol{\beta}_1^{(c)}\right]_j \quad \forall\, s < t.$$

Dies bedeutet, dass sich die Erwartungswertvektoren in mindestens einer der p Dimensionen linear mit der Zeit verändern.

Bezüglich der Kovarianzmatrizen aller M Klassen wird in Hinblick auf die Annahmen der LDA angenommen, dass diese identisch für alle Klassen sind ((LDA2) auf Seite 39/ (FDA2) auf Seite 45). Zusätzlich seien die Kovarianzmatrizen unverändert über die Zeit, es wird also kein concept drift bezüglich der Varianzen und Kovarianzen unterstellt.

Solch ein linearer Trend (7.1) des Erwartungswertvektors über die Zeit ist in Abbildung 7.1 für eine zweidimensionale Verteilung einer einzelnen Klasse $c = 1$ veranschaulicht. In diesem Beispiel unterliegen beide Dimensionen einem linearen Trend, die Kovarianzmatrix bleibt über die Zeit jedoch unverändert.

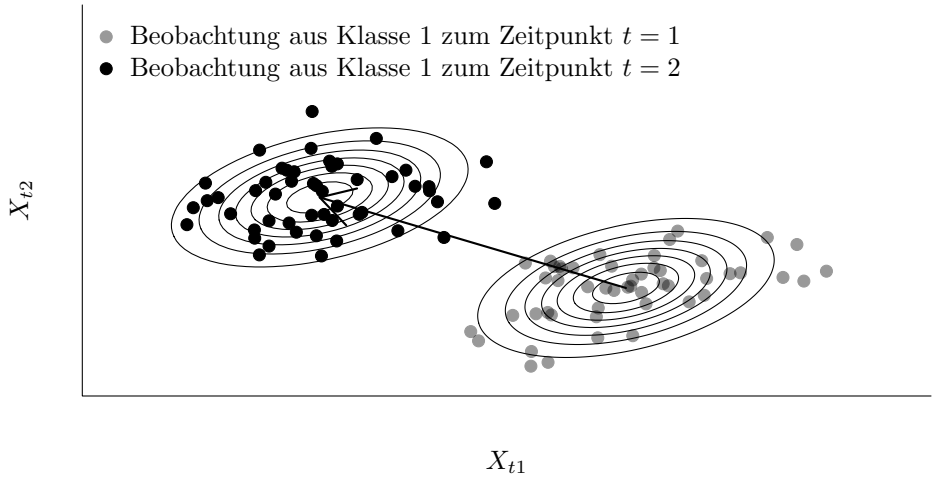


Abbildung 7.1: Beispiel für linearen Trend des Erwartungswertvektors einer zweidimensionalen Verteilung für eine Klasse von einem zum nächsten Zeitpunkt. Die Kovarianzmatrix bleibt über die Zeit unverändert.

7.3 Modellierung von zeitabhängigem Concept Drift durch lokales lineares Regressionsmodell

In diesem Abschnitt wird eine Methodik entwickelt, mithilfe derer die Prognosegüte der stetig aktualisierten Klassifikatoren bei der Online Diskriminanzanalyse im Fall des Vorliegens von concept drift unter den Modellannahmen aus Abschnitt 7.2 verbessert werden kann. Als erster Schritt wird dazu der zeitabhängige concept drift modelliert. Dieses Modell wird später (in Abschnitt 7.5) herangezogen, um bestehende Klassifikatoren zu verbessern.

Die Idee der Verbesserung besteht darin, dass unter oben beschriebenen Modellannahmen der jeweilige Maximum-Likelihood-Schätzer $\boldsymbol{m}^{(c)}_{n_t^{(c)}}$ für den Erwartungswertvektor der Klasse c der verschiedenen Methoden in der jeweiligen Klassifikationsregel der Diskriminanzanalyse ((3.45) bzw. (3.47) für Fisher LDA (*Sequential ILDA*), (4.30) für *OLDC*, (4.63) für *QDA-AF* und (4.64) für *LDA-AF*) durch einen „besser geeigneten“ Schätzer basierend auf der Vorhersage eines lokalen linearen Regressionsmodells ersetzt wird. Mittels dieser „besseren“ Schätzung der Erwartungswerte der Klassen des zukünftigen Zeitpunktes soll das in Abschnitt 7.1 beschriebene Problem gelöst werden.

Zunächst wird dazu der zeitabhängige concept drift, also im Speziellen der Trend der Erwartungswerte, mittels eines lokalen linearen Regressionsmodells modelliert. Die lokal beschränkte Betrachtung hat den Vorteil, dass in der Praxis auch beliebige nicht-lineare Drifts linear approximiert werden können. Die Klassenmittel $\boldsymbol{m}^{(c)}_{n_t^{(c)}}$ werden dabei als Schätzer für die Erwartungswertvektoren herangezogen. Es handelt sich hierbei um ML-Schätzer, die im Falle keiner betrachteten Lernrate oder Gewichtung unverzerrt sind.

Sei zunächst der Spezialfall betrachtet, dass bei der Klassifikationsmethode keine Anpassung an einen concept drift erfolgt. Dies sind die Methoden *Sequential Incremental LDA*

(Abschnitt 4.2) und *OLDC* mit fester Lernrate $\lambda = 1/2$ (Abschnitt 4.3). Die Verallgemeinerung auf alle Methoden inklusive Lernrate folgt in Abschnitt 7.5.

Zum Zeitpunkt t werden die Mittelwertvektoren aus den letzten N_{trend} Aktualisierungsschritten der LDA betrachtet. Um den linearen Trend der Erwartungswertvektoren zu modellieren, wird ein lokales lineares Regressionsmodell an jene Mittelwertvektoren – repräsentativ für die Erwartungswertvektoren – der Zeitpunkte aus dieser Menge angepasst, zu denen wirklich ein Update erfolgt ist, also eine Beobachtung aus Klasse c im Datenstrom aufgetreten ist. Da nicht zu jedem Zeitpunkt eine Beobachtung in Klasse c im Datenstrom realisiert wird, seien dies nach Definition $n_{\text{trend}}^{(c)} \leq N_{\text{trend}}$ Beobachtungen:[1]

$$I = \underbrace{\{k : g(\boldsymbol{x}_k) = c,\ k = t - N_{\text{trend}} + 1, \ldots, t\}}_{n_{\text{trend}}^{(c)} \text{ Zeitpunkte/Beobachtungen}}. \tag{7.2}$$

Zu beachten sei hier, dass $n_{\text{trend}}^{(c)}$ zufällig ist, da die Beobachtungen im Datenstrom zufällig in den verschiedenen Klassen auftreten. Im Folgenden wird daher Annahme 1 betrachtet, da nur in diesem Fall die Anpassung eines linearen Regressionsmodells sinnvoll ist:

> **Annahme 1.** Zum Zeitpunkt t seien mindestens zwei Beobachtungen aus Klasse c in dem Intervall $t - N_{\text{trend}} + 1, \ldots, t$ realisiert, d. h. $n_{\text{trend}}^{(c)} \geq 2$. Falls dies nicht der Fall ist, wird kein lineares Regressionsmodell angepasst.

Implementierungstechnisch bedeutet dies, dass zum Zeitpunkt t die Menge von $n_{\text{trend}}^{(c)}$ aktualisierten Mittelwerten betrachtet wird. Nur falls $n_{\text{trend}}^{(c)} \geq 2$, wird ein lokales lineares Regressionsmodell angepasst. Andernfalls wird das bisherige Regressionsmodell betrachtet bzw. falls bisher keines modelliert wurde, werden weiterhin die standardmäßigen Mittelwertschätzer aus der LDA in der Klassifikationsregel der Diskriminanzanalyse verwendet.

Das lokale lineare Regressionsmodell sieht für $c = 1, \ldots, M$ folgendermaßen aus:

$$\boldsymbol{m}_{n_i^{(c)}}^{(c)} = \boldsymbol{\beta}_{0t}^{(c)} + \boldsymbol{\beta}_{1t}^{(c)} z_i^{(c)} + \boldsymbol{\epsilon}_i^{(c)}, \quad i \in I. \tag{7.3}$$

Bei der Modellierung wird jede der $j \in \{1, \ldots, p\}$ Dimensionen unabhängig betrachtet. Es wird also ein einzelnes lineares Regressionsmodell für jede der p Dimensionen unterstellt:

$$\left[\boldsymbol{m}_{n_i^{(c)}}^{(c)}\right]_j = \left[\boldsymbol{\beta}_{0t}^{(c)}\right]_j + \left[\boldsymbol{\beta}_{1t}^{(c)}\right]_j z_i^{(c)} + \epsilon_{ij}^{(c)}, \quad i \in I. \tag{7.4}$$

In Matrixschreibweise für alle $i \in I$ Beobachtungen gleichzeitig formuliert gilt:

$$\boldsymbol{m}_j^{(c)} = \boldsymbol{Z}_j^{(c)} \begin{pmatrix} \left[\boldsymbol{\beta}_{0t}^{(c)}\right]_j \\ \left[\boldsymbol{\beta}_{1t}^{(c)}\right]_j \end{pmatrix} + \boldsymbol{\epsilon}_j^{(c)} \Leftrightarrow: \underbrace{\boldsymbol{m}_j^{(c)}}_{(n_{\text{trend}}^{(c)} \times 1)\text{-dim.}} = \underbrace{\boldsymbol{Z}_j^{(c)}}_{(n_{\text{trend}}^{(c)} \times 2)\text{-dim.}} \underbrace{\boldsymbol{\beta}_j^{(c)}}_{(2 \times 1)\text{-dim.}} + \underbrace{\boldsymbol{\epsilon}_j^{(c)}}_{(n_{\text{trend}}^{(c)} \times 1)\text{-dim.}}, \tag{7.5}$$

wobei die erste Spalte von $\boldsymbol{Z}_j^{(c)}$ aus Einsen besteht.

[1] Wenn in jedem der letzten N_{trend} Aktualisierungsschritte eine neue Beobachtung aus Klasse c eingeflossen ist, gilt $n_{\text{trend}}^{(c)} := N_{\text{trend}}$.

Vielfach wird für lineare Regressionsmodelle die Annahme einer nicht-stochastischen Designmatrix unterstellt. Durch diese Annahme werden viele Beweise theoretischer Eigenschaften des Modells vereinfacht, u. a. jener des Gauß-Markov-Theorems (Toutenburg, 2003, S. 102/108 f.; Groß, 2003, S. 34/51):

Satz 10. *Gauß-Markov-Theorem*
Im klassischen linearen Regressionsmodell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ist der gewöhnliche Kleinste-Quadrate-Schätzer (KQ-Schätzer)

$$\hat{\boldsymbol{\beta}}_{\mathrm{KQ}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

unter den Annahmen

(1) $\boldsymbol{X}$ ist nicht-stochastisch mit vollem Spaltenrang,
(2) die Elemente des Vektors $\boldsymbol{y}$ sind (beobachtbare) Zufallsvariablen,
(3) $\boldsymbol{\epsilon} \sim \left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right)$, also $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ und $\mathrm{Cov}(\boldsymbol{\epsilon}) = \mathrm{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right) = \sigma^2\boldsymbol{I}$ mit $\sigma^2 > 0$,

die beste (homogene) lineare erwartungstreue Schätzung des wahren Parametervektors $\boldsymbol{\beta}$. Dies bedeutet, dass der Schätzer minimale Varianz unter allen linearen erwartungstreuen Schätzern aufweist mit $\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}\right) = \sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$.

Diese Eigenschaft wird auch häufig als BLUE (best linear unbiased estimator) bezeichnet (Fahrmeir et al., 1996b, S. 99).

Diese Annahme fester, nicht zufälliger Werte der Einflussvariablen, also nicht-stochastischer Regressoren aus Annahme (1), kann jedoch nur in geplanten Experimenten erfüllt werden. In Beobachtungsstudien resultieren die Werte der Regressoren zufällig als Realisationen von Zufallsvariablen. So sind auch hier die $z_i^{(c)}$ aus (7.3) bzw. (7.4) zufällig und zwar als Resultat des zufälligen Auftretens oder Nicht-Auftretens einer Beobachtung in Klasse c zu den Zeitpunkten $i \in I$ (siehe nächste Seite).

Fahrmeir et al. (2009, S. 61 ff.) weisen darauf hin, dass das lineare Regressionsmodell auch bei (1*) stochastischen Regressoren angewandt werden kann. Wenn die Verteilung der Fehlerterme bedingt auf die Designmatrix $\boldsymbol{X}$ betrachtet wird, lassen sich die theoretischen Eigenschaften des KQ-Schätzers weiterhin beweisen. Für stochastische Regressoren $\boldsymbol{X}$ und der entgegen (3) strengeren Annahme (3*) $\boldsymbol{\epsilon}|\boldsymbol{X} \sim \left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right)$ gilt das Gauß-Markov-Theorem weiterhin. Diese Annahme leitet sich aus unterstellter Unabhängigkeit der bedingten Verteilung der Fehler $\boldsymbol{\epsilon}$ von $\boldsymbol{X}$ her (Fahrmeir et al., 1996b, S. 96). Da dies äquivalent zu stochastischer Unabhängigkeit zwischen Fehler $\boldsymbol{\epsilon}$ und Regressoren $\boldsymbol{X}$ ist, ist das Gauß-Markov-Theorem auch bei dieser strengeren Annahme gültig.

In der Klasse aller bedingten linearen erwartungstreuen Schätzer ist der KQ-Schätzer weiterhin BLUE, da $\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}$ für jede Realisation von $\boldsymbol{X}$ BLUE ist (Shaffer, 1991, S. 269). Zudem impliziert dies auch eine unbedingte erwartungstreue Schätzung, da $\mathrm{E}\left(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}\right) = \mathrm{E}\left(\mathrm{E}\left(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}|\boldsymbol{X}\right)\right) = \mathrm{E}\left(\boldsymbol{\beta}\right) = \boldsymbol{\beta}$. Bei unbedingter Betrachtung gilt jedoch $\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}\right) = \sigma^2\,\mathrm{E}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$ (Fahrmeir et al., 1996b, S. 99).

Shaffer (1991) analysiert das Gauß-Markov-Theorem hinsichtlich der Gültigkeit im Falle stochastischer Regressoren. Sie leitet Bedingungen her, unter denen der KQ-Schätzer auch in der größeren Klasse aller unbedingten linearen erwartungstreuen Schätzer weiterhin BLUE ist: Falls $(\boldsymbol{X}, \boldsymbol{y})$ der multivariaten Normalverteilung mit unbekannten Parametern oder falls $\boldsymbol{X}$ einer stetigen, nicht-degenerierten, aber nicht weiter spezifizierten unbekannten Verteilung folgt. Zudem falls $\boldsymbol{X}$ Realisationen einer einfachen Zufallsstichprobe ohne Zurücklegen aus einer beschränkten Grundgesamtheit sind, wobei gewisse schwache Bedingungen unterstellt werden. Darüber hinaus wird bewiesen, dass das Theorem jedoch nicht gilt, falls der Erwartungswert $\mathrm{E}\left(\boldsymbol{X}^T\boldsymbol{X}\right)$ bekannt ist.

In einigen Fällen kann der KQ-Schätzer aus Satz 10 somit auch im Falle stochastischer Regressoren BLUE sein. Für theoretische Herleitungen sei auf Shaffer (1991) verwiesen.

Zu beachten ist, dass im linearen Regressionsmodell (7.3) bzw. (7.4) verschobene Zeitpunkte $z_i^{(c)}$ anstelle der Zeitpunkte i verwendet werden. Diese verschobenen Zeitpunkte $z_i^{(c)}$ sind in allen p Dimensionen, also für alle p Regressionsmodelle identisch. Der verschobene Zeitpunkt $z_i^{(c)}$ ergibt sich durch den Mittelwert der Zeitpunkte aller Beobachtungen aus Klasse c, die insgesamt in die Berechnung des aktualisierten Mittelwertvektors $\boldsymbol{m}^{(c)}_{n_i^{(c)}}$ der Klasse c zum Zeitpunkt i eingeflossen sind:

$$z_i^{(c)} = \frac{1}{n_i^{(c)}} \sum_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j \le i}} j = \frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}} \sum_{j=1}^{i} \left(j \cdot \mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}} \right). \tag{7.6}$$

Der Grund dafür ist, dass die Klassenmittel zum Zeitpunkt i keine adäquaten Schätzer für die Erwartungswertvektoren zum Zeitpunkt i sind. Dies liegt daran, dass die Klassenmittel im Datenstrom bei der Online Diskriminanzanalyse mit jeder neuen Beobachtung aktualisiert werden. Das Klassenmittel der Klasse c zum Zeitpunkt i basiert daher auf allen Beobachtungen von Zeitpunkt 1 bis i, die in Klasse c aufgetreten sind. Daher repräsentiert der Schätzer vielmehr den Erwartungswertvektor der Klasse c zum mittleren Zeitpunkt des Auftretens dieser Beobachtungen (vgl. (7.6)) als zum Zeitpunkt i. Diese Idee der verschobenen Zeitpunkte ist in Abbildung 7.2 visualisiert und erklärt.

Im einfachen Fall, dass zu jedem Zeitpunkt $i = 1, \ldots, t$ Beobachtungen einer einzelnen Klasse c auftreten bedeutet dies, dass der Mittelwertvektor $\boldsymbol{m}^{(c)}_{n_t^{(c)}}$ ein guter Schätzer für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)}_{\frac{t+1}{2}}$ zum Zeitpunkt $\frac{t+1}{2}$ anstelle von $\boldsymbol{\mu}_t^{(c)}$ ist.

Auch die Zeitpunkte (7.6) können im Modell schrittweise rekursiv aktualisiert werden:

$$z_i^{(c)} = \begin{cases} z_{i-1}^{(c)}, & \text{falls } g(\boldsymbol{x}_i) \neq c, \\ \dfrac{n_{i-1}^{(c)} z_{i-1}^{(c)} + i}{n_{i-1}^{(c)} + 1}, & \text{falls } g(\boldsymbol{x}_i) = c \in \{1, \ldots, M\}, \\ i, & \text{falls } g(\boldsymbol{x}_i) = c = M + 1. \end{cases} \tag{7.7}$$

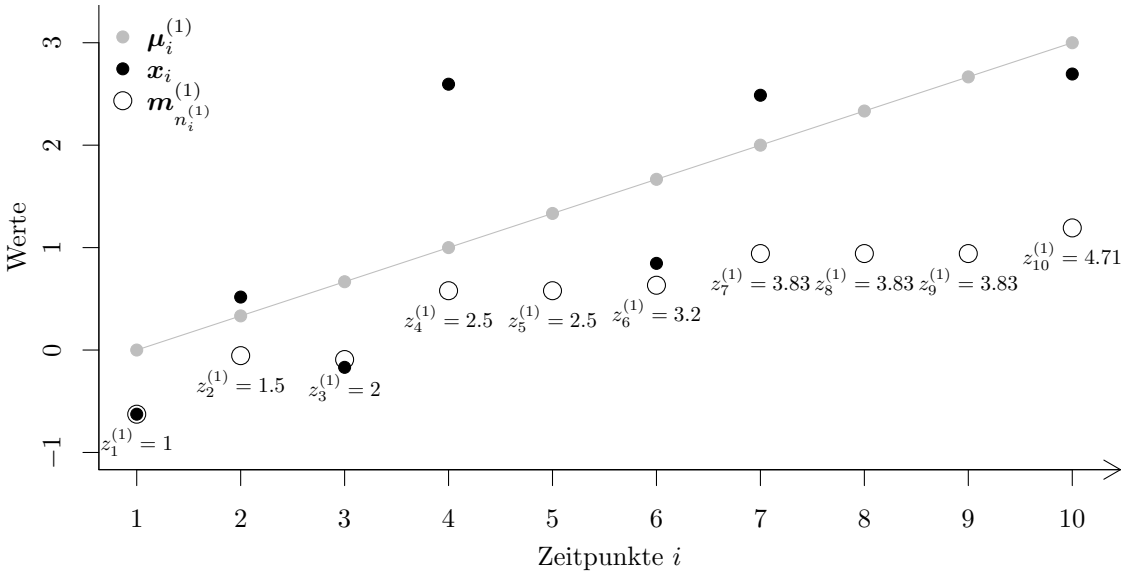


Abbildung 7.2: Beispiel für die verschobenen Zeitpunkte (7.6) bzw. (7.7) für eine Klasse $c = 1$. Die grauen Punkte stellen die wahren Erwartungswerte $\boldsymbol{\mu}_i^{(c)}$ einer eindimensionalen Verteilung für den Zeitpunkt i dar (linearer Trend). Die schwarzen Punkte veranschaulichen Beobachtungen $\boldsymbol{x}_i$ in Klasse 1 zum Zeitpunkt i (hier $i \in \{1, 2, 3, 4, 6, 7, 10\}$) und folgen einer Verteilung mit Erwartungswert $\boldsymbol{\mu}_i^{(c)}$. Die Kreise repräsentieren die aktualisierten Mittelwertschätzer $\boldsymbol{m}_{n_i^{(1)}}^{(1)}$.
Wie zu sehen sind diese adäquate Schätzer für die Erwartungswerte zum Zeitpunkt $z_i^{(1)}$ anstelle von i.

7.4 Vorhersage der Verteilung der Variablen

Zur Verbesserung der Prognose bei der Diskriminanzanalyse wird in einem nächsten Schritt das lokale lineare Regressionsmodell verwendet, um die Verteilung der Variablen des kommenden Zeitpunktes vorherzusagen. Da lediglich ein linearer Trend bezüglich der Erwartungswertvektoren der Variablen angenommen wird, wird daher für jede Klasse der Erwartungswertvektor der Variablen zum kommenden Zeitpunkt $t+1$ mithilfe des folgenden Schätzers prognostiziert:

$$\hat{\boldsymbol{m}}_{n_{t+1}^{(c)}}^{(c)} = \hat{\boldsymbol{\beta}}_{0t}^{(c)} + \hat{\boldsymbol{\beta}}_{1t}^{(c)}(t+1), \quad c = 1, \ldots, M. \tag{7.8}$$

Die Zeitpunkte $i \in I$ bilden auf die verschobenen Zeitpunkte $z_i^{(c)}$ im linearen Regressionsmodell zum Zeitpunkt t ab bzw. die verschobenen Zeitpunkte sind eine Transformation der wahren Zeitpunkte i nach (7.6). $z_i^{(c)}$ kann demnach auch als Funktion von i aufgefasst werden durch

$$z^{(c)} : \mathbb{N} \to \mathbb{R},\ i \mapsto z_i^{(c)} := z^{(c)}(i) \ \text{ mit } \ z^{(c)}(i) = \begin{cases} \text{z. B. (7.6) (abh. von Methode)}, & i \leq t, \\ i & i > t. \end{cases}$$

Bei der Prognose durch das lineare Regressionsmodell (7.3) bzw. (7.4) gilt daher $z_{t+1}^{(c)} = t+1$. Die Parameter des unterstellten linearen Trends der Erwartungswertvektoren (7.1) und jene aus dem linearen Regressionsmodell (7.3) sind somit identisch, da die Mittelwertvektoren $\boldsymbol{m}_{n_i^{(c)}}^{(c)}$ beobachtbare Realisationen der Erwartungswertvektoren von Klasse c zu den Zeitpunkten $z_i^{(c)}$ sind (formal $\boldsymbol{\mu}_{z_i^{(c)}}^{(c)}$): $\boldsymbol{\beta}_{0t}^{(c)} := \boldsymbol{\beta}_0^{(c)}$, $\boldsymbol{\beta}_{1t}^{(c)} := \boldsymbol{\beta}_1^{(c)}$.

Aufgrund des stochastischen Regressors im linearen Modell werden wie oben in Annahme (3*) beschrieben alle Modellannahmen bedingt auf die Designmatrix $\boldsymbol{Z}_j^{(c)}$ betrachtet (Fahrmeir et al., 2009, S. 62). Es wird angenommen, dass der bedingte Erwartungswert der Fehler des einfachen linearen Regressionsmodells (7.4) bzw. (7.5) $\mathrm{E}\big(\epsilon_{ij}^{(c)} | z_i^{(c)}\big) = 0$ für alle $i \in I$ ist, also $\mathrm{E}\big(\boldsymbol{\epsilon}_j^{(c)} | \boldsymbol{Z}_j^{(c)}\big) = \boldsymbol{0}$.

Da die Mittelwertschätzer $\boldsymbol{m}_{n_i^{(c)}}^{(c)}$ jedoch rekursiv bestimmt werden und somit auf immer mehr, jedoch teilweise denselben Beobachtungen basieren, sind diese nicht unabhängig voneinander. Es liegt eine Korrelationsstruktur vor, welche durch den festen Parametervektor $\boldsymbol{\beta}_j^{(c)}$ im linearen Regressionsmodell nicht abgebildet wird und somit in den Fehlerterm $\boldsymbol{\epsilon}_j^{(c)}$ von (7.5) bzw. die Elemente $\epsilon_{ij}^{(c)}$ übergeht. Daher sind die (bedingten) Fehler autokorreliert:

$$\mathrm{Cov}\left(\epsilon_{ij}^{(c)}, \epsilon_{kj}^{(c)} | z_i^{(c)}, z_k^{(c)}\right) \neq 0 \quad \text{für alle } i \neq k,\ i,k \in I.$$

Die Bedingung (3*) (Fahrmeir et al., 1996b, S. 96) $\mathrm{Cov}\big(\boldsymbol{\epsilon}_j^{(c)} | \boldsymbol{Z}_j^{(c)}\big) = \sigma^2 \boldsymbol{I}$ der Gauß-Markov-Annahmen ist somit verletzt. Es gilt stattdessen $\mathrm{Cov}\big(\boldsymbol{\epsilon}_j^{(c)} | \boldsymbol{Z}_j^{(c)}\big) = \sigma^2 \boldsymbol{\Psi}$ mit $\boldsymbol{\Psi} \neq \boldsymbol{I}$ (Groß, 2003, S. 34/266). Der KQ-Schätzer ist nicht mehr bester linearer erwartungstreuer Schätzer (BLUE) für den wahren Parametervektor $\boldsymbol{\beta}_j^{(c)} = \left(\left[\boldsymbol{\beta}_{0t}^{(c)}\right]_j, \left[\boldsymbol{\beta}_{1t}^{(c)}\right]_j\right)^T$ (Groß, 2003, S. 51). Auch nicht in der Klasse der bedingten linearen erwartungstreuen Schätzer bei Vorliegen stochastischer Regressoren.

Einen Ausweg bietet die Verwendung des verallgemeinerten KQ-Schätzers (Toutenburg, 2003, S. 303/312; Groß, 2003, S. 34/266 ff.):

Satz 11. *Gauß-Markov-Aitken-Theorem*
Im verallgemeinerten linearen Regressionsmodell $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ ist der verallgemeinerte KQ-Schätzer (auch als Aitken-Schätzer bezeichnet)

$$\tilde{\boldsymbol{\beta}}_{\mathrm{KQ}} = \left(\boldsymbol{X}^T \boldsymbol{\Psi}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{\Psi}^{-1} \boldsymbol{y}$$

unter den Annahmen

(1) $\boldsymbol{X}$ ist nicht-stochastisch mit vollem Spaltenrang,
(2) die Elemente des Vektors $\boldsymbol{y}$ sind (beobachtbare) Zufallsvariablen,
(3) $\boldsymbol{\epsilon} \sim \left(\boldsymbol{0}, \sigma^2 \boldsymbol{\Psi}\right)$, also $\mathrm{E}(\boldsymbol{\epsilon}) = \boldsymbol{0}$ und $\mathrm{Cov}(\boldsymbol{\epsilon}) = \mathrm{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right) = \sigma^2 \boldsymbol{\Psi}$ mit $\sigma^2 > 0$ und $\boldsymbol{\Psi}$ positiv definit und bekannt,

die beste lineare erwartungstreue Schätzung des wahren Parametervektors $\boldsymbol{\beta}$.

Dies ist äquivalent zur Transformation des linearen Regressionsmodells, um die Struktur in den Fehlertermen zu korrigieren (Groß, 2003, S. 267), wobei im Folgenden aufgrund der stochastischen Regressoren wieder die durch $\boldsymbol{Z}_j^{(c)}$ bedingte Verteilung der Fehler betrachtet wird. Es muss eine Transformationsmatrix $\boldsymbol{P}$ gefunden werden, für die $\boldsymbol{\Psi}^{-1} = \boldsymbol{P}^T\boldsymbol{P}$ gilt. Durch Multiplikation beider Seiten des linearen Regressionsmodells (7.5) mit $\boldsymbol{P}$ kann das Modell so transformiert werden, dass der Fehlerterm die Gauß-Markov-Annahme $\tilde{\boldsymbol{\epsilon}}_j^{(c)}|\boldsymbol{Z}_j^{(c)} := \boldsymbol{P}\boldsymbol{\epsilon}_j^{(c)}|\boldsymbol{Z}_j^{(c)} \sim (\boldsymbol{0}, \sigma^2\boldsymbol{I})$ aus Satz 10 erfüllt:

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{P}\boldsymbol{\epsilon}_j^{(c)}|\boldsymbol{Z}_j^{(c)}\right) &= \boldsymbol{P}\,\mathrm{E}\left(\boldsymbol{\epsilon}_j^{(c)}|\boldsymbol{Z}_j^{(c)}\right) = \boldsymbol{0}, \\
\mathrm{Cov}\left(\boldsymbol{P}\boldsymbol{\epsilon}_j^{(c)}|\boldsymbol{Z}_j^{(c)}\right) &= \boldsymbol{P}\,\mathrm{Cov}\left(\boldsymbol{\epsilon}_j^{(c)}|\boldsymbol{Z}_j^{(c)}\right)\boldsymbol{P}^T = \sigma^2\boldsymbol{P}\boldsymbol{\Psi}\boldsymbol{P}^T = \sigma^2\boldsymbol{P}\boldsymbol{P}^{-1}\left(\boldsymbol{P}^T\right)^{-1}\boldsymbol{P}^T = \sigma^2\boldsymbol{I}.
\end{aligned}$$

Der resultierende KQ-Schätzer des transformierten Modells $\boldsymbol{P}\boldsymbol{m}_j^{(c)} = \boldsymbol{P}\boldsymbol{Z}_j^{(c)}\boldsymbol{\beta}_j^{(c)} + \boldsymbol{P}\boldsymbol{\epsilon}_j^{(c)}$ bzw. der verallgemeinerte KQ-Schätzer $\tilde{\boldsymbol{\beta}}_j^{(c)}$ des ursprünglichen Modells (7.5)

$$\tilde{\boldsymbol{\beta}}_j^{(c)} = \left(\left(\boldsymbol{P}\boldsymbol{Z}_j^{(c)}\right)^T\boldsymbol{P}\boldsymbol{Z}_j^{(c)}\right)^{-1}\left(\boldsymbol{P}\boldsymbol{Z}_j^{(c)}\right)^T\boldsymbol{P}\boldsymbol{m}_j^{(c)} = \left(\left(\boldsymbol{Z}_j^{(c)}\right)^T\boldsymbol{\Psi}^{-1}\boldsymbol{Z}_j^{(c)}\right)^{-1}\left(\boldsymbol{Z}_j^{(c)}\right)^T\boldsymbol{\Psi}^{-1}\boldsymbol{m}_j^{(c)}$$

ist unter der Annahme $\boldsymbol{\epsilon}_j^{(c)}\big|\boldsymbol{Z}_j^{(c)} \sim \left(\boldsymbol{0}, \sigma^2\boldsymbol{\Psi}\right)$ nach Satz 11 erwartungstreu für $\boldsymbol{\beta}_j^{(c)}$, da $\mathrm{E}\big(\tilde{\boldsymbol{\beta}}_j^{(c)}\big) = \mathrm{E}\big(\mathrm{E}\big(\tilde{\boldsymbol{\beta}}_j^{(c)}\big|\boldsymbol{Z}_j^{(c)}\big)\big) = \mathrm{E}\big(\boldsymbol{\beta}_j^{(c)}\big) = \boldsymbol{\beta}_j^{(c)}$. Mit den Ergebnissen von Shaffer (1991) lässt sich schlussfolgern, dass der verallgemeinerte KQ-Schätzer $\tilde{\boldsymbol{\beta}}_j^{(c)}$ BLUE in der Klasse der bedingten linearen erwartungstreuen Schätzer ist. In der Klasse der unbedingten linearen erwartungstreuen Schätzer gilt die BLUE Eigenschaft ohne weitere Annahmen nicht im Allgemeinen (vgl. Seite 164). Für die Kovarianzmatrix gilt: $\mathrm{Cov}\big(\tilde{\boldsymbol{\beta}}_j^{(c)}\big) = \sigma^2\mathrm{E}\big(\big(\big(\boldsymbol{Z}_j^{(c)}\big)^T\boldsymbol{\Psi}^{-1}\boldsymbol{Z}_j^{(c)}\big)^{-1}\big) = \sigma^2\mathrm{E}\big(\big(\big(\boldsymbol{P}\boldsymbol{Z}_j^{(c)}\big)^T\boldsymbol{P}\boldsymbol{Z}_j^{(c)}\big)^{-1}\big)$ (Groß, 2003, S. 267 f.; Fahrmeir et al., 1996b, S. 99, 2009, S. 127).

Tabelle 7.1 fasst das *Gauß-Markov-Theorem*, das *Gauß-Markov-Aitken-Theorem* und die Ergebnisse von Shaffer (1991) zusammen. Es wird gegenübergestellt, welche Eigenschaften für den einfachen KQ-Schätzer $\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}$ und den verallgemeinerten KQ-Schätzer $\tilde{\boldsymbol{\beta}}_{\mathrm{KQ}}$ bei Vorliegen bzw. Nicht-Vorliegen von autokorrelierten Fehlern und stochastischen Regressoren unter verschiedenen Voraussetzungen gelten.

In der Praxis ist die Matrix $\boldsymbol{\Psi}$ zur Beschreibung der Kovarianzstruktur des (bedingten) Fehlers $\boldsymbol{\epsilon}_j^{(c)}$ allerdings unbekannt. Eine Schätzung von $\boldsymbol{\Psi}$ und Herleitung der Transformationsmatrix $\boldsymbol{P}$ ist ohne theoretische Annahmen aufwändig. Fahrmeir et al. (1996b, S. 102) erläutern ein zweistufiges Verfahren, bei welchem zunächst der normale KQ-Schätzer bestimmt wird. Mithilfe von Erkenntnissen aus den Residualplots können unbekannte Korrelationsparameter geschätzt werden, um daraufhin die verallgemeinerte KQ-Schätzung durchzuführen. Groß (2003, S. 270) weist darauf hin, dass der verallgemeinerte KQ-Schätzer bei Verwendung einer Schätzung von $\boldsymbol{\Psi}$ nicht mehr dieselben statistischen Eigenschaften aufweist. Mit Verweis auf Greene (2002) wird betont, dass diese unter bestimmten Annahmen nur asymptotisch gelten.

Tabelle 7.1: Gegenüberstellung von Eigenschaften des KQ-Schätzers $\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}$ und des verallgemeinerten KQ-Schätzers $\tilde{\boldsymbol{\beta}}_{\mathrm{KQ}}$ unter bestimmten Voraussetzungen.

	nicht-stochastische Regressoren (1)	stochastische Regressoren (1*)
keine Autokorrelation	(3) $\boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2 \boldsymbol{I})$ aus Satz 10 ⇒ Gewöhnlicher KQ-Schätzer $\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}$ ist BLUE nach *Gauß-Markov-Theorem*: $\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}) = \boldsymbol{\beta}$, $\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}$.	(3*) $\boldsymbol{\epsilon} \vert \boldsymbol{X} \sim (\mathbf{0}, \sigma^2 \boldsymbol{I})$ ⇒ $\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}$ ist BLUE in Klasse aller bedingten linearen unverzerrten Schätzer. ⇒ (Unbedingte) unverzerrte Schätzung $\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}) = \boldsymbol{\beta}$. ⇒ $\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}$ BLUE in Klasse aller unbedingten linearen unverzerrten Schätzer nur unter weiteren Bedingungen.
Autokorrelation	(3) $\boldsymbol{\epsilon} \sim (\mathbf{0}, \sigma^2 \boldsymbol{\Psi})$ aus Satz 11 ⇒ Verallgemeinerter KQ-Schätzer $\tilde{\boldsymbol{\beta}}_{\mathrm{KQ}}$ ist BLUE nach *Gauß-Markov-Aitken-Theorem*: $\mathrm{E}(\tilde{\boldsymbol{\beta}}_{\mathrm{KQ}}) = \boldsymbol{\beta}$, $\mathrm{Cov}(\tilde{\boldsymbol{\beta}}_{\mathrm{KQ}}) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{\Psi}^{-1} \boldsymbol{X}\right)^{-1}$. ⇒ $\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}$ weiterhin unverzerrt: $\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}) = \boldsymbol{\beta}$.	(3*) $\boldsymbol{\epsilon} \vert \boldsymbol{X} \sim (\mathbf{0}, \sigma^2 \boldsymbol{\Psi})$ ⇒ $\tilde{\boldsymbol{\beta}}_{\mathrm{KQ}}$ ist BLUE in Klasse aller bedingten linearen unverzerrten Schätzer. ⇒ (Unbedingte) unverzerrte Schätzung $\mathrm{E}(\tilde{\boldsymbol{\beta}}_{\mathrm{KQ}}) = \boldsymbol{\beta}$. ⇒ $\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}$ weiterhin (unbedingt) unverzerrt: $\mathrm{E}(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}) = \boldsymbol{\beta}$.

Der einfache KQ-Schätzer des Modells (7.5)

$$\hat{\boldsymbol{\beta}}_j^{(c)} = \left(\left(\boldsymbol{Z}_j^{(c)}\right)^T \boldsymbol{Z}_j^{(c)}\right)^{-1} \left(\boldsymbol{Z}_j^{(c)}\right)^T \boldsymbol{m}_j^{(c)} = \left(\left(\boldsymbol{Z}_j^{(c)}\right)^T \boldsymbol{Z}_j^{(c)}\right)^{-1} \left(\boldsymbol{Z}_j^{(c)}\right)^T \left(\boldsymbol{Z}_j^{(c)} \boldsymbol{\beta}_j^{(c)} + \boldsymbol{\epsilon}_j^{(c)}\right) \tag{7.9}$$

ist auch bei Vorliegen von autokorrelierten Fehlern, also den Bedingungen aus Satz 11, weiterhin erwartungstreu für den wahren Parametervektor $\boldsymbol{\beta}_j^{(c)}$ (Groß, 2003, S. 268; Fahrmeir et al., 1996b, S. 99):

$$\mathrm{E}\left(\hat{\boldsymbol{\beta}}_j^{(c)}\right) = \mathrm{E}\left(\mathrm{E}\left(\hat{\boldsymbol{\beta}}_j^{(c)} \middle| \boldsymbol{Z}_j^{(c)}\right)\right) = \mathrm{E}\left(\boldsymbol{\beta}_j^{(c)}\right) = \boldsymbol{\beta}_j^{(c)}.$$

Im Falle nicht-stochastischer Regressoren $\boldsymbol{X}$ ist die Varianz des einfachen KQ-Schätzers aus Satz 10 unter den Bedingungen aus Satz 11 jedoch nicht mehr minimal, sondern (Groß, 2003, S. 268)

$$\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}_{\mathrm{KQ}}\right) = \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{\Psi} \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}.$$

Im Falle stochastischer Regressoren wie $\boldsymbol{Z}_j^{(c)}$ gilt die BLUE Eigenschaft folglich bereits in der Klasse aller bedingten linearen erwartungstreuen Schätzer nicht.

Da die Bestimmung des verallgemeinerten KQ-Schätzers in der Praxis ohne Kenntnisse über die Kovarianzstruktur $\boldsymbol{\Psi}$ aufwändig ist und selbst der verallgemeinerte KQ-Schätzer

bei Vorliegen stochastischer Regressoren nicht in jedem Fall BLUE in der Klasse aller unbedingten linearen erwartungstreuen Schätzer ist, wird der einfache KQ-Schätzer (7.9) aus praktischen Gründen in einem ersten Schritt zur Schätzung herangezogen.

Es können keine Hypothesentests durchgeführt werden, der wahre Erwartungswertvektor wird im Mittel jedoch richtig geschätzt. Anhand der Simulationsstudie in Kapitel 9 wird deutlich, dass ein linearer Trend der Erwartungswertvektoren durch das einfache lineare Regressionsmodell auf den aktualisierten Mittelwertvektoren angemessen modelliert werden kann. Die Eigenschaft der Erwartungstreue ist ausreichend, um die Prognosegüte der Update-Methoden der Diskriminanzanalyse durch Integration lokaler linearer Regressionsmodelle verbessern zu können.

Es wird somit der KQ-Schätzer (7.9) betrachtet. Dieser minimiert (Groß, 2003, S. 38)

$$\begin{aligned}\hat{\boldsymbol{\beta}}_j^{(c)} = \left(\left[\hat{\boldsymbol{\beta}}_{0t}^{(c)}\right]_j, \left[\hat{\boldsymbol{\beta}}_{1t}^{(c)}\right]_j\right)^T &= \underset{\boldsymbol{\beta}^*=(\beta_0^*,\beta_1^*)^T}{\arg\min} \left(\boldsymbol{m}_j^{(c)} - \boldsymbol{Z}_j^{(c)}\boldsymbol{\beta}^*\right)^T \left(\boldsymbol{m}_j^{(c)} - \boldsymbol{Z}_j^{(c)}\boldsymbol{\beta}^*\right) \\ &= \underset{\boldsymbol{\beta}^*=(\beta_0^*,\beta_1^*)^T}{\arg\min} \sum_{i\in I} \left(\left[\boldsymbol{m}_{n_i^{(c)}}^{(c)}\right]_j - \beta_0^* - \beta_1^* z_i^{(c)}\right)^2.\end{aligned}$$

Ableiten der Summe nach β_0^* und β_1^* und Gleichsetzen mit Null liefert die getrennten standardmäßigen geschlossenen Formen für $\left[\hat{\boldsymbol{\beta}}_{0t}^{(c)}\right]_j$ und $\left[\hat{\boldsymbol{\beta}}_{1t}^{(c)}\right]_j$ für das einfache lineare Regressionsmodell (7.4) bzw. (7.5). Bei gleichzeitiger Betrachtung aller p Dimensionen ergeben sich die folgenden Schätzer für das gesamte Regressionsmodell (7.3):

$$\hat{\boldsymbol{\beta}}_{1t}^{(c)} = \frac{\sum_{i\in I}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\boldsymbol{m}_{n_i^{(c)}}^{(c)} - \bar{\boldsymbol{m}}_t^{(c)}\right)}{\sum_{i\in I}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2}, \tag{7.10}$$

$$\hat{\boldsymbol{\beta}}_{0t}^{(c)} = \bar{\boldsymbol{m}}_t^{(c)} - \hat{\boldsymbol{\beta}}_{1t}^{(c)}\bar{z}_t^{(c)}, \tag{7.11}$$

wobei die herangezogenen Mittelwerte folgendermaßen berechnet werden:

$$\begin{aligned}\bar{z}_t^{(c)} &= \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i\in I} z_i^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i\in I}\left(\frac{1}{n_i^{(c)}}\sum_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j\le i}} j\right) \\ &= \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i\in I}\left(\frac{1}{\sum_{k=1}^{i}\mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}\sum_{j=1}^{i}\left(j\cdot\mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}}\right)\right), \\ \bar{\boldsymbol{m}}_t^{(c)} &= \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i\in I}\boldsymbol{m}_{n_i^{(c)}}^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i\in I}\left(\frac{1}{n_i^{(c)}}\sum_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j\le i}} \boldsymbol{x}_j\right) \\ &= \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i\in I}\left(\frac{1}{\sum_{k=1}^{i}\mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}\sum_{j=1}^{i}\left(\boldsymbol{x}_j\cdot\mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}}\right)\right).\end{aligned}$$

Insgesamt ergibt sich die Vorhersage $\hat{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t+1}}$ aus (7.8) dann durch

$$\hat{\boldsymbol{\beta}}^{(c)}_{0t}+\hat{\boldsymbol{\beta}}^{(c)}_{1t}(t+1)=\bar{\boldsymbol{m}}^{(c)}_t-\left(\frac{\sum_{i\in I}\left(z^{(c)}_i-\bar{z}^{(c)}_t\right)\left(\boldsymbol{m}^{(c)}_{n^{(c)}_i}-\bar{\boldsymbol{m}}^{(c)}_t\right)}{\sum_{i\in I}\left(z^{(c)}_i-\bar{z}^{(c)}_t\right)^2}\right)\left(\bar{z}^{(c)}_t-(t+1)\right). \quad (7.12)$$

7.5 Einbindung in existierende Methoden

Die Idee ist für alle existierenden Methoden für Online Diskriminanzanalyse dieselbe. Der zeitabhängige concept drift wird durch ein lokales lineares Regressionsmodell wie in Abschnitt 7.3 modelliert. Danach folgt für jede Klasse eine Prognose des Erwartungswertvektors der Variablen zum kommenden Zeitpunkt mithilfe dieses Regressionsmodells durch (7.8) bzw. (7.12). Diese Prognosen ersetzen die bisherigen Schätzer für die Erwartungswertvektoren in der Klassifikationsregel der Diskriminanzanalyse.

Die allgemeine Form des Regressionsmodells für Klasse c ist für alle Methoden durch (7.3) gegeben:

$$\boldsymbol{m}^{(c)}_{n^{(c)}_i}=\boldsymbol{\beta}^{(c)}_{0t}+\boldsymbol{\beta}^{(c)}_{1t}z^{(c)}_i+\boldsymbol{\epsilon}^{(c)}_i, \quad i\in I.$$

Hier werden jeweils die Mittelwertvektoren $\boldsymbol{m}^{(c)}_{n^{(c)}_i}$, welche durch die jeweilige Methode der Online Diskriminanzanalyse bestimmt werden, betrachtet (vgl. (4.8), (4.27) und (4.38)).

Allerdings sind ebenfalls die verschobenen Zeitpunkte nicht für alle Methoden identisch, da diese auch auf den jeweils betrachteten Gewichten bzw. den Ideen zur Anpassung an einen concept drift basieren. Der Spezialfall, dass bei der Klassifikationsmethode keine Anpassung an einen concept drift erfolgt (vgl. *Sequential Incremental LDA* (Abschnitt 4.2) und *OLDC (Online Linear Discriminant Classifier)* mit fester Lernrate $\lambda = 1/2$ (Abschnitt 4.3)) wurde bereits in Abschnitt 7.3 betrachtet (vgl. (7.7)):

$$z^{(c)}_i=\begin{cases} z^{(c)}_{i-1}, & \text{falls } g(\boldsymbol{x}_i)\neq c,\\ \dfrac{n^{(c)}_{i-1}z^{(c)}_{i-1}+i}{n^{(c)}_{i-1}+1}, & \text{falls } g(\boldsymbol{x}_i)=c\in\{1,\dots,M\},\\ i, & \text{falls } g(\boldsymbol{x}_i)=c=M+1. \end{cases}$$

Die Formeln für die verschobenen Zeitpunkte der übrigen in dieser Arbeit betrachteten Methoden werden im Folgenden hergeleitet.

OLDC (Online Linear Discriminant Classifier) Für die Bestimmung der verschobenen Zeitpunkte für *OLDC (Online Linear Discriminant Classifier)* (Abschnitt 4.3) im allgemeinen Fall, also mit Lernrate λ, wird die rekursive Formel zur Bestimmung der Mittelwert-

vektoren betrachtet. In Analogie zu dieser Update-Formel der Klassenmittelwertvektoren (4.27) lassen sich die verschobenen Zeitpunkte demnach folgendermaßen ermitteln:

$$z_i^{(c)} = \begin{cases} z_{i-1}^{(c)}, & \text{falls } g(\boldsymbol{x}_i) \neq c, \\ \dfrac{(1-\lambda) n_{i-1}^{(c)} z_{i-1}^{(c)} + \lambda i}{(1-\lambda) n_{i-1}^{(c)} + \lambda}, & \text{falls } g(\boldsymbol{x}_i) = c \in \{1, \ldots, M\}, \\ i, & \text{falls } g(\boldsymbol{x}_i) = c = M+1. \end{cases}$$

Die jeweiligen Zeitpunkte werden genauso gewichtet wie die entsprechenden Beobachtungen bzw. vergangenen Mittelwertvektoren bei der Mittelwertbestimmung. Der neue Zeitpunkt erhält das Gewicht λ, der Anteil der vergangenen Zeitpunkte, bzw. die (gewichtete) „Summe aller alten Zeitpunkte“ $n_{i-1}^{(c)} z_{i-1}^{(c)}$, fließt mit dem Gewicht $1-\lambda$ in die Berechnung ein. Ebenso wird der Vorfaktor zur Normierung entsprechend gewichtet.

Die Formel lässt sich auch nicht-rekursiv in Abhängigkeit aller Zeitpunkte herleiten, wobei die iterative Formulierung der Formel aufgrund der Gewichtung recht komplex ist (vgl. (6.8) für iterative Variante zur Bestimmung der Mittelwertvektoren):

$$\begin{aligned} z_i^{(c)} = {} & \frac{1}{(1-\lambda)(n_i^{(c)} - 1) + \lambda} \cdot \\ & \Bigg(\frac{(1-\lambda)^{n_i^{(c)}-1} (n_i^{(c)} - 1)!}{\prod\limits_{\substack{k:\, g(\boldsymbol{x}_k)=c \,\cap\, n_k^{(c)}>2 \\ 3 \leq k \leq i}} ((1-\lambda)(n_k^{(c)} - 2) + \lambda)} \cdot \underset{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j<i}}{\operatorname{argmin}} (\boldsymbol{x}_j) \\ & + \sum_{k=\underset{j:\, g(\boldsymbol{x}_j)=c;\, j<i}{\operatorname{argmin}}(\boldsymbol{x}_j)+1}^{\underset{j:\, g(\boldsymbol{x}_j)=c;\, j \leq i}{\operatorname{argmax}}(\boldsymbol{x}_j)-1} \frac{\frac{(n_i^{(c)}-1)!}{(n_k^{(c)}-1)!} \cdot (1-\lambda)^{n_i^{(c)} - n_k^{(c)}} \lambda}{\prod\limits_{\substack{l:\, g(\boldsymbol{x}_l)=c \,\cap\, n_l^{(c)}>2 \\ k+1 \leq l \leq i}} ((1-\lambda)(n_l^{(c)} - 2) + \lambda)} \cdot k \cdot \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}} \\ & + \lambda \underset{j:\, g(\boldsymbol{x}_j)=c;\, j \leq i}{\operatorname{argmax}} (\boldsymbol{x}_j) \Bigg) \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i) \in \{1,\ldots,M\}\}} + i \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=M+1\}} + z_{i-1}^{(c)} \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i) \neq c\}}. \end{aligned}$$

Für den Spezialfall $g(\boldsymbol{x}_j) = c$, $1 \leq j \leq i$, gilt vereinfacht (vgl. (6.9)):

$$\begin{aligned} z_i^{(c)} = {} & \frac{1}{(1-\lambda)(i-1) + \lambda} \cdot \\ & \left(\frac{(1-\lambda)^{i-1} \prod\limits_{j=1}^{i-1} j}{\prod\limits_{j=3}^{i} ((1-\lambda)(j-2) + \lambda)} \cdot 1 + \sum_{j=2}^{i-1} \frac{(1-\lambda)^{i-j} \lambda \prod\limits_{k=j}^{i-1} k}{\prod\limits_{k=j+1}^{i} ((1-\lambda)(k-2) + \lambda)} \cdot j + \lambda\, i \right). \end{aligned} \tag{7.13}$$

Online Diskriminanzanalyse mit adaptivem Vergessen Die verschobenen Zeitpunkte werden hier hergeleitet, indem die Zeitpunkte $1 \leq j \leq i$ mit den entsprechenden analogen Gewichten gewichtet werden, welche ebenfalls bei der Berechnung der Mittelwertvektoren der Klassen zum Zeitpunkt i einfließen. Es wird also eine gewichtete Summe der Zeitpunkte betrachtet. Mit Hinblick auf (4.34) ergibt sich:

$$z_i^{(c)} = \frac{1}{N_i^{(c)}} \sum_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j \leq i}} v_j^{(c)} j = \frac{1}{N_i^{(c)}} \sum_{j=1}^{i} \left(v_j^{(c)} j \cdot \mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}} \right) \tag{7.14}$$

$$\overset{(4.37)}{=} \frac{1}{\sum\limits_{\substack{l:\, g(\boldsymbol{x}_l)=c \\ l \leq i-1}} v_l^{(c)} + 1} \left(\sum_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j \leq i-1}} v_j^{(c)} j + i \right)$$

$$\overset{(4.33)}{=} \frac{1}{\sum\limits_{\substack{l:\, g(\boldsymbol{x}_l)=c \\ l \leq i-1}} \left(\prod\limits_{k=\sum_{m=1}^{l} \mathbb{1}_{\{g(\boldsymbol{x}_m)=c\}}}^{n_i^{(c)}-1} \lambda_{(k)}^{(c)} \right) + 1} \left(\sum_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j \leq i-1}} \left(\prod_{k=\sum_{m=1}^{j} \mathbb{1}_{\{g(\boldsymbol{x}_m)=c\}}}^{n_i^{(c)}-1} \lambda_{(k)}^{(c)} \right) j + i \right).$$

Zur Herleitung der rekursiven Formel wird zunächst der Spezialfall betrachtet, dass bis zum Zeitpunkt i nur Beobachtungen aus Klasse c aufgetreten sind, also $g(\boldsymbol{x}_j) = c,\ 1 \leq j \leq i$.

Für die ersten Zeitpunkte gilt mit (7.14) und $N_i^{(c)}$ aus (4.37) in Abschnitt 4.4:

- $i = 1$:

$$z_1^{(c)} = \frac{1}{N_1^{(c)}} \cdot v_1^{(c)} \cdot 1 = \frac{1}{1} \cdot 1 \cdot 1 = 1 = \left(1 - \frac{1}{N_1^{(c)}} \right) z_0^{(c)} + \frac{1}{N_1^{(c)}} \cdot 1 \text{ mit } z_0^{(c)} := 0.$$

- $i = 2$:

$$z_2^{(c)} = \frac{1}{N_2^{(c)}} \left(v_1^{(c)} \cdot 1 + v_2^{(c)} \cdot 2 \right) = \frac{1}{\lambda_{(1)}^{(c)} + 1} \cdot \left(\lambda_{(1)}^{(c)} \cdot 1 + 1 \cdot 2 \right) = \frac{\lambda_{(1)}^{(c)} + 2}{\lambda_{(1)}^{(c)} + 1}$$
$$= \left(1 - \frac{1}{\lambda_{(1)}^{(c)} + 1} \right) \cdot 1 + \frac{1}{\lambda_{(1)}^{(c)} + 1} \cdot 2 = \left(1 - \frac{1}{N_2^{(c)}} \right) z_1^{(c)} + \frac{1}{N_2^{(c)}} \cdot 2.$$

- $i = 3$:

$$z_3^{(c)} = \frac{1}{N_3^{(c)}} \left(v_1^{(c)} \cdot 1 + v_2^{(c)} \cdot 2 + v_3^{(c)} \cdot 3 \right) = \frac{1}{\lambda_{(2)}^{(c)} N_2^{(c)} + 1} \left(\lambda_{(1)}^{(c)} \lambda_{(2)}^{(c)} \cdot 1 + \lambda_{(2)}^{(c)} \cdot 2 + 1 \cdot 3 \right)$$
$$= \frac{\lambda_{(1)}^{(c)} \lambda_{(2)}^{(c)} + \lambda_{(2)}^{(c)} \cdot 2 + 3}{\lambda_{(1)}^{(c)} \lambda_{(2)}^{(c)} + \lambda_{(2)}^{(c)} + 1} = \frac{\left(\lambda_{(1)}^{(c)} + 1 \right) \left(\lambda_{(1)}^{(c)} + 2 \right) \cdot \lambda_{(2)}^{(c)}}{\left(\lambda_{(1)}^{(c)} + 1 \right) \left(\lambda_{(1)}^{(c)} \lambda_{(2)}^{(c)} + \lambda_{(2)}^{(c)} + 1 \right)} + \frac{3}{\lambda_{(1)}^{(c)} \lambda_{(2)}^{(c)} + \lambda_{(2)}^{(c)} + 1}$$
$$= \frac{\lambda_{(1)}^{(c)} \lambda_{(2)}^{(c)} + \lambda_{(2)}^{(c)} + 1 - 1}{\lambda_{(1)}^{(c)} \lambda_{(2)}^{(c)} + \lambda_{(2)}^{(c)} + 1} \cdot \frac{\lambda_{(1)}^{(c)} + 2}{\lambda_{(1)}^{(c)} + 1} + \frac{1}{\lambda_{(1)}^{(c)} \lambda_{(2)}^{(c)} + \lambda_{(2)}^{(c)} + 1} \cdot 3$$

$$= \left(1 - \frac{1}{\lambda_{(1)}^{(c)}\lambda_{(2)}^{(c)} + \lambda_{(2)}^{(c)} + 1}\right) z_2^{(c)} + \frac{1}{\lambda_{(1)}^{(c)}\lambda_{(2)}^{(c)} + \lambda_{(2)}^{(c)} + 1} \cdot 3$$
$$= \left(1 - \frac{1}{N_3^{(c)}}\right) z_2^{(c)} + \frac{1}{N_3^{(c)}} \cdot 3.$$

- ...

Insgesamt lässt sich die Formel zur Berechnung der verschobenen Zeitpunkte damit auch rekursiv formulieren durch:

$$z_i^{(c)} = \left(1 - \frac{1}{N_i^{(c)}}\right) z_{i-1}^{(c)} + \frac{1}{N_i^{(c)}} \cdot i \quad \text{mit } N_i^{(c)} \text{ aus (4.41).}$$

Die Formel ist damit analog zu jener zur rekursiven Aktualisierung der Klassenmittelwertvektoren (vgl. (4.38)).

Zu beachten ist, dass diese Formel zwar unter dem Spezialfall hergeleitet wurde, dass nur Beobachtungen aus einer Klasse c auftreten, die Formel aber unter Betrachtung der folgenden Fallunterscheidung allgemeingültig ist, da im Falle des Nichtauftretens der Klasse c zu einem Zeitpunkt der entsprechende verschobene Zeitpunkt fortgeschrieben wird:

$$z_i^{(c)} = \begin{cases} z_{i-1}^{(c)}, & \text{falls } g(\boldsymbol{x}_i) \neq c, \\ \left(1 - \frac{1}{N_i^{(c)}}\right) z_{i-1}^{(c)} + \frac{1}{N_i^{(c)}} \cdot i, & \text{falls } g(\boldsymbol{x}_i) = c \in \{1, \ldots, M\}, \\ i, & \text{falls } g(\boldsymbol{x}_i) = c = M + 1. \end{cases}$$

Für alle drei in dieser Arbeit betrachteten Methoden für Online Diskriminanzanalyse sind die entsprechenden Aktualisierungsformeln für die verschobenen Zeitpunkte $z_i^{(c)}$ noch einmal in Tabelle 7.2 zusammengefasst.

Tabelle 7.2: Berechnung der verschobenen Zeitpunkte $z_i^{(c)}$ zum Zeitpunkt i bei den verschiedenen Methoden für alle Klassen $c \in \{1, \ldots, M[+1]\}$.

	$g(\boldsymbol{x}_i) \neq c$	$g(\boldsymbol{x}_i) = c \in \{1, \ldots, M\}$	$g(\boldsymbol{x}_i) = M + 1$
Sequential ILDA (Abschnitt 4.2)	$z_i^{(c)} = z_{i-1}^{(c)}$	$z_i^{(c)} = \frac{n_{i-1}^{(c)} z_{i-1}^{(c)} + i}{n_{i-1}^{(c)} + 1}$	$z_i^{(M+1)} = i$
OLDC (Abschnitt 4.3)	$z_i^{(c)} = z_{i-1}^{(c)}$	$z_i^{(c)} = \frac{(1-\lambda) n_{i-1}^{(c)} z_{i-1}^{(c)} + \lambda i}{(1-\lambda) n_{i-1}^{(c)} + \lambda}$	$z_i^{(M+1)} = i$
LDA-AF/QDA-AF (Abschnitt 4.4)	$z_i^{(c)} = z_{i-1}^{(c)}$	$z_i^{(c)} = \left(1 - \frac{1}{N_i^{(c)}}\right) z_{i-1}^{(c)} + \frac{1}{N_i^{(c)}} \cdot i$	$z_i^{(M+1)} = i$

7.6 Anpassung der aktualisierten Kovarianzmatrizen

Auch wenn sich die Annahme eines concept drifts auf die Erwartungswertvektoren beschränkt – im Speziellen ein linearer Trend über die Zeit (7.1) unterstellt wird – und zeitinvariante Kovarianzmatrizen betrachtet werden, können die Schätzer für die Kovarianzmatrizen über die Zeit sehr stark von den wahren Kovarianzmatrizen abweichen, da in diese die Schätzer für die Erwartungswertvektoren einfließen. Insbesondere im Falle der Fisher LDA können dadurch im Laufe der Aktualisierungen bei bestimmten Datensituationen Probleme auftreten und folglich kann der Prognosefehler der LDA recht groß werden. In kleineren Vorsimulationen hat sich gezeigt, dass insbesondere der Schätzer für die Zwischen-den-Klassen Kovarianzmatrix recht schnell degeneriert, falls keine Anpassung vorgenommen wird. Da die Zwischen-den-Klassen Kovarianzmatrix essentiell für die Bestimmung der Diskriminanzkomponenten in der Fisher LDA ist, ist es daher wichtig repräsentative Schätzer für die „aktuellen“ Erwartungswertvektoren bzw. jene des folgenden Zeitpunktes $t+1$ heranzuziehen.

Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ Vor der Bestimmung der Diskriminanzkomponenten und dem Lösen des Eigenwertproblems von $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}\boldsymbol{\Sigma}^{-1/2}$ (vgl. Seite 48) wird daher in jedem Schritt die Zwischen-den-Klassen Kovarianzmatrix neu geschätzt, indem der Schätzer für den Erwartungswertvektor aus Klasse c durch die Prognose $\hat{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t+1}}$ (vgl. (7.8) bzw. (7.12)) aus dem linearen Modell (7.3) ersetzt wird, falls diese Prognose für Klasse c bereits existiert. Ansonsten wird in der Summe weiterhin der Mittelwertvektor betrachtet:

$$\tilde{\boldsymbol{m}}^{(c)}_{t+1} = \hat{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t+1}} \cdot \mathbb{1}_{\left\{\exists\, \hat{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t+1}}\right\}} + \boldsymbol{m}^{(c)}_{n^{(c)}_{t}} \cdot \mathbb{1}_{\left\{\nexists\, \hat{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t+1}}\right\}}. \tag{7.15}$$

Anstelle des Schätzers aus (4.12) bzw. des Schätzers (4.14) für die Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ wird somit zum Zeitpunkt t folgende Form bei der Erweiterung der Sequential Incremental LDA betrachtet:

$$\hat{\boldsymbol{B}}_t = \frac{1}{n_t} \sum_{c=1}^{M} \left(n_t^{(c)} \left(\tilde{\boldsymbol{m}}^{(c)}_{t+1} - \hat{\boldsymbol{m}}_{t+1} \right) \left(\tilde{\boldsymbol{m}}^{(c)}_{t+1} - \hat{\boldsymbol{m}}_{t+1} \right)^T \right), \tag{7.16}$$

wobei auch der gesamte Mittelwertvektor zunächst neu berechnet wird:

$$\hat{\boldsymbol{m}}_{t+1} = \frac{1}{n_t} \sum_{c=1}^{M} n_t^{(c)} \tilde{\boldsymbol{m}}^{(c)}_{t+1}.$$

Gepoolte Kovarianzmatrix innerhalb der Klassen $\boldsymbol{S}$ Natürlich kann auch die gepoolte Kovarianzmatrix innerhalb der Klassen $\boldsymbol{S}$ im Laufe der Zeit degenerieren. Insbesondere

im Falle des unterstellten linearen Trends der Erwartungswertvektoren der Klassen resultieren im Laufe der Zeit sehr „gestreckte“ Kovarianzmatrizen. Besonders, wenn Update-Methoden für die Fisher LDA (*Sequential ILDA*) oder die Kanonische LDA ohne Gewichtungen (*OLDC* mit fester Lernrate $\lambda = 0.5$) herangezogen werden. Diese Schätzer für die Kovarianzmatrizen repräsentieren natürlich nicht mehr die aktuellen theoretischen Kovarianzmatrizen der Klassen. Anders als bei den Schätzern für die Erwartungswertvektoren hängen diese dann nicht nur „zeitlich hinterher“, sondern nehmen eine ganz andere Form an. Je nach Zusammenspiel der Trends in allen Klassen bzw. der speziellen Datensituation kann sich dies so auf die gepoolte Kovarianzmatrix zwischen den Klassen $\boldsymbol{S}$ auswirken, dass eine unerwünschte Rotation der Trenngeraden durch die Klassifikationsregel resultiert.

Im Gegensatz zur Zwischen-den-Klassen Kovarianzmatrix fließen in die gepoolte Kovarianzmatrix jedoch die einzelnen neuen Beobachtungen ein. Daher wird sie bei den Update-Methoden für Diskriminanzanalyse anhand dieser neuen Beobachtungen immer aktualisiert (vgl. z. B. (4.13) bei *Sequential ILDA*) und nicht wie die Zwischen-den-Klassen Kovarianzmatrix in (4.12) komplett neu berechnet. Die aktualisierte gepoolte Kovarianzmatrix muss anders als die aktuelle Zwischen-den-Klassen Kovarianzmatrix also immer für die nächste Aktualisierung gespeichert werden.

Die Idee der Erweiterung der Methoden für Online Diskriminanzanalyse besteht nun darin, dass lediglich in die Konstruktion der Klassifikationsregel eingegriffen wird, indem verbesserte Schätzer für die Erwartungswertvektoren der Klassen (und nun auch der Zwischen-den-Klassen Kovarianzmatrix) eingesetzt werden. Es wird allerdings nicht in die zuvor ausgeführten Aktualisierungsschritte zur Bestimmung der benötigten Größen für die Klassifikationsregel eingegriffen. An den beispielhaft in Abschnitt 4.2–4.4 vorgestellten Methoden für Online Diskriminanzanalyse selbst wird bis zur Konstruktion der Klassifikationsregel nichts verändert.

Vor diesem Hintergrund ist es nicht möglich die gepoolte Kovarianzmatrix neu anzupassen. Da sie auf allen bisherigen Beobachtungen basiert, die im Datenstrom nicht gespeichert werden (müssen), kann sie nicht wie $\boldsymbol{B}$ vor Konstruktion der Klassifikationsregel neu berechnet werden. Würden in der Aktualisierungsformel die Schätzer für die Erwartungswertvektoren der Klassen ersetzt (z. B. in (4.13)), dann müsste in die bestehenden Methoden eingegriffen werden. Daher wird dies im ersten Schritt vermieden. Bei der folgenden Simulationsstudie in Kapitel 9 hat sich der Verzicht auf eine Anpassung der Matrix $\boldsymbol{S}$ nicht deutlich negativ auf die resultierenden Klassifikationsgrenzen ausgewirkt. Auch ohne diese Anpassung ist eine deutliche Verbesserung der Prognosegüte der Klassifikatoren möglich.

7.7 Verbesserung der Prognosegüte der Klassifikatoren

Zur Verbesserung der Prognosegüte des stetig aktualisierten Klassifikators in der Online Diskriminanzanalyse wird der bisherige schrittweise aktualisierte Mittelwertvektor $\boldsymbol{m}^{(c)}_{n_t^{(c)}}$

durch den aus der Vorhersage des lokalen linearen Regressionsmodells resultierenden Schätzer (7.12)

$$\hat{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t+1}} = \hat{\boldsymbol{\beta}}^{(c)}_{0t} + \hat{\boldsymbol{\beta}}^{(c)}_{1t}(t+1), \quad c = 1, \ldots, M, \tag{7.17}$$

bzw. genauer $\tilde{\boldsymbol{m}}^{(c)}_{t+1}$ aus (7.15) ersetzt, um die Beobachtung des Zeitpunktes $t+1$ zu prognostizieren.

Abhängig davon, ob eine Online Variante der Fisher Diskriminanzanalyse (hier: *Sequential Incremental LDA* (Abschnitt 4.2)) oder der Kanonischen Linearen (oder Quadratischen) Diskriminanzanalyse (hier: *OLDC* (Abschnitt 4.3) und *Online Diskriminanzanalyse mit adaptivem Vergessen* (Abschnitt 4.4)) betrachtet wird, sieht die Klassifikationsregel im Datenstrom zum Zeitpunkt t folgendermaßen aus:

Sequential Incremental LDA Die *Sequential Incremental LDA* basiert auf der Fisher Diskriminanzanalyse, deren Klassifikationsregel in (3.45) bzw. (3.47) vorgestellt wurde. In der speziellen Klassifikationsregel zum Zeitpunkt t zur Prognose der Klasse der nachfolgenden Beobachtung $\boldsymbol{x}_{t+1}$ werden die theoretischen Größen durch die empirischen Schätzer des Zeitpunktes t ersetzt. Für den Erwartungswertvektor der Klasse c wird (7.17) bzw. (7.15) als Schätzer herangezogen, sodass die Prognose folgendermaßen gegeben ist:

$$\hat{c}_{t+1} = \underset{c=1,\ldots,M}{\arg\min} \sum_{j=1}^{r} \left(\boldsymbol{\alpha}_j^T \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}^{(c)}_{t+1} \right) \right)^2 - 2 \log P_t^{(c)}. \tag{7.18}$$

Zur Bestimmung der Diskriminanzkomponenten $\boldsymbol{\alpha}_j$ wird der Schätzer für die Zwischen-den-Klassen Kovarianzmatrix wie in Abschnitt 7.6 beschrieben durch den Schätzer (7.16) ersetzt, welcher ebenfalls die Prognose (7.17) aus dem linearen Regressionsmodell mit einbezieht. Es wird somit das Eigenwertproblem des Matrizenproduktes $\tilde{\boldsymbol{S}}_t^{-1/2} \hat{\boldsymbol{B}}_t \tilde{\boldsymbol{S}}_t^{-1/2}$ gelöst. Nach (3.41) ergeben sich mithilfe der resultierenden Eigenvektoren die Diskriminanzkomponenten durch Transformation:

$$\boldsymbol{\alpha}_j := \tilde{\boldsymbol{S}}_t^{-1/2} \boldsymbol{\nu}_j,$$

wobei die $\boldsymbol{\nu}_j,\ j = 1, \ldots, r$, die normalisierten Eigenvektoren mit $\boldsymbol{\nu}_j^T \boldsymbol{\nu}_j = 1$ zu den entsprechenden r positiven Eigenwerten des Eigenwertproblems sind.

OLDC (Online Linear Discriminant Classifier) Die Methode *OLDC* ist eine Erweiterung der Kanonischen LDA für Datenströme, bei welcher die Klassifikationsregel auf der Annahme normalverteilter Beobachtungen und daraus resultierend auf Betrachtung der Diskriminanzfunktion (3.22) bzw. (3.23) basiert. In der Klassifikationsregel (4.30) werden die Schätzer der Erwartungswertvektoren der Klassen durch (7.17) bzw. (7.15) ersetzt, sodass für die zukünftige Beobachtung $\boldsymbol{x}_{t+1}$ die Klasse $\hat{c}_{t+1}$ prognostiziert wird durch

$$\hat{c}_{t+1} = \underset{c=1,\ldots,M}{\arg\max} \left(\boldsymbol{x}_{t+1}^T \boldsymbol{S}_t^{-1} \tilde{\boldsymbol{m}}^{(c)}_{t+1} - \frac{1}{2} \left(\tilde{\boldsymbol{m}}^{(c)}_{t+1} \right)^T \boldsymbol{S}_t^{-1} \tilde{\boldsymbol{m}}^{(c)}_{t+1} + \log P_t^{(c)} \right). \tag{7.19}$$

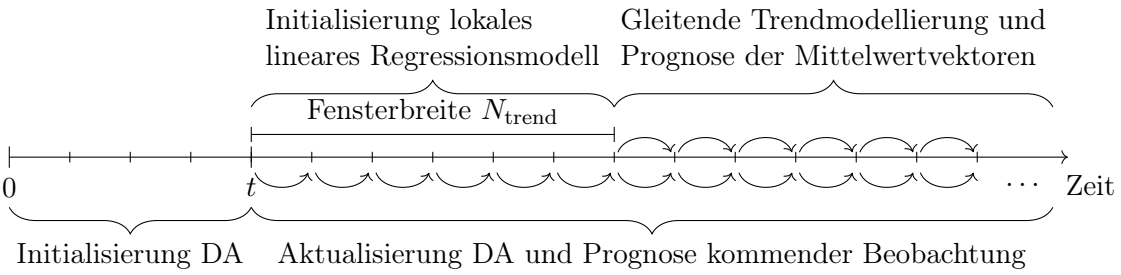


Abbildung 7.3: Schaubild zur Darstellung der Idee der Integration lokaler linearer Regressionsmodelle zur Verbesserung der Vorhersage bei Methoden der Online DA.

Online Diskriminanzanalyse mit adaptivem Vergessen Die Klassifikationsregel bei der Online Diskriminanzanalyse mit adaptivem Vergessen wurde in (4.63) auf Seite 86 vorgestellt. Ersetzen des bisherigen Mittelwertvektors führt zu der erweiterten Klassifikationsregel im Falle der Quadratischen Diskriminanzanalyse für die kommende Beobachtung $\boldsymbol{x}_{t+1}$:

$$\tilde{c}_{t+1} = \underset{c=1,\ldots,M}{\arg\min} \left(\frac{1}{2}\, d_t^{(c)} + \frac{1}{2} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{t+1}^{(c)} \right)^T \boldsymbol{G}_t^{(c)} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{t+1}^{(c)} \right) - \log \tilde{P}_t^{(c)} \right). \quad (7.20)$$

Für die Lineare Diskriminanzanalyse wird folgende aus (4.64) erweiterte Klassifikationsregel betrachtet:

$$\tilde{c}_{t+1} = \underset{c=1,\ldots,M}{\arg\min} \left(\frac{1}{2}\, d_t^{(P)} + \frac{1}{2} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{t+1}^{(c)} \right)^T \boldsymbol{G}_t^{(P)} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{t+1}^{(c)} \right) - \log \tilde{P}_t^{(c)} \right). \quad (7.21)$$

Die Erweiterung von Methoden für Online Diskriminanzanalyse für Datenströme mit concept drift bzw. die Einbindung eines lokalen linearen Regressionsmodells zur Modellierung des Trends der Erwartungswertvektoren ist in Algorithmus 5 zusammengefasst. Dieser Algorithmus ist in Abbildung 7.3 schematisch veranschaulicht.

Algorithmus 5 Erweiterung von Methoden für Online Diskriminanzanalyse für Datenströme mit Concept Drift

Require: Datenstrom $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots\}$, N_{trend}.

1: Initialisierung der Online Diskriminanzanalyse anhand der ersten t Beobachtungen.
2: Wenn neue Beobachtung $\boldsymbol{x}_{t+1}$ im Datenstrom ankommt
→ Aktualisierung der Größen der Online DA (z. B. Abschnitte 4.2, 4.3, 4.4).
3: Anpassung eines lokalen linearen Regressionsmodells (7.3) an die $n_{\text{trend}}^{(c)}$ Mittelwertvektoren jeder Klasse des Intervalls (7.2) der Fensterbreite N_{trend}
(zum ersten Mal nach N_{trend} Aktualisierungen der Diskriminanzanalyse).
4: Schätzen des kommenden Erwartungswertvektors durch Prognose des Mittelwertvektors (7.17) für jede Klasse c.
5: Erstellung der Klassifikationsregel der Diskriminanzanalyse durch prognostizierte Mittelwertvektoren (7.17) bzw. (7.15) und Vorhersage der neuen Beobachtung im Datenstrom durch z. B. (7.18), (7.19), (7.20) oder (7.21).
6: Wiederhole 2–5 für gesamten Datenstrom.

8 Untersuchung der Erwartungstreue der erweiterten Schätzfunktionen

In diesem Kapitel werden die erweiterten Schätzfunktionen für die Erwartungswertvektoren der Klassen aus Kapitel 7, welche bei der Prognose in der Diskriminanzanalyse herangezogen werden, miteinander verglichen. Es wird gezeigt, dass diese im Gegensatz zu den ursprünglichen Schätzfunktionen (vgl. Kapitel 6) auch im Falle eines linearen Trends der Erwartungswertvektoren unter bestimmten Annahmen weiterhin erwartungstreu für den Erwartungswertvektor der Verteilung des kommenden Zeitpunktes sein können. Dadurch werden erwartungstreue Schätzer bei der Prognose der LDA/QDA betrachtet. Gleichzeitig bleibt die Eigenschaft der Erwartungstreue im Falle stabiler Verteilungen erhalten.

Die allgemeine Form des erweiterten Schätzers wurde in Kapitel 7 mittels der geschlossenen Formen der KQ-Schätzer des linearen Modells hergeleitet (vgl. (7.8) und (7.10)–(7.12)):

$$\begin{aligned}\hat{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t+1}} =: \hat{\boldsymbol{y}}^{(c)}_{t+1} &= \hat{\boldsymbol{\beta}}^{(c)}_{0t} + \hat{\boldsymbol{\beta}}^{(c)}_{1t}(t+1) \\ &= \bar{\boldsymbol{y}}^{(c)}_t - \left(\frac{\sum_{i\in I}\left(z^{(c)}_i - \bar{z}^{(c)}_t\right)\left(\boldsymbol{y}^{(c)}_i - \bar{\boldsymbol{y}}^{(c)}_t\right)}{\sum_{i\in I}\left(z^{(c)}_i - \bar{z}^{(c)}_t\right)^2} \right)\left(\bar{z}^{(c)}_t - (t+1)\right), \end{aligned} \tag{8.1}$$

wobei I das Intervall der Zeitpunkte für das lokale lineare Regressionsmodell ist (vgl. (7.2)):

$$I = \underbrace{\{k : g(\boldsymbol{x}_k) = c,\ k = t - N_{\text{trend}} + 1, \ldots, t\}}_{n^{(c)}_{\text{trend}} \text{ Zeitpunkte/Beobachtungen}}. \tag{8.2}$$

Der Schätzer (8.1) ist die Prognose des Erwartungswertvektors für Klasse c des Zeitpunktes $t+1$ durch ein lokales lineares Regressionsmodell, welches an die letzten $n^{(c)}_{\text{trend}}$ Mittelwertvektoren aus den Aktualisierungsschritten der Diskriminanzanalyse angepasst wird.

Die entsprechende zunächst von der Methode für Online Diskriminanzanalyse unabhängige Formulierung der Schätzfunktion zum Zeitpunkt t sieht demnach folgendermaßen aus:

$$T^{(c)}_2(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) = \bar{\boldsymbol{Y}}^{(c)}_t - \left(\frac{\sum_{i\in I}\left(Z^{(c)}_i - \bar{Z}^{(c)}_t\right)\left(\boldsymbol{Y}^{(c)}_i - \bar{\boldsymbol{Y}}^{(c)}_t\right)}{\sum_{i\in I}\left(Z^{(c)}_i - \bar{Z}^{(c)}_t\right)^2} \right)\left(\bar{Z}^{(c)}_t - (t+1)\right). \tag{8.3}$$

Wenn eine Einschränkung auf den Spezialfall $g(\boldsymbol{x}_i) = c$, $i = 1, \ldots, t$, aus Voraussetzung 4 (Seite 138) erfolgt, werden im linearen Regressionsmodell nur nicht-stochastische Regres-

soren betrachtet (vgl. Abschnitte 7.3 und 7.4). In diesem Spezialfall nimmt die allgemeine Schätzfunktion folgende Form an:

$$T_2^{(c)}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)}) = \bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1)\right). \tag{8.4}$$

Je nach Methode variiert die Form der Zufallsvariablen der verschobenen Zeitpunkte $Z_i^{(c)}$ und folglich auch der Mittelwerte $\bar{Z}_t^{(c)}$. Die $\boldsymbol{y}_i^{(c)}$ in (8.1) sind jeweils Schätzer für die Erwartungswertvektoren aus Klasse c zum Zeitpunkt i, an welche das lokale lineare Regressionsmodell angepasst wird. Der Vektor $\bar{\boldsymbol{y}}_t^{(c)}$ ist der Mittelwert über alle $n_{\text{trend}}^{(c)}$ vergangenen Schätzwerte $\boldsymbol{y}_i^{(c)}$ mit $i \in I$. $\boldsymbol{Y}_i^{(c)}$ sowie $\bar{\boldsymbol{Y}}_t^{(c)}$ entsprechen den jeweiligen Schätzfunktionen. Es gilt somit

$$\boldsymbol{Y}_i^{(c)} = T_1^{(c)}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_i) \tag{8.5}$$

mit $T_1^{(c)}$ der jeweiligen ursprünglichen Schätzfunktion (z. B. $T_1^{(c),P}$, $T_1^{(c),K}$ und $T_1^{(c),A}$) der verschiedenen betrachteten Methoden für Online Diskriminanzanalyse (vgl. Kapitel 6).

Für die Analyse der Schätzfunktionen werden wieder die Situationen aus Abschnitt 6.1 betrachtet. Es wird somit untersucht wie sich die Erwartungswerte der Schätzfunktionen unter der Annahme eines linearen Trends der Erwartungswertvektoren in den Klassen sowie bei stabiler Verteilung ohne Vorliegen eines concept drifts verhalten. Der folgende Abschnitt 8.1 umfasst zunächst einen kurzen Beweis für die Erwartungstreue der Schätzfunktionen (8.4) im Spezialfall, dass bis zum Zeitpunkt t lediglich Beobachtungen in Klasse c realisiert werden (Voraussetzung 4 auf Seite 138). In den Abschnitten 8.2–8.4 folgen ausführlichere Beweise für die Erwartungstreue der Schätzfunktionen der Erweiterungen von *Sequential ILDA*, *OLDC* und *QDA-AF/LDA-AF* unter bestimmten Voraussetzungen. In Abschnitt 8.5 werden die Ergebnisse diskutiert und mit jenen aus Kapitel 6, in welchem die Erwartungstreue der Schätzfunktionen der nicht-erweiterten Methoden für Online Diskriminanzanalyse untersucht wurde, verglichen.

Im Folgenden seien autokorrelierte Fehler für (7.5) unterstellt, sodass $\boldsymbol{\epsilon}_j^{(c)} \sim (\boldsymbol{0}, \sigma^2\boldsymbol{\Psi})$ bzw. $\boldsymbol{\epsilon}_j^{(c)} | \boldsymbol{Z}_j^{(c)} \sim (\boldsymbol{0}, \sigma^2\boldsymbol{\Psi})$ im allgemeinen Fall bei stochastischen Regressoren (vgl. Abschnitt 7.4).

8.1 Kurzer Beweis im Spezialfall

Da im Falle nicht-stochastischer Regressoren $z_i^{(c)}$ und autokorrelierter Fehler der gewöhnliche KQ-Schätzer $\left(\hat{\boldsymbol{\beta}}_{0t}^{(c)}, \hat{\boldsymbol{\beta}}_{1t}^{(c)}\right)^T$ aus (7.10)/(7.11) erwartungstreu für den wahren Parametervektor ist (vgl. Abschnitt 7.4), gilt für den Erwartungswert der Prognose (7.17):

$$\mathrm{E}\left(\hat{\boldsymbol{m}}_{n_{t+1}^{(c)}}^{(c)}\right) = \mathrm{E}\left(\hat{\boldsymbol{\beta}}_{0t}^{(c)} + \hat{\boldsymbol{\beta}}_{1t}^{(c)}(t+1)\right) = \mathrm{E}\left(\hat{\boldsymbol{\beta}}_{0t}^{(c)}\right) + \mathrm{E}\left(\hat{\boldsymbol{\beta}}_{1t}^{(c)}\right)(t+1) = \boldsymbol{\beta}_{0t}^{(c)} + \boldsymbol{\beta}_{1t}^{(c)}(t+1).$$

Da bei Unterstellung des linearen Trends (7.1) und Betrachtung der verschobenen Zeitpunkte im linearen Regressionsmodell $\boldsymbol{\beta}_{0t}^{(c)} := \boldsymbol{\beta}_0^{(c)}$ und $\boldsymbol{\beta}_{1t}^{(c)} := \boldsymbol{\beta}_1^{(c)}$ gilt (vgl. Seite 166), folgt:

$$\mathrm{E}\left(\hat{\boldsymbol{m}}_{n_{t+1}^{(c)}}^{(c)}\right) = \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(t+1) \overset{(7.1)}{=} \boldsymbol{\mu}_{t+1}^{(c)}.$$

Falls eine stabile Verteilung unterstellt wird, gilt in (7.1) $\boldsymbol{\beta}_1^{(c)} := \mathbf{0}$ bzw. $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_1^{(c)} = \boldsymbol{\mu}_2^{(c)} = \ldots$, sodass weiterhin Erwartungstreue der Prognose für $\boldsymbol{\mu}_{t+1}^{(c)}$ vorliegt.

Da in diesem Abschnitt für den kompakten Beweis im Spezialfall sehr viele Annahmen unterstellt werden, wird die Erwartungstreue der Schätzfunktion im Folgenden für alle drei erweiterten Methoden – teilweise auch unter allgemeineren Voraussetzungen – nochmals ausführlicher und anschaulicher hergeleitet.

8.2 Sequential Incremental LDA (Sequential ILDA)

Als Erstes wird die Online LDA ohne Vergessen bzw. Gewichtung (alle Beobachtungen haben ein identisches Gewicht), also die Methode von Pang et al. (2005b) betrachtet, welche in Abschnitt 4.2 vorgestellt wurde. In Abschnitt 6.2 konnte die Erwartungstreue der ursprünglichen Schätzfunktion $T_1^{(c),P}$ zum Zeitpunkt t für den Erwartungswertvektor der Verteilung des Zeitpunktes $t+1$ im Falle stabiler Verteilungen über die Zeit bewiesen werden. Gleichzeitig zeigte sich, dass diese Erwartungstreue für $\boldsymbol{\mu}_{t+1}^{(c)}$ bei linearem Trend der Erwartungswertvektoren nicht mehr gilt. In diesem Abschnitt wird gezeigt, dass die Schätzfunktion (8.3) der erweiterten Methode in beiden Fällen erwartungstreu ist.

Zum Zeitpunkt t im Datenstrom sehen die letzten $n_{\text{trend}}^{(c)}$ standardmäßigen Mittelwertvektoren (4.3) als Schätzer für die Erwartungswertvektoren der Klassen in iterativer Form folgendermaßen aus (vgl. (6.2) und (4.8) für entsprechende rekursive Variante):

$$\boldsymbol{y}_i^{(c)} := \boldsymbol{m}_{n_i^{(c)}}^{(c)} = \bar{\boldsymbol{x}}_i^{(c)} = \frac{1}{n_i^{(c)}} \sum_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j \le i}} \boldsymbol{x}_j = \frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}} \sum_{j=1}^{i} \left(\boldsymbol{x}_j \cdot \mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}}\right), \quad i \in I.$$

Diese geschätzten Erwartungswertvektoren werden jeweils durch ein lineares Regressionsmodell modelliert (vgl. (7.3) in Abschnitt 7.3). Da die Klassenmittel bei der Online LDA im Datenstrom mit jeder neuen Beobachtung aktualisiert werden, werden dabei verschobene Zeitpunkte $z_i^{(c)}$ betrachtet, die in das lineare Regressionsmodell einfließen (vgl. (7.6)). Die Überlegung dazu wurde in Abschnitt 7.3 erläutert:

$$z_i^{(c)} = \frac{1}{n_i^{(c)}} \sum_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j \le i}} j = \frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}} \sum_{j=1}^{i} \left(j \cdot \mathbb{1}_{\{g(\boldsymbol{x}_j)=c\}}\right). \tag{8.6}$$

Die Zeitpunkte $z_i^{(c)}$ sind zufällig (stochastische Regressoren). Die entsprechende Zufallsvariable hat die folgende Form:

$$Z_i^{(c)} = \frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}}} \sum_{j=1}^{i} \left(j \cdot \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} \right). \tag{8.7}$$

Der Mittelwert über alle $n_{\text{trend}}^{(c)}$ Zufallsvariablen aus dem Intervall I zum Zeitpunkt t ergibt sich durch

$$\bar{Z}_t^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i \in I} \left(\frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}}} \sum_{j=1}^{i} \left(j \cdot \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} \right) \right). \tag{8.8}$$

Die Schätzfunktion der Erweiterung von *Sequential Incremental LDA* (vgl. (8.3))

$$T_2^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) = \bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1) \right) \tag{8.9}$$

setzt sich aus den Zufallsvariablen und -vektoren $\boldsymbol{Y}_i^{(c)} := T_1^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i)$ (vgl. (6.3) auf Seite 139), $\bar{\boldsymbol{Y}}_t^{(c)}$, $Z_i^{(c)}$ und $\bar{Z}_t^{(c)}$ zusammen.

8.2.1 Situation: Stabile Verteilung

Im Falle einer stabilen Verteilung in Klasse c über die Zeit gilt (vgl. Seite 138): $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_1^{(c)} = \boldsymbol{\mu}_2^{(c)} = \ldots$.

Zunächst wird die Erwartungstreue der Schätzfunktion (8.4) unter der Voraussetzung 4 (vgl. Seite 138), dass bis zum Zeitpunkt t im Datenstrom nur Beobachtungen aus Klasse c auftreten ($g(\boldsymbol{x}_i) = c$, $i = 1, \ldots, t$) im Vergleich zu Abschnitt 8.1 genauer untersucht.

Satz 12. Unter der Annahme einer stabilen Verteilung und dem Spezialfall aus Voraussetzung 4 (Seite 138) ist die bei der Erweiterung der Update-Methode *Sequential ILDA* (Pang et al., 2005b) verwendete Schätzfunktion $T_2^{(c),P}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})$ aus (8.4) bzw. (8.9) zum Zeitpunkt t erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c:

$$\mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)}) \right) = \boldsymbol{\mu}^{(c)}.$$

Beweis. Der Zufallsvektor $\boldsymbol{Y}_i^{(c)}$ entspricht der ursprünglichen Schätzfunktion (8.5) bzw. (6.3) des Zeitpunktes i: $\boldsymbol{Y}_i^{(c)} = T_1^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i)$. Die Zufallsvektoren, auf denen die Schätzfunktion basiert, sowie deren Erwartungswerte sehen daher folgendermaßen aus:

$$\boldsymbol{Y}_i^{(c)} = \frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}}} \sum_{j=1}^{i} \left(\boldsymbol{X}_j \cdot \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} \right) \overset{(6.3)}{=} T_1^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i) = \bar{\boldsymbol{X}}_i^{(c)},$$

$$\bar{\boldsymbol{Y}}_t^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i \in I} \left(\frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}}} \sum_{j=1}^{i} \left(\boldsymbol{X}_j \cdot \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} \right) \right), \tag{8.10}$$

$$\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) = \mathrm{E}\left(\bar{\boldsymbol{X}}_i^{(c)}\right) \overset{(6.5)}{=} \boldsymbol{\mu}^{(c)},$$

$$\begin{aligned}
\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) &= \mathrm{E}\left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{i \in I} \left(\frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}}} \sum_{j=1}^{i} \left(\boldsymbol{X}_j \cdot \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} \right) \right) \right) \\
&= \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i \in I} \mathrm{E}\left(\frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}}} \sum_{j=1}^{i} \left(\boldsymbol{X}_j \cdot \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} \right) \right) \\
&\overset{(6.5)}{=} \frac{1}{n_{\text{trend}}^{(c)}} \underbrace{\sum_{i \in I} \boldsymbol{\mu}^{(c)}}_{n_{\text{trend}}^{(c)} \text{ Summanden}} = \frac{n_{\text{trend}}^{(c)} \boldsymbol{\mu}^{(c)}}{n_{\text{trend}}^{(c)}} = \boldsymbol{\mu}^{(c)}.
\end{aligned}$$

Im Spezialfall der Voraussetzung 4, dass $g(\boldsymbol{x}_i) = c,\ i = 1, \ldots, t$, gilt $\boldsymbol{X}_i := \boldsymbol{X}_i^{(c)}$. Zudem vereinfachen sich in diesem Spezialfall die Zeitpunkte $z_i^{(c)}$ aus (8.6) zu:

$$z_i^{(c)} = \frac{1}{n_i^{(c)}} \sum_{\substack{j:\, g(\boldsymbol{x}_j) = c \\ j \leq i}} j = \frac{1}{i} \sum_{j=1}^{i} j = \frac{i(i+1)}{2i} = \frac{i+1}{2}, \quad i = t - n_{\text{trend}}^{(c)} + 1, \ldots, t. \tag{8.11}$$

Der Mittelwert über alle Zeitpunkte $i \in I$ ist dann ebenfalls fest:

$$\begin{aligned}
\bar{z}_t^{(c)} &= \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \frac{i+1}{2} = \frac{1}{2n_{\text{trend}}^{(c)}} \left(\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} i + n_{\text{trend}}^{(c)} \right) \\
&= \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t - n_{\text{trend}}^{(c)})(t - n_{\text{trend}}^{(c)} + 1)}{2} \right) + \frac{1}{2}.
\end{aligned} \tag{8.12}$$

Die Schätzfunktion $T_2^{(c),P}$ aus (8.4) bzw. (8.9) hängt somit nur noch von den Zufallsvektoren $\boldsymbol{Y}_i^{(c)}$ und $\bar{\boldsymbol{Y}}_t^{(c)}$ bzw. indirekt $\boldsymbol{X}_i := \boldsymbol{X}_i^{(c)}$ ab. Der Ausdruck von Produkten und Quotienten von Zufallsvariablen und Zufallsvektoren entfällt. Der Erwartungswert der Schätzfunktion $T_2^{(c),P}$ vereinfacht sich durch Einsetzen der obigen Überlegungen zu:

$$\begin{aligned}
&\mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)\right) = \mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})\right) \\
&= \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1) \right) \right) \\
&= \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} \right) - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right) \left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)} \right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} \right) \right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1) \right) \\
&= \boldsymbol{\mu}^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right) \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{\mu}^{(c)} \right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1) \right) = \boldsymbol{\mu}^{(c)} = \boldsymbol{\mu}_{t+1}^{(c)}.
\end{aligned} \tag{8.13}$$

□

Die Schätzfunktion $T_2^{(c),P}$ ist demnach im Falle stabiler Verteilungen in den Klassen weiterhin erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose des Zeitpunktes $t+1$, falls bis zum Zeitpunkt t alle Beobachtungen Klassenlabel c besitzen.

Die Schätzfunktion $T_2^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ ist auch im allgemeinen Fall erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose. Dazu wird die folgende Voraussetzung betrachtet, die durch die praktische Anwendbarkeit der Schätzwerte begründet ist:

Voraussetzung 7. Zum Zeitpunkt t seien innerhalb der letzten N_{trend} Zeitpunkte mindestens zwei Beobachtungen aus Klasse c realisiert, d. h. es existieren mindestens zwei Zeitpunkte $i \in \{t - N_{\text{trend}} + 1, \ldots, t\} : g(\boldsymbol{x}_i) = c$. Dies ist gleichbedeutend damit, dass das Intervall $I = \{k : g(\boldsymbol{x}_k) = c,\ k = t - N_{\text{trend}} + 1, \ldots, t\}$ aus (8.2) aus mindestens zwei Beobachtungen besteht und demnach $n_{\text{trend}}^{(c)} \geq 2$ ist. Falls dies nicht der Fall ist, wird kein lokales lineares Regressionsmodell angepasst und folglich der Schätzwert nicht berechnet. Die Schätzfunktion ist somit nicht definiert.

Es gelten demnach die folgenden Annahmen für die Auftrittswahrscheinlichkeiten (a-priori Wahrscheinlichkeiten der Klassen):

„Wahrscheinlichkeit, dass innerhalb von I keine Beobachtung in Klasse c realisiert wird“:

$$\mathrm{P}\left(\bigcap_{i=t-N_{\text{trend}}+1}^{t} \left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 0\right)\right) = 0 \Longleftrightarrow \prod_{i=t-N_{\text{trend}}+1}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 0\right) = 0.$$

„Wahrscheinlichkeit, dass innerhalb von I nur eine Beobachtung in Klasse c realisiert wird“:

$$\mathrm{P}\left(\bigcup_{i=t-N_{\text{trend}+1}}^{t} \left(\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1\right) \cap \bigcap_{\substack{j=t-N_{\text{trend}}+1 \\ j \neq i}}^{t} \left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 0\right)\right)\right) = 0$$

$$\Longleftrightarrow \sum_{i=t-N_{\text{trend}+1}}^{t} \left(\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1\right) \prod_{\substack{j=t-N_{\text{trend}}+1 \\ j \neq i}}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 0\right)\right) = 0.$$

Da die Summe über die Wahrscheinlichkeiten aller Möglichkeiten der Zusammensetzung der Zufallsvektoren von Zeitpunkt 1 bis t Eins ist, gilt:

$$\left(\sum_{l_1=0}^{1} \cdots \sum_{l_{t-N_{\text{trend}}}=0}^{1} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = l_1\right) \cdots \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_{t-N_{\text{trend}}} \sim F_c\}} = l_{t-N_{\text{trend}}}\right)\right)$$

$$\cdot \left(\sum_{i=t-N_{\text{trend}}+1}^{t} \sum_{\substack{j=t-N_{\text{trend}}+1 \\ j \neq i}}^{t} \left(\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1\right) \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 1\right) \right.\right.$$

$$\left.\left. \cdot \sum_{l_{t-N_{\text{trend}}+1}=0}^{1} \cdots \sum_{l_t=0}^{1} \prod_{\substack{k=t-N_{\text{trend}}+1 \\ k \neq i,\, k \neq j}}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}} = l_k\right)\right)\right)$$

$$
\begin{aligned}
&+ \left(\sum_{l_1=0}^{1} \cdots \sum_{l_{t-N_{\text{trend}}}=0}^{1} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = l_1\right) \cdots \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_{t-N_{\text{trend}}} \sim F_c\}} = l_{t-N_{\text{trend}}}\right) \right) \\
&\quad \cdot \left(\sum_{i=t-N_{\text{trend}}+1}^{t} \left(\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1\right) \prod_{\substack{j=t-N_{\text{trend}}+1 \\ j \neq i}}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 0\right) \right) \right) \\
&+ \left(\sum_{l_1=0}^{1} \cdots \sum_{l_{t-N_{\text{trend}}}=0}^{1} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = l_1\right) \cdots \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_{t-N_{\text{trend}}} \sim F_c\}} = l_{t-N_{\text{trend}}}\right) \right) \\
&\quad \cdot \prod_{i=t-N_{\text{trend}}+1}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 0\right) \\
&= 1 \\
&\Rightarrow \left(\sum_{l_1=0}^{1} \cdots \sum_{l_{t-N_{\text{trend}}}=0}^{1} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = l_1\right) \cdots \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_{t-N_{\text{trend}}} \sim F_c\}} = l_{t-N_{\text{trend}}}\right) \right) \\
&\quad \cdot \left(\sum_{i=t-N_{\text{trend}}+1}^{t} \sum_{\substack{j=t-N_{\text{trend}}+1 \\ j \neq i}}^{t} \left(\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1\right) \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 1\right) \right. \right. \\
&\qquad \left. \left. \cdot \sum_{l_{t-N_{\text{trend}}+1}=0}^{1} \cdots \sum_{l_t=0}^{1} \prod_{\substack{k=t-N_{\text{trend}}+1 \\ k \neq i,\, k \neq j}}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}} = l_k\right) \right) \right) = 1.
\end{aligned}
$$

Unter dieser praktischen Voraussetzung gilt die Erwartungstreue der Schätzfunktion auch im allgemeinen Fall (mit stochastischen Regressoren), was durch den folgenden Satz beschrieben wird.

Satz 13. Unter der Annahme einer stabilen Verteilung und der Voraussetzung 7 ist die bei der Erweiterung der Update-Methode *Sequential ILDA* (Pang et al., 2005b) verwendete Schätzfunktion $T_2^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ aus (8.9) zum Zeitpunkt t erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c:

$$\mathrm{E}\left(T_2^{(c),P}\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t\right)\right) = \boldsymbol{\mu}^{(c)}.$$

Beweis. Der Erwartungswert der Schätzfunktion lässt sich durch den bedingten Erwartungswert ausdrücken:

$$
\begin{aligned}
&\mathrm{E}\left(T_2^{(c),P}\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t\right)\right) \\
&= \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1)\right) \right) \\
&= \mathrm{E}\left(\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1)\right) \middle| \begin{matrix} \mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}}, \ldots, \\ \mathbb{1}_{\{\boldsymbol{X}_t \sim F_c\}} \end{matrix} \right) \right).
\end{aligned}
\tag{8.14}
$$

Aus Darstellungsgründen bezeichne im Folgenden: $B_i := \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}}$, $i = 1, \ldots, t$. Der innere Erwartungswert von (8.14) sieht damit folgendermaßen aus:

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1) \right) \middle| B_1, \ldots, B_t \right).$$

Da $\boldsymbol{Y}_i^{(c)} = T_1^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i)$, gilt $\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) = \boldsymbol{\mu}^{(c)}$ für alle verschiedenen möglichen $2^t - N_{\text{trend}} - 1$ Kombinationen an Ausprägungen der B_i, $i = 1, \ldots, t$ (vgl. Voraussetzung 7). Vergleiche dazu den Beweis von Satz 4 ab Seite 140.

Der bedingte Erwartungswert von $\bar{\boldsymbol{Y}}_t^{(c)}$ für alle $2^t - N_{\text{trend}} - 1$ verschiedenen betrachteten Kombinationen an Ausprägungen der B_i für $i = 1, \ldots, t$ ergibt sich durch
(mit Fallunterscheidungen $A_1 := \{B_i = 1 \ (i = t-1, t),\ B_j = 0 \ (j \neq i)\}$,
$A_2 := \{B_i = 1 \ (i = t - N_{\text{trend}} + 1, t - N_{\text{trend}} + 2),\ B_j = 0 \ (j \neq i)\}$,
$A_3 := \{B_i = 1 \ (i = 1, t-1, t),\ B_j = 0 \ (j \neq i)\}$,
$A_4 := \{B_i = 1 \ (i = 1, \ldots, t)\}$)

$$\begin{aligned}
&\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} \middle| B_1, \ldots, B_{t-N_{\text{trend}}}, B_{t-N_{\text{trend}}+1}, \ldots, B_t \right) \\
&= \mathrm{E}\left(\frac{\sum_{i \in I} \left(\frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}}} \sum_{j=1}^{i} \left(\boldsymbol{X}_j \cdot \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} \right) \right)}{\sum_{l=t-N_{\text{trend}}+1}^{t} \mathbb{1}_{\{\boldsymbol{X}_l \sim F_c\}}} \middle| B_1, \ldots, B_{t-N_{\text{trend}}}, B_{t-N_{\text{trend}}+1}, \ldots, B_t \right) \\
&= \begin{cases}
\mathrm{E}\left(\frac{\boldsymbol{X}_{t-1}^{(c)} + \frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)} + \boldsymbol{X}_t^{(c)} \right)}{2} \middle| A_1 \right), & A_1, \\
\vdots & \\
\mathrm{E}\left(\frac{\boldsymbol{X}_{t-N_{\text{trend}}+1}^{(c)} + \frac{1}{2}\left(\boldsymbol{X}_{t-N_{\text{trend}}+1}^{(c)} + \boldsymbol{X}_{t-N_{\text{trend}}+2}^{(c)} \right)}{2} \middle| A_2 \right), & A_2, \\
\vdots & \\
\mathrm{E}\left(\frac{\frac{1}{2}\left(\boldsymbol{X}_1^{(c)} + \boldsymbol{X}_{t-1}^{(c)} \right) + \frac{1}{3}\left(\boldsymbol{X}_1^{(c)} + \boldsymbol{X}_{t-1}^{(c)} + \boldsymbol{X}_t^{(c)} \right)}{2} \middle| A_3 \right), & A_3, \\
\vdots & \\
\mathrm{E}\left(\frac{\sum_{i=t-N_{\text{trend}}+1}^{t} \left(\frac{1}{i} \sum_{j=1}^{i} \boldsymbol{X}_j^{(c)} \right)}{N_{\text{trend}}} \middle| A_4 \right), & A_4,
\end{cases} \\
&= \begin{cases}
\mathrm{E}\left(\boldsymbol{X}^{(c)} \middle| B_i = 1 \ (i = t-1, t),\ B_j = 0 \ (j \neq i) \right), & B_i = 1 \ (i = t-1, t), \\
& B_j = 0 \ (j \neq i), \\
\vdots & \\
\mathrm{E}\left(\boldsymbol{X}^{(c)} \middle| \begin{matrix} B_i = 1 \ (i = t - N_{\text{trend}} + 1, t - N_{\text{trend}} + 2), \\ B_j = 0 \ (j \neq i) \end{matrix} \right), & B_i = 1 \ (i = t - N_{\text{trend}} + 1, \\
& \qquad t - N_{\text{trend}} + 2), \\
& B_j = 0 \ (j \neq i), \\
\vdots & \\
\mathrm{E}\left(\boldsymbol{X}^{(c)} \middle| B_i = 1 \ (i = 1, t-1, t),\ B_j = 0 \ (j \neq i) \right), & B_i = 1 \ (i = 1, t-1, t), \\
& B_j = 0 \ (j \neq i), \\
\vdots & \\
\mathrm{E}\left(\boldsymbol{X}^{(c)} \middle| B_i = 1 \ (i = 1, \ldots, t) \right), & B_i = 1 \ (i = 1, \ldots, t),
\end{cases}
\end{aligned}$$

$$= \begin{cases} \boldsymbol{\mu}^{(c)}, & B_i = 1 \ (i = t-1, t), \ B_j = 0 \ (j \neq i), \\ \vdots & \\ \boldsymbol{\mu}^{(c)}, & B_i = 1 \ (i = t - N_{\text{trend}} + 1, t - N_{\text{trend}} + 2), \ B_j = 0 \ (j \neq i), \\ \vdots & \\ \boldsymbol{\mu}^{(c)}, & B_i = 1 \ (i = 1, t-1, t), \ B_j = 0 \ (j \neq i), \\ \vdots & \\ \boldsymbol{\mu}^{(c)}, & B_i = 1 \ (i = 1, \ldots, t). \end{cases}$$

Es werden dabei alle $2^t - N_{\text{trend}} - 1$ möglichen Kombinationen an Ausprägungen der $B_i := \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}}$ für $i = 1, \ldots, t$ betrachtet, bei denen $B_j = 1$ für mindestens zwei Zeitpunkte $j \in \{t - N_{\text{trend}} + 1, \ldots, t\}$ (Voraussetzung 7). Die durch A bedingten Zufallsvariablen $Z_i^{(c)}$ bzw. $\bar{Z}_t^{(c)}$, wobei A eine der $2^t - N_{\text{trend}} - 1$ möglichen Kombinationen repräsentiert, sind nicht mehr zufällig und werden durch $z_i^{(c)}$ bzw. $\bar{z}_t^{(c)}$ ausgedrückt.

Insgesamt ergibt sich der innere Erwartungswert von (8.14) durch die Betrachtung aller Kombinationen folgendermaßen
(mit Fallunterscheidungen $A_1 := \{B_i = 1 \ (i = t-1, t), \ B_j = 0 \ (j \neq i)\}$,
$A_2 := \{B_i = 1 \ (i = t - N_{\text{trend}} + 1, t - N_{\text{trend}} + 2), \ B_j = 0 \ (j \neq i)\}$,
$A_3 := \{B_i = 1 \ (i = 1, t-1, t), \ B_j = 0 \ (j \neq i)\}$,
$A_4 := \{B_i = 1 \ (i = 1, \ldots, t)\}$ wie oben):

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1) \right) \middle| B_1, \ldots, B_t \right)$$

$$= \begin{cases} \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1) \right) \middle| A_1 \right), & A_1, \\ \vdots & \\ \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1) \right) \middle| A_2 \right), & A_2, \\ \vdots & \\ \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1) \right) \middle| A_3 \right), & A_3, \\ \vdots & \\ \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1) \right) \middle| A_4 \right), & A_4, \end{cases}$$

$$
= \begin{cases}
\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\middle|A_1\right) - \left(\frac{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\middle|A_1\right)-\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\middle|A_1\right)\right)}{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)}-(t+1)\right), & A_1,\\
\vdots\\
\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\middle|A_2\right) - \left(\frac{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\middle|A_2\right)-\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\middle|A_2\right)\right)}{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)}-(t+1)\right), & A_2,\\
\vdots\\
\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\middle|A_3\right) - \left(\frac{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\middle|A_3\right)-\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\middle|A_3\right)\right)}{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)}-(t+1)\right), & A_3,\\
\vdots\\
\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\middle|A_4\right) - \left(\frac{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\middle|A_4\right)-\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\middle|A_4\right)\right)}{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)}-(t+1)\right), & A_4,
\end{cases}
$$

$$
\overset{\text{Beweis zu Satz 4}}{=} \begin{cases}
\boldsymbol{\mu}^{(c)} - \left(\frac{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)\left(\boldsymbol{\mu}^{(c)}-\boldsymbol{\mu}^{(c)}\right)}{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)}-(t+1)\right), & A_1,\\
\vdots\\
\boldsymbol{\mu}^{(c)} - \left(\frac{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)\left(\boldsymbol{\mu}^{(c)}-\boldsymbol{\mu}^{(c)}\right)}{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)}-(t+1)\right), & A_2,\\
\vdots\\
\boldsymbol{\mu}^{(c)} - \left(\frac{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)\left(\boldsymbol{\mu}^{(c)}-\boldsymbol{\mu}^{(c)}\right)}{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)}-(t+1)\right), & A_3,\\
\vdots\\
\boldsymbol{\mu}^{(c)} - \left(\frac{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)\left(\boldsymbol{\mu}^{(c)}-\boldsymbol{\mu}^{(c)}\right)}{\sum_{i\in I}\left(z_i^{(c)}-\bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)}-(t+1)\right), & A_4,
\end{cases}
$$

$$
= \begin{cases}
\boldsymbol{\mu}^{(c)}, & A_1,\\
\vdots\\
\boldsymbol{\mu}^{(c)}, & A_2,\\
\vdots\\
\boldsymbol{\mu}^{(c)}, & A_3,\\
\vdots\\
\boldsymbol{\mu}^{(c)}, & A_4.
\end{cases}
$$

Dieser (innere) Erwartungswert ist eine diskrete Zufallsvariable mit $2^t - N_{\text{trend}} - 1$ Ausprägungen. Insgesamt gilt für den gesamten Erwartungswert (8.14) damit:

$$
\begin{aligned}
&\mathrm{E}\left(T_2^{(c),P}\left(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t\right)\right)\\
&= \mathrm{E}\left(\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i\in I}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)}-\bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i\in I}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)}-(t+1)\right)\middle| B_1,\ldots,B_t\right)\right)\\
&= \boldsymbol{\mu}^{(c)}\,\mathrm{P}\left((B_{t-1}=1)\cap(B_t=1)\cap\bigcap_{i=1}^{t-2}(B_i=0)\right) + \ldots + \boldsymbol{\mu}^{(c)}\,\mathrm{P}\left(\bigcap_{i=1}^{t}(B_i=1)\right)\\
&= \boldsymbol{\mu}^{(c)}\prod_{j=t-1}^{t}\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}}=1\right)\prod_{i=1}^{t-2}\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}=0\right) + \ldots + \boldsymbol{\mu}^{(c)}\prod_{i=1}^{t}\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}=1\right)
\end{aligned}
$$

$$= \boldsymbol{\mu}^{(c)} \left(\sum_{l_1=0}^{1} \cdots \sum_{l_{t-N_{\text{trend}}}=0}^{1} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = l_1\right) \cdots \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_{t-N_{\text{trend}}} \sim F_c\}} = l_{t-N_{\text{trend}}}\right) \right)$$
$$\cdot \left(\sum_{i=t-N_{\text{trend}}+1}^{t} \sum_{\substack{j=t-N_{\text{trend}}+1 \\ j \neq i}}^{t} \left(\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1\right) \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 1\right) \right.\right.$$
$$\left.\left. \cdot \sum_{l_{t-N_{\text{trend}}+1}=0}^{1} \cdots \sum_{l_t=0}^{1} \prod_{\substack{k=t-N_{\text{trend}}+1 \\ k \neq i,\, k \neq j}}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}} = l_k\right) \right) \right)$$
$$\overset{\text{Vor. } 7}{=} \boldsymbol{\mu}^{(c)} = \boldsymbol{\mu}_{t+1}^{(c)}.$$

□

Die Schätzfunktion $T_2^{(c),P}$ ist demnach im Falle stabiler Verteilungen in den Klassen unter der Voraussetzung 7 weiterhin erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose des Zeitpunktes $t+1$.

8.2.2 Situation: Linearer Trend der Erwartungswertvektoren

Als Nächstes wird die Situation unter concept drift, im Speziellen unter einem linearen Trend (7.1) der Erwartungswertvektoren (vgl. auch Seite 138), betrachtet.

Da die $\boldsymbol{X}_i^{(c)}$ nun nicht mehr alle identisch wie $\boldsymbol{X}^{(c)}$ verteilt sind, sehen die Erwartungswerte der Zufallsvektoren $\boldsymbol{Y}_i^{(c)}$ und $\bar{\boldsymbol{Y}}_t^{(c)}$ allgemein folgendermaßen aus:

$$\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) = \mathrm{E}\left(T_1^{(c),P}\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i\right)\right) = \mathrm{E}\left(\frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}}} \sum_{j=1}^{i} \left(\boldsymbol{X}_j \cdot \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}}\right)\right)$$
$$\overset{(6.6)}{=} \boldsymbol{\mu}_1^{(c)} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = 1\right) \prod_{j=2}^{i} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 0\right) + \ldots + \tag{8.15}$$
$$+ \left(\frac{1}{i} \sum_{j=1}^{i} \boldsymbol{\mu}_j^{(c)}\right) \prod_{j=1}^{i} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 1\right)$$
$$\overset{(7.1)}{=} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}\right) \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = 1\right) \prod_{j=2}^{i} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 0\right) + \ldots +$$
$$+ \left(\frac{1}{i} \sum_{j=1}^{i} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j\right)\right) \prod_{j=1}^{i} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 1\right),$$

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) = \mathrm{E}\left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{i \in I} \left(\frac{1}{\sum_{k=1}^{i} \mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}}} \sum_{j=1}^{i} \left(\boldsymbol{X}_j \cdot \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}}\right)\right)\right) \tag{8.16}$$
$$= \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i \in I} \left(\left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}\right) \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = 1\right) \prod_{j=2}^{i} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 0\right) \right.$$
$$\left. + \ldots + \left(\frac{1}{i} \sum_{j=1}^{i} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j\right)\right) \prod_{j=1}^{i} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 1\right) \right).$$

Die Aussage über die Erwartungstreue der Schätzfunktion $T_2^{(c),P}$ aus (8.9) wird auch hier zunächst für den Spezialfall $g(\boldsymbol{x}_i) = c$, $i = 1, \ldots, t$, aus Voraussetzung 4 getroffen:

Satz 14. Unter der Annahme eines linearen Trends (6.1)/(7.1) der Erwartungswertvektoren der Klassen ist die bei der Erweiterung der Update-Methode *Sequential ILDA* (Pang et al., 2005b) verwendete Schätzfunktion $T_2^{(c),P}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})$ aus (8.4) bzw. (8.9) zum Zeitpunkt t im Spezialfall aus Voraussetzung 4 (Seite 138) erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c des kommenden Zeitpunktes $t+1$:

$$\mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})\right) = \boldsymbol{\mu}_{t+1}^{(c)}.$$

Beweis. In dem Fall, dass alle Beobachtungen des Zeitraums 1 bis t aus Klasse c kommen (Voraussetzung 4), gilt für die verschobenen Zeitpunkte $z_i^{(c)}$ (vgl. Herleitung in (8.11)):

$$z_i^{(c)} = \frac{i+1}{2}, \quad i = t - n_{\text{trend}}^{(c)} + 1, \ldots, t.$$

Die Erwartungswerte (8.15) und (8.16) der Zufallsvektoren $\boldsymbol{Y}_i^{(c)}$ und $\bar{\boldsymbol{Y}}_t^{(c)}$ vereinfachen sich unter den Annahmen zu:

$$\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) = \mathrm{E}\left(\frac{1}{i}\sum_{j=1}^{i}\boldsymbol{X}_j^{(c)}\right) = \frac{1}{i}\sum_{j=1}^{i}\mathrm{E}\left(\boldsymbol{X}_j^{(c)}\right) = \frac{1}{i}\sum_{j=1}^{i}\boldsymbol{\mu}_j^{(c)} \overset{(7.1)}{=} \frac{1}{i}\sum_{j=1}^{i}\left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j\right),$$

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{1}{i}\sum_{j=1}^{i}\left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j\right)\right).$$

Der Erwartungswert $\mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})\right)$ der Schätzfunktion $T_2^{(c),P}$ aus (8.4) bzw. (8.9) lässt sich durch Einsetzen der obigen Überlegungen folgendermaßen berechnen:

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right)\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)} - (t+1)\right)$$

$$= \overbrace{\frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{1}{i}\sum_{j=1}^{i}\left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j\right)\right)}^{=:a \ \left(=\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right)\right)}$$

$$
- \left(\frac{\overbrace{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\left(\frac{i+1}{2} - \left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{j=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{j+1}{2} \right) \right) \right) \right)}^{=:b \ \left(= \left(z_i^{(c)} - \bar{z}_t^{(c)} \right) \right)}}{\underbrace{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i+1}{2} - \left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{j=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{j+1}{2} \right) \right) \right)^2}_{=:c \ \left(= \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right)^2 \right)}} \right.
$$

$$
\left. \cdot \frac{\overbrace{\left(\left(\frac{1}{i} \sum_{j=1}^{i} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j \right) \right) - \left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{k=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\left(\frac{1}{k} \sum_{j=1}^{k} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j \right) \right) \right) \right) \right)}^{=:d \ \left(= \left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)} \right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} \right) \right) \right)}}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i+1}{2} - \left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{j=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{j+1}{2} \right) \right) \right)^2} \right)
$$

$$
\cdot \underbrace{\left(\left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i+1}{2} \right) \right) - (t+1) \right)}_{=:e \ \left(= \left(\bar{z}_t^{(c)} - (t+1) \right) \right)}.
$$

Dabei sind:

$$
\begin{aligned}
a &= \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{1}{i} \sum_{j=1}^{i} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j \right) \right) \\
&= \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{1}{i} \left(i\, \boldsymbol{\beta}_0^{(c)} + \frac{i(i+1)}{2} \cdot \boldsymbol{\beta}_1^{(c)} \right) \right) \\
&= \frac{1}{n_{\text{trend}}^{(c)}} \left(n_{\text{trend}}^{(c)} \boldsymbol{\beta}_0^{(c)} + \frac{1}{2} \boldsymbol{\beta}_1^{(c)} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} (i+1) \right) \\
&= \boldsymbol{\beta}_0^{(c)} + \frac{1}{2n_{\text{trend}}^{(c)}} \cdot \boldsymbol{\beta}_1^{(c)} \left(\left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) + n_{\text{trend}}^{(c)} \right) \\
&= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} \left(\frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) + \frac{1}{2} \right), \\
b &= \frac{i+1}{2} - \left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{j=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{j+1}{2} \right) \right) \\
&\overset{(8.12)}{=} \frac{i}{2} - \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right),
\end{aligned}
$$

$$
\begin{aligned}
c &= \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i+1}{2} - \left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{j=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{j+1}{2} \right) \right) \right)^2 \\
&\overset{\text{b)}}{=} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i}{2} - \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) \right)^2, \\
d &= \left(\frac{1}{i} \sum_{j=1}^{i} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j \right) \right) - \left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{k=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{1}{k} \sum_{j=1}^{k} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} j \right) \right) \right) \\
&\overset{\text{a)}}{=} \left(\boldsymbol{\beta}_0^{(c)} + \frac{i+1}{2} \cdot \boldsymbol{\beta}_1^{(c)} \right) \\
&\quad - \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} \left(\frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) + \frac{1}{2} \right) \right) \\
&= \boldsymbol{\beta}_1^{(c)} \left(\frac{i}{2} - \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) \right), \\
e &= \left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i+1}{2} \right) \right) - (t+1) \\
&\overset{(8.12)}{=} \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) + \frac{1}{2} - (t+1).
\end{aligned}
$$

Einsetzen von a – e ergibt

$$
\begin{aligned}
&\mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) \right) = \mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)}) \right) \\
&= \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} \right) - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right) \left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)} \right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} \right) \right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1) \right) \\
&= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} \left(\frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) + \frac{1}{2} \right) \\
&\quad - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\left(\frac{i}{2} - \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) \right) \right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i}{2} - \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) \right)^2} \right. \\
&\qquad \left. \frac{\cdot \left(\boldsymbol{\beta}_1^{(c)} \left(\frac{i}{2} - \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) \right) \right) \right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i}{2} - \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) \right)^2} \right) \\
&\quad \cdot \left(\frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) + \frac{1}{2} - (t+1) \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} \left(\frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t - n_{\text{trend}}^{(c)})(t - n_{\text{trend}}^{(c)} + 1)}{2} \right) + \frac{1}{2} \right) \\
&\quad - \left(\frac{\boldsymbol{\beta}_1^{(c)} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i}{2} - \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) \right)^2}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{i}{2} - \frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2} \right) \right)^2} \right) \\
&\qquad \cdot \left(\frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t - n_{\text{trend}}^{(c)})(t - n_{\text{trend}}^{(c)} + 1)}{2} \right) + \frac{1}{2} - (t+1) \right) \\
&= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} \left(\frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t - n_{\text{trend}}^{(c)})(t - n_{\text{trend}}^{(c)} + 1)}{2} \right) + \frac{1}{2} \right) \\
&\quad - \boldsymbol{\beta}_1^{(c)} \left(\frac{1}{2n_{\text{trend}}^{(c)}} \left(\frac{t(t+1)}{2} - \frac{(t - n_{\text{trend}}^{(c)})(t - n_{\text{trend}}^{(c)} + 1)}{2} \right) + \frac{1}{2} - (t+1) \right) \\
&= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(t+1).
\end{aligned}
$$

Wegen (7.1) folgt insgesamt unter den getroffenen Annahmen:

$$
\mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) \right) = \mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)}) \right) = \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(t+1) = \boldsymbol{\mu}_{t+1}^{(c)}.
$$

□

Das heißt unter Annahme eines linearen Trends der Erwartungswertvektoren (7.1) ist die Schätzfunktion $T_2^{(c),P}$ erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose des Zeitpunktes $t+1$, falls nur Beobachtungen mit Klassenlabel c auftreten.

Auch im Falle eines linearen Trends der Erwartungswertvektoren lässt sich die Aussage über die Erwartungstreue der Schätzfunktion unter der praktischen Voraussetzung 7 ohne Beschränkung auf den Spezialfall $g(\boldsymbol{x}_i)$, $i = 1, \ldots, t$, aus Voraussetzung 4 treffen.

Satz 15. Unter der Annahme eines linearen Trends (6.1)/(7.1) der Erwartungswertvektoren der Klassen und der Voraussetzung 7 ist die bei der Erweiterung der Update-Methode *Sequential Incremental LDA* (Pang et al., 2005b) verwendete Schätzfunktion $T_2^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ aus (8.9) zum Zeitpunkt t erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c des kommenden Zeitpunktes $t+1$:

$$
\mathrm{E}\left(T_2^{(c),P}\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t\right) \right) = \boldsymbol{\mu}_{t+1}^{(c)}.
$$

Beweis. Wie im Falle einer stabilen Verteilung in Abschnitt 8.2.1 (Beweis zu Satz 13) wird für den Beweis der bedingte Erwartungswert der Schätzfunktion betrachtet:

$$
\mathrm{E}\left(T_2^{(c),P}\left(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t\right) \right)
$$

$$= \mathrm{E}\left(\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i\in I}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i\in I}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)} - (t+1)\right)\middle|\begin{matrix}\mathbb{1}_{\{\boldsymbol{X}_1\sim F_c\}},\ldots,\\ \mathbb{1}_{\{\boldsymbol{X}_t\sim F_c\}}\end{matrix}\right)\right). \tag{8.17}$$

Für den inneren Erwartungswert werden die $2^t - N_{\text{trend}} - 1$ verschiedenen möglichen Kombinationen der Ausprägungen von $\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}}$ für $i = 1,\ldots,t$ betrachtet (vgl. Voraussetzung 7).

Es sei als Beispiel die Ausprägung des Erwartungswertes für die Kombination $\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}} = 1$ $(i = t-1, t)$ und $\mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}} = 0$ $(j \neq i)$ (Bedingung unten bezeichnet als A) betrachtet:

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-1}^{t}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i=t-1}^{t}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)} - (t+1)\right)\middle|\begin{matrix}\mathbb{1}_{\{\boldsymbol{X}_i\sim F_c\}} = 1\ (i = t-1, t),\\ \mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}} = 0\ (j \neq i)\end{matrix}\right).$$

Die einzelnen Größen sehen unter der Bedingung folgendermaßen aus:

$$\bar{\boldsymbol{Y}}_t^{(c)} \overset{(8.10)}{=} \frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)} + \frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)} + \boldsymbol{X}_t^{(c)}\right)\right), \tag{8.18}$$

$$\sum_{i=t-1}^{t}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2 \overset{(8.7)/(8.8)}{=} \left((t-1) - \frac{1}{2}\left((t-1) + \frac{1}{2}\left((t-1)+t\right)\right)\right)^2 \tag{8.19}$$
$$+ \left(\frac{1}{2}\left((t-1)+t\right) - \frac{1}{2}\left((t-1) + \frac{1}{2}\left((t-1)+t\right)\right)\right)^2,$$

$$\left(\bar{Z}_t^{(c)} - (t+1)\right) \overset{(8.8)}{=} \frac{1}{2}\left((t-1) + \frac{1}{2}\left((t-1)+t\right)\right) - (t+1),$$

$$\sum_{i=t-1}^{t}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right) = \left((t-1) - \frac{1}{2}\left((t-1) + \frac{1}{2}\left((t-1)+t\right)\right)\right) \tag{8.20}$$
$$\cdot\left(\boldsymbol{X}_{t-1}^{(c)} - \frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)} + \frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)} + \boldsymbol{X}_t^{(c)}\right)\right)\right)$$
$$+ \left(\frac{1}{2}\left((t-1)+t\right) - \frac{1}{2}\left((t-1) + \frac{1}{2}\left((t-1)+t\right)\right)\right)$$
$$\cdot\left(\frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)} + \boldsymbol{X}_t^{(c)}\right) - \frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)} + \frac{1}{2}\left(\boldsymbol{X}_{t-1}^{(c)} + \boldsymbol{X}_t^{(c)}\right)\right)\right).$$

Aufgrund der Bedingung lassen sich die additiven und multiplikativen Konstanten aus dem Erwartungswert herausziehen. Nur die Größen $\bar{\boldsymbol{Y}}_t^{(c)}$ und $\boldsymbol{Y}_i^{(c)}$ basieren auf den Zufallsvektoren $\boldsymbol{X}_{t-1}^{(c)}$ und $\boldsymbol{X}_t^{(c)}$. Die Zufallsvariablen $Z_i^{(c)}$ bzw. $\bar{Z}_t^{(c)}$ sind unter Bedingung durch A nicht mehr zufällig. Die betrachtete Ausprägung des bedingten Erwartungswertes lässt sich umformen zu:

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-1}^{t}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i=t-1}^{t}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)} - (t+1)\right)\middle|A\right)$$
$$= \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}|A\right) - \left(\frac{\sum_{i=t-1}^{t}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}|A\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}|A\right)\right)}{\sum_{i=t-1}^{t}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)} - (t+1)\right). \tag{8.21}$$

Hier berechnen sich die einzelnen bedingten Erwartungswerte aus (8.18) und (8.20). Dabei kann auch gezeigt werden, dass $\sum_{i=t-1}^{t}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}|A\right)-\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}|A\right)\right)=\boldsymbol{\beta}_1^{(c)}\sum_{i=t-1}^{t}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)^2$.

Es gilt:

$$
\begin{aligned}
\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}|A\right) &\overset{(8.18)}{=} \frac{1}{2}\left(\mathrm{E}\left(\boldsymbol{X}_{t-1}^{(c)}\right)+\frac{1}{2}\left(\mathrm{E}\left(\boldsymbol{X}_{t-1}^{(c)}\right)+\mathrm{E}\left(\boldsymbol{X}_t^{(c)}\right)\right)\right)=\frac{1}{2}\left(\boldsymbol{\mu}_{t-1}^{(c)}+\frac{1}{2}\left(\boldsymbol{\mu}_{t-1}^{(c)}+\boldsymbol{\mu}_t^{(c)}\right)\right)\\
&\overset{(7.1)}{=} \frac{1}{2}\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}(t-1)+\frac{1}{2}\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}(t-1)+\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}t\right)\right)\\
&= \boldsymbol{\beta}_0^{(c)}+\frac{\boldsymbol{\beta}_1^{(c)}}{2}\left((t-1)+\frac{1}{2}\left((t-1)+t\right)\right),
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{i=t-1}^{t}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}|A\right)-\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}|A\right)\right)\\
&\overset{(8.20)}{=}\left((t-1)-\frac{1}{2}\left((t-1)+\frac{1}{2}\left((t-1)+t\right)\right)\right)\\
&\quad\cdot\left(\mathrm{E}\left(\boldsymbol{X}_{t-1}^{(c)}\right)-\frac{1}{2}\left(\mathrm{E}\left(\boldsymbol{X}_{t-1}^{(c)}\right)+\frac{1}{2}\left(\mathrm{E}\left(\boldsymbol{X}_{t-1}^{(c)}\right)+\mathrm{E}\left(\boldsymbol{X}_t^{(c)}\right)\right)\right)\right)\\
&\quad+\left(\frac{1}{2}\left((t-1)+t\right)-\frac{1}{2}\left((t-1)+\frac{1}{2}\left((t-1)+t\right)\right)\right)\\
&\qquad\cdot\left(\frac{1}{2}\left(\mathrm{E}\left(\boldsymbol{X}_{t-1}^{(c)}\right)+\mathrm{E}\left(\boldsymbol{X}_t^{(c)}\right)\right)-\frac{1}{2}\left(\mathrm{E}\left(\boldsymbol{X}_{t-1}^{(c)}\right)+\frac{1}{2}\left(\mathrm{E}\left(\boldsymbol{X}_{t-1}^{(c)}\right)+\mathrm{E}\left(\boldsymbol{X}_t^{(c)}\right)\right)\right)\right)\\
&\overset{(7.1)}{=}\left((t-1)-\frac{1}{2}\left((t-1)+\frac{1}{2}\left((t-1)+t\right)\right)\right)\\
&\quad\cdot\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}(t-1)-\frac{1}{2}\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}(t-1)+\frac{1}{2}\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}(t-1)+\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}t\right)\right)\right)\\
&\quad+\left(\frac{1}{2}\left((t-1)+t\right)-\frac{1}{2}\left((t-1)+\frac{1}{2}\left((t-1)+t\right)\right)\right)\\
&\qquad\cdot\left(\frac{1}{2}\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}(t-1)+\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}t\right)\right.\\
&\qquad\quad\left.-\frac{1}{2}\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}(t-1)+\frac{1}{2}\left(\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}(t-1)+\boldsymbol{\beta}_0^{(c)}+\boldsymbol{\beta}_1^{(c)}t\right)\right)\right)\\
&=\boldsymbol{\beta}_1^{(c)}\left(\left((t-1)-\frac{1}{2}\left((t-1)+\frac{1}{2}\left((t-1)+t\right)\right)\right)^2\right.\\
&\qquad\left.+\left(\frac{1}{2}\left((t-1)+t\right)-\frac{1}{2}\left((t-1)+\frac{1}{2}\left((t-1)+t\right)\right)\right)^2\right)\\
&\overset{(8.19)}{=}\boldsymbol{\beta}_1^{(c)}\sum_{i=t-1}^{t}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)^2.
\end{aligned}
$$

Damit lässt sich (8.21) weiter vereinfachen:

$$
\begin{aligned}
&\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}-\left(\frac{\sum_{i=t-1}^{t}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)}-\bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i=t-1}^{t}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)}-(t+1)\right)\Big|A\right)\\
&=\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}|A\right)-\left(\frac{\sum_{i=t-1}^{t}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}|A\right)-\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}|A\right)\right)}{\sum_{i=t-1}^{t}\left(Z_i^{(c)}-\bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)}-(t+1)\right)
\end{aligned}
$$

$$= \boldsymbol{\beta}_0^{(c)} + \frac{\boldsymbol{\beta}_1^{(c)}}{2}\left((t-1) + \frac{1}{2}\left((t-1)+t\right)\right) - \boldsymbol{\beta}_1^{(c)}\left(\frac{1}{2}\left((t-1) + \frac{1}{2}\left((t-1)+t\right)\right) - (t+1)\right)$$
$$= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(t+1) \overset{(7.1)}{=} \boldsymbol{\mu}_{t+1}^{(c)}.$$

Diese Umformungen lassen sich für alle möglichen Kombinationen der Ausprägungen von $\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}}$ für $i = 1, \ldots, t$ des inneren Erwartungswertes aus (8.17) durchführen. Damit gilt für diesen:

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i\in I}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i\in I}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)} - (t+1)\right) \middle| \begin{matrix}\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}}, \ldots, \\ \mathbb{1}_{\{\boldsymbol{X}_t \sim F_c\}}\end{matrix}\right)$$
$$= \begin{cases} \boldsymbol{\mu}_{t+1}^{(c)}, & \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1 \ (i = t-1, t), \ \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 0 \ (j \neq i), \\ \vdots & \\ \boldsymbol{\mu}_{t+1}^{(c)}, & \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1 \ (i = t - N_{\text{trend}} + 1, t - N_{\text{trend}} + 2), \ \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 0 \ (j \neq i), \\ \vdots & \\ \boldsymbol{\mu}_{t+1}^{(c)}, & \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1 \ (i = 1, t-1, t), \ \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 0 \ (j \neq i), \\ \vdots & \\ \boldsymbol{\mu}_{t+1}^{(c)}, & \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1 \ (i = 1, \ldots, t). \end{cases}$$

Dieser Erwartungswert ist eine diskrete Zufallsvariable mit $2^t - N_{\text{trend}} - 1$ Ausprägungen. Insgesamt gilt damit für (8.17):

$$\mathrm{E}\left(T_2^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)\right)$$
$$= \mathrm{E}\left(\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i\in I}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i\in I}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)} - (t+1)\right) \middle| \begin{matrix}\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}}, \ldots, \\ \mathbb{1}_{\{\boldsymbol{X}_t \sim F_c\}}\end{matrix}\right)\right)$$
$$= \boldsymbol{\mu}_{t+1}^{(c)} \mathrm{P}\left(\left(\mathbb{1}_{\{\boldsymbol{X}_{t-1} \sim F_c\}} = 1\right) \cap \left(\mathbb{1}_{\{\boldsymbol{X}_t \sim F_c\}} = 1\right) \cap \bigcap_{i=1}^{t-2}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 0\right)\right)$$
$$+ \ldots + \boldsymbol{\mu}_{t+1}^{(c)} \mathrm{P}\left(\bigcap_{i=1}^{t}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1\right)\right)$$
$$= \boldsymbol{\mu}_{t+1}^{(c)} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_{t-1} \sim F_c\}} = 1\right) \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_t \sim F_c\}} = 1\right) \prod_{i=1}^{t-2} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 0\right)$$
$$+ \ldots + \boldsymbol{\mu}_{t+1}^{(c)} \prod_{i=1}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1\right)$$

$$
\begin{aligned}
&\overset{\text{Vor. } 7}{=} \boldsymbol{\mu}_{t+1}^{(c)} \left(\sum_{l_1=0}^{1} \cdots \sum_{l_{t-N_{\text{trend}}}=0}^{1} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_1 \sim F_c\}} = l_1\right) \cdots \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_{t-N_{\text{trend}}} \sim F_c\}} = l_{t-N_{\text{trend}}}\right) \right) \\
&\quad \cdot \left(\sum_{i=t-N_{\text{trend}}+1}^{t} \sum_{\substack{j=t-N_{\text{trend}}+1 \\ j \neq i}}^{t} \left(\mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} = 1\right) \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}} = 1\right) \right. \right. \\
&\qquad \left. \left. \cdot \sum_{l_{t-N_{\text{trend}}+1}=0}^{1} \cdots \sum_{l_t=0}^{1} \prod_{\substack{k=t-N_{\text{trend}}+1 \\ k \neq i,\, k \neq j}}^{t} \mathrm{P}\left(\mathbb{1}_{\{\boldsymbol{X}_k \sim F_c\}} = l_k\right) \right) \right) \\
&\overset{\text{Vor. } 7}{=} \boldsymbol{\mu}_{t+1}^{(c)}.
\end{aligned}
$$

□

Die Schätzfunktion $T_2^{(c),P}$ ist demnach auch im Falle eines linearen Trends (6.1)/(7.1) der Erwartungswertvektoren in den Klassen unter der Voraussetzung 7 weiterhin erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose des Zeitpunktes $t+1$.

8.3 Online Linear Discriminant Classifier (OLDC)

Eine Update-Methode für die Kanonische Lineare Diskriminanzanalyse wurde von Kuncheva und Plumpton (2008) eingeführt. Diese Methode *OLDC* (Online Linear Discriminant Classifier) mit fester und adaptiver Lernrate wurde in Abschnitt 4.3 vorgestellt.

In Abschnitt 6.3 wurde für spezielle Situationen herausgestellt, dass die ursprüngliche Schätzfunktion $T_1^{(c),K}$ aus (6.10) erwartungstreu für den Erwartungswertvektor der Prognose ist, falls kein concept drift vorliegt und demnach eine stabile Verteilung über die Zeit unterstellt wird. Es zeigte sich jedoch, dass diese Erwartungstreue der Schätzfunktion nicht mehr gilt, falls ein linearer Trend der Erwartungswertvektoren in den Klassen vorliegt.

Im Folgenden wird unter der Voraussetzung 4 (Seite 138) entgegen Abschnitt 8.1 ausführlicher gezeigt, dass die Schätzfunktion der erweiterten Methode auch bei einem linearen Trend der Erwartungswertvektoren erwartungstreu für den Erwartungswertvektor der Prognose ist und die Erwartungstreue auch im Falle stabiler Verteilungen weiterhin gilt.

Für die Untersuchung des alternativ eingeführten Schätzers (vgl. (8.1))

$$
\begin{aligned}
\hat{\boldsymbol{m}}_{n_{t+1}^{(c)}}^{(c)} =: \hat{\boldsymbol{y}}_{t+1}^{(c)} &= \hat{\boldsymbol{\beta}}_{0t}^{(c)} + \hat{\boldsymbol{\beta}}_{1t}^{(c)}(t+1) \\
&= \bar{\boldsymbol{y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\boldsymbol{y}_i^{(c)} - \bar{\boldsymbol{y}}_t^{(c)}\right)}{\sum_{i \in I} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1)\right)
\end{aligned}
$$

unter einer stabilen Verteilung in Abschnitt 8.3.1 und bei Unterstellung eines linearen Trends der Erwartungswertvektoren in den Klassen in Abschnitt 8.3.2 wird dazu wie in Abschnitt 6.3 der Spezialfall $g(\boldsymbol{x}_i) = c,\ i = 1, \ldots, t$, aus Voraussetzung 4 betrachtet.

Die letzten $n_{\text{trend}}^{(c)}$ geschätzten Erwartungswertvektoren der Klasse c aus $OLDC$ sehen unter diesem Spezialfall in iterativer Form für $i \in I,\ i \geq 3$, folgendermaßen aus (vgl. (6.9)):

$$\begin{aligned}\boldsymbol{y}_i^{(c)} &:= \boldsymbol{m}_{n_i^{(c)}}^{(c)} \\ &= \frac{1}{(1-\lambda)(i-1)+\lambda} \cdot \\ &\left(\frac{(1-\lambda)^{i-1} \prod_{j=1}^{i-1} j}{\prod_{j=3}^{i} ((1-\lambda)(j-2)+\lambda)} \cdot \boldsymbol{x}_1 + \sum_{j=2}^{i-1} \frac{(1-\lambda)^{i-j} \lambda \prod_{k=j}^{i-1} k}{\prod_{k=j+1}^{i} ((1-\lambda)(k-2)+\lambda)} \cdot \boldsymbol{x}_j + \lambda \boldsymbol{x}_i \right).\end{aligned}$$

In dem Fall $g(\boldsymbol{x}_i) = c,\ i = 1, \ldots, t$, sind die verschobenen Zeitpunkte $z_i^{(c)}$ fest (nichtstochastische Regressoren) und lassen sich analog folgendermaßen formulieren (vgl. (7.13)):

$$\begin{aligned}z_i^{(c)} &= \frac{1}{(1-\lambda)(i-1)+\lambda} \cdot \\ &\left(\frac{(1-\lambda)^{i-1} \prod_{j=1}^{i-1} j}{\prod_{j=3}^{i} ((1-\lambda)(j-2)+\lambda)} \cdot 1 + \sum_{j=2}^{i-1} \frac{(1-\lambda)^{i-j} \lambda \prod_{k=j}^{i-1} k}{\prod_{k=j+1}^{i} ((1-\lambda)(k-2)+\lambda)} \cdot j + \lambda i \right) \\ &= \frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1} ((1-\lambda)j+\lambda)} + \sum_{j=2}^{i-1} \frac{\frac{(i-1)!}{(j-1)!} \cdot (1-\lambda)^{i-j} \lambda j}{\prod_{k=j-1}^{i-1} ((1-\lambda)k+\lambda)} + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda}. \qquad (8.22)\end{aligned}$$

Ebenso ist folglich der Mittelwert $\bar{z}_t^{(c)}$ zum Zeitpunkt t über alle $n_{\text{trend}}^{(c)}$ Zeitpunkte (8.22) des betrachteten Intervalls für das lokale lineare Regressionsmodell I (vgl. (8.2)) fest:

$$\begin{aligned}\bar{z}_t^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} &\left(\frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1} ((1-\lambda)j+\lambda)} + \sum_{j=2}^{i-1} \frac{\frac{(i-1)!}{(j-1)!} \cdot (1-\lambda)^{i-j} \lambda j}{\prod_{k=j-1}^{i-1} ((1-\lambda)k+\lambda)} \right. \\ &\left. + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda} \right). \qquad (8.23)\end{aligned}$$

Die erweiterte Schätzfunktion für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ des folgenden Zeitpunktes $t+1$ lässt sich im Spezialfall daher folgendermaßen formulieren:

$$\begin{aligned}&T_2^{(c),K}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) = T_2^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)}) \qquad (8.24) \\ &= \bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1) \right),\end{aligned}$$

wobei hier $\boldsymbol{Y}_i^{(c)}$ der Schätzfunktion $T_1^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_i^{(c)})$ (vgl. (6.10)) der ursprünglichen Methode entspricht und damit:

$$\boldsymbol{Y}_i^{(c)} = \frac{1}{(1-\lambda)(i-1)+\lambda} \cdot \left(\frac{(1-\lambda)^{i-1} \prod_{j=1}^{i-1} j}{\prod_{j=3}^{i} ((1-\lambda)(j-2)+\lambda)} \cdot \boldsymbol{X}_1^{(c)} + \sum_{j=2}^{i-1} \frac{(1-\lambda)^{i-j} \lambda \prod_{k=j}^{i-1} k}{\prod_{k=j+1}^{i} ((1-\lambda)(k-2)+\lambda)} \cdot \boldsymbol{X}_j^{(c)} + \lambda \boldsymbol{X}_i^{(c)} \right),$$

$$\bar{\boldsymbol{Y}}_t^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \boldsymbol{Y}_i^{(c)}.$$

8.3.1 Situation: Stabile Verteilung

Im Falle einer stabilen Verteilung in Klasse c über die Zeit gilt (vgl. Seite 138): $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_1^{(c)} = \boldsymbol{\mu}_2^{(c)} = \ldots$.

Der folgende Satz trifft eine Aussage über die Erwartungstreue der Schätzfunktion $T_2^{(c),K}$ unter bestimmten Voraussetzungen.

Satz 16. Unter der Annahme einer stabilen Verteilung und dem Spezialfall aus Voraussetzung 4 (Seite 138) ist die bei der Erweiterung der Update-Methode *OLDC* (Kuncheva und Plumpton, 2008) verwendete Schätzfunktion $T_2^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})$ aus (8.4) bzw. (8.24) zum Zeitpunkt t erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c des kommenden Zeitpunktes $t+1$ für alle Zeitpunkte $t \geq 3$ im Datenstrom:

$$\mathrm{E}\left(T_2^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})\right) = \boldsymbol{\mu}^{(c)}.$$

Beweis. Da $\boldsymbol{Y}_i^{(c)}$ der ursprünglichen Schätzfunktion $T_1^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_i^{(c)})$ zum Zeitpunkt i entspricht, ergeben sich die Erwartungswerte von $\boldsymbol{Y}_i^{(c)}$ und $\bar{\boldsymbol{Y}}_t^{(c)}$ in dieser Situation stabiler Verteilungen und im Spezialfall aus Voraussetzung 4 (Seite 138) durch

$$\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) = \mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i)\right) = \mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_i^{(c)})\right) \overset{(6.12)}{=} \boldsymbol{\mu}^{(c)} \text{ für } i \geq 3,$$

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) = \frac{n_{\text{trend}}^{(c)} \boldsymbol{\mu}^{(c)}}{n_{\text{trend}}^{(c)}} = \boldsymbol{\mu}^{(c)} \text{ für } t - n_{\text{trend}}^{(c)} + 1 \geq 3.$$

Wie bei der Schätzfunktion $T_2^{(c),P}$ der erweiterten Methode von Pang et al. (2005b) (vgl. (8.13) in Abschnitt 8.2.1) vereinfacht sich der Erwartungswert der Schätzfunktion $T_2^{(c),K}$ durch die Annahme und Einsetzen der obigen Überlegungen zu:

$$\mathrm{E}\left(T_2^{(c),K}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)\right) = \mathrm{E}\left(T_2^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})\right)$$

$$= \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right)\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)} - (t+1)\right)$$

$$= \boldsymbol{\mu}^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\boldsymbol{\mu}^{(c)} - \boldsymbol{\mu}^{(c)}\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)} - (t+1)\right) = \boldsymbol{\mu}^{(c)} = \boldsymbol{\mu}_{t+1}^{(c)}.$$

□

Das heißt die Schätzfunktion $T_2^{(c),K}$ ist weiterhin erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose von Klasse c im Falle einer stabilen Verteilung in den Klassen, falls alle bisherigen Beobachtungen Klassenlabel c besitzen.

8.3.2 Situation: Linearer Trend der Erwartungswertvektoren

Für die Untersuchung der Erwartungstreue der erweiterten Schätzfunktion (8.24) unter der Annahme eines linearen Trends (7.1) der Erwartungswertvektoren der Klassen wird ebenfalls der Spezialfall $g(\boldsymbol{x}_i) = c,\ i = 1, \ldots, t$, aus Voraussetzung 4 (Seite 138) betrachtet.

Unter der Voraussetzung 4 kann gezeigt werden, dass die Schätzfunktion $T_2^{(c),K}$ zum Zeitpunkt t auch in dieser Situation von concept drift immer noch erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ aus Klasse c des folgenden Zeitpunktes $t+1$ ist.

Satz 17. Unter der Annahme eines linearen Trends (6.1)/(7.1) der Erwartungswertvektoren der Klassen ist die bei der Erweiterung der Update-Methode *OLDC* (Kuncheva und Plumpton, 2008) verwendete Schätzfunktion $T_2^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})$ aus (8.4) bzw. (8.24) zum Zeitpunkt t im Spezialfall aus Voraussetzung 4 (Seite 138) erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c des kommenden Zeitpunktes $t+1$:

$$\mathrm{E}\left(T_2^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})\right) = \boldsymbol{\mu}_{t+1}^{(c)}.$$

Beweis. In der Schätzfunktion $T_2^{(c),K}$ (vgl. (8.24)) wird die Differenz aus einzelnen verschobenen Zeitpunkten (8.22) und dem Mittelwert (8.23) betrachtet:

$$\begin{aligned} & z_i^{(c)} - \bar{z}_t^{(c)} \\ &= \frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1}\left((1-\lambda)j + \lambda\right)} + \sum_{j=2}^{i-1}\frac{\frac{(i-1)!}{(j-1)!}\cdot(1-\lambda)^{i-j}\lambda j}{\prod_{k=j-1}^{i-1}\left((1-\lambda)k+\lambda\right)} + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda} \\ &\quad - \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1}\left((1-\lambda)j+\lambda\right)} + \sum_{j=2}^{i-1}\frac{\frac{(i-1)!}{(j-1)!}\cdot(1-\lambda)^{i-j}\lambda j}{\prod_{k=j-1}^{i-1}\left((1-\lambda)k+\lambda\right)} + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda}\right). \end{aligned} \tag{8.25}$$

Zur Vereinfachung des Erwartungswertes der Schätzfunktion $T_2^{(c),K}$ aus (8.24) wird gezeigt, dass die Differenz der Erwartungswerte $\mathrm{E}\big(\boldsymbol{Y}_i^{(c)}\big) - \mathrm{E}\big(\bar{\boldsymbol{Y}}_t^{(c)}\big)$, welche eine Komponente in der Schätzfunktion $T_2^{(c),K}$ ist, gleich $\boldsymbol{\beta}_1^{(c)}\big(z_i^{(c)} - \bar{z}_t^{(c)}\big)$ ist. Dazu werden zunächst die Erwartungswerte von $\boldsymbol{Y}_i^{(c)}$ und $\bar{\boldsymbol{Y}}_t^{(c)}$ bestimmt:

$$\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) = \mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_i)\right) = \mathrm{E}\left(T_1^{(c),K}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_i^{(c)})\right)$$

$$\overset{(6.13)}{=} \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}\left(\frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{i-1}\frac{\frac{(i-1)!}{(j-1)!}\cdot(1-\lambda)^{i-j}\lambda j}{\prod_{k=j-1}^{i-1}((1-\lambda)k+\lambda)} + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda}\right),$$

$$\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) = \mathrm{E}\left(\frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\boldsymbol{Y}_i^{(c)}\right) = \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right)$$

$$= \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}\left(\frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{i-1}\frac{\frac{(i-1)!}{(j-1)!}\cdot(1-\lambda)^{i-j}\lambda j}{\prod_{k=j-1}^{i-1}((1-\lambda)k+\lambda)} + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda}\right)\right)$$

$$= \boldsymbol{\beta}_0^{(c)} + \frac{\boldsymbol{\beta}_1^{(c)}}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{i-1}\frac{\frac{(i-1)!}{(j-1)!}\cdot(1-\lambda)^{i-j}\lambda j}{\prod_{k=j-1}^{i-1}((1-\lambda)k+\lambda)} + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda}\right).$$

Mit diesen Erwartungswerten lässt sich die Differenz bestimmen:

$$\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right)$$

$$= \boldsymbol{\beta}_1^{(c)}\left(\frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{i-1}\frac{\frac{(i-1)!}{(j-1)!}\cdot(1-\lambda)^{i-j}\lambda j}{\prod_{k=j-1}^{i-1}((1-\lambda)k+\lambda)} + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda}\right.$$

$$\left. - \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{i-1}\frac{\frac{(i-1)!}{(j-1)!}\cdot(1-\lambda)^{i-j}\lambda j}{\prod_{k=j-1}^{i-1}((1-\lambda)k+\lambda)} + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda}\right)\right)$$

$$\overset{(8.25)}{=} \boldsymbol{\beta}_1^{(c)}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right).$$

Das Produkt aus beiden Differenzen lässt sich somit vereinfachen zu:

$$\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right)\right) = \boldsymbol{\beta}_1^{(c)}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2. \tag{8.26}$$

Für den Erwartungswert der Schätzfunktion $T_2^{(c),K}$ aus (8.4) bzw. (8.24) gilt unter den Voraussetzungen im Gesamten:

$$\begin{aligned}
&\mathrm{E}\left(T_2^{(c),K}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t)\right) = \mathrm{E}\left(T_2^{(c),K}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right)\\
&\quad= \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right)\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)} - (t+1)\right)\\
&\overset{(8.23)/(8.26)}{=} \boldsymbol{\beta}_0^{(c)} + \frac{\boldsymbol{\beta}_1^{(c)}}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{i-1}\frac{\frac{(i-1)!}{(j-1)!}\cdot(1-\lambda)^{i-j}\lambda j}{\prod_{k=j-1}^{i-1}((1-\lambda)k+\lambda)}\right.\\
&\qquad\left. + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda}\right)\\
&\qquad - \boldsymbol{\beta}_1^{(c)}\left(\frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{(1-\lambda)^{i-1}(i-1)!}{\prod_{j=1}^{i-1}((1-\lambda)j+\lambda)} + \sum_{j=2}^{i-1}\frac{\frac{(i-1)!}{(j-1)!}\cdot(1-\lambda)^{i-j}\lambda j}{\prod_{k=j-1}^{i-1}((1-\lambda)k+\lambda)}\right.\right.\\
&\qquad\left.\left. + \frac{\lambda i}{(1-\lambda)(i-1)+\lambda}\right) - (t+1)\right)\\
&\quad= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(t+1) \overset{(7.1)}{=} \boldsymbol{\mu}_{t+1}^{(c)}.
\end{aligned}$$

□

Das heißt unter Annahme eines linearen Trends der Erwartungswertvektoren (7.1) ist die Schätzfunktion $T_2^{(c),K}$ erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose des Zeitpunktes $t+1$, falls nur Beobachtungen mit Klassenlabel c auftreten.

8.4 Online Diskriminanzanalyse mit adaptivem Vergessen

Die Update-Methode für Online Diskriminanzanalyse von Anagnostopoulos et al. (2012), welche auf der Idee von adaptivem exponentiellem Vergessen beruht, wurde in Abschnitt 4.4 erläutert.

In Abschnitt 6.4 wurde für spezielle Situationen herausgestellt, dass die ursprüngliche Schätzfunktion $T_1^{(c),A}$ der Methode erwartungstreu für den Erwartungswertvektor der Prognose ist, falls kein concept drift vorliegt und demnach eine stabile Verteilung über die Zeit unterstellt wird. Allerdings zeigte sich, dass diese Erwartungstreue der Schätzfunktion unabhängig von Faktoren bzw. Gewichten in keiner Situation erfüllt werden kann, falls ein linearer Trend der Erwartungswertvektoren in den Klassen vorliegt.

In den folgenden beiden Abschnitten soll vergleichend dazu der Erwartungswert der Schätzfunktion der erweiterten Methode für beide unterstellten Situationen entgegen Abschnitt 8.1 ausführlicher untersucht werden. Es wird dabei jeweils für die Analyse der Schätzfunktion unter einer stabilen Verteilung in Abschnitt 8.4.1 sowie bei Unterstellung eines linearen Trends der Erwartungswertvektoren in den Klassen in Abschnitt 8.4.2 wieder der Spezialfall $g(\boldsymbol{x}_i) = c,\ i = 1, \ldots, t$, aus Voraussetzung 4 (Seite 138) betrachtet.

Unter der Voraussetzung 4 sind auch hier die verschobenen Zeitpunkte nicht mehr zufällig (nicht-stochastische Regressoren im linearen Regressionsmodell). Die Formel (7.14) für die verschobenen Zeitpunkte kann folgendermaßen umgeschrieben werden:

$$\begin{aligned} z_i^{(c)} &= \frac{1}{N_i^{(c)}} \sum_{j=1}^{i} v_j^{(c)} j = \frac{1}{\sum_{l=1}^{i-1} v_l^{(c)} + 1} \sum_{j=1}^{i} v_j^{(c)} j \\ &= \frac{1}{\sum_{l=1}^{i-1} \left(\prod_{k=l}^{i-1} \lambda_{(k)}^{(c)} \right) + 1} \left(\sum_{j=1}^{i-1} \left(v_j^{(c)} j \right) + \underbrace{v_i^{(c)}}_{:=1} i \right) \\ &= \frac{1}{\sum_{l=1}^{i-1} \left(\prod_{k=l}^{i-1} \lambda_{(k)}^{(c)} \right) + 1} \left(\sum_{j=1}^{i-1} \left(\left(\prod_{k=j}^{i-1} \lambda_{(k)}^{(c)} \right) j \right) + i \right). \end{aligned} \tag{8.27}$$

Der Mittelwert über alle $n_{\text{trend}}^{(c)}$ Zeitpunkte $z_i^{(c)}$ im Intervall I (vgl. (8.2), im Spezialfall identisch zu $I = \{t - n_{\text{trend}}^{(c)} + 1, \ldots, t\}$) ergibt sich durch

$$\bar{z}_t^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{1}{\sum_{l=1}^{i-1} \left(\prod_{k=l}^{i-1} \lambda_{(k)}^{(c)} \right) + 1} \left(\sum_{j=1}^{i-1} \left(\left(\prod_{k=j}^{i-1} \lambda_{(k)}^{(c)} \right) j \right) + i \right) \right). \tag{8.28}$$

Mit

$$\begin{aligned} \boldsymbol{Y}_i^{(c)} &\overset{(6.14)}{=} T_1^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_i^{(c)}) \\ &= \frac{1}{\sum_{l=1}^{i-1} \left(\prod_{k=l}^{i-1} \lambda_{(k)}^{(c)} \right) + 1} \left(\sum_{j=1}^{i-1} \left(\left(\prod_{k=j}^{i-1} \lambda_{(k)}^{(c)} \right) \boldsymbol{X}_j^{(c)} \right) + \boldsymbol{X}_i^{(c)} \right), \\ \bar{\boldsymbol{Y}}_t^{(c)} &= \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \boldsymbol{Y}_i^{(c)} \end{aligned}$$

lässt sich die erweiterte Schätzfunktion für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ des Zeitpunktes $t+1$ hier analog wie bei den anderen Methoden formulieren:

$$\begin{aligned} &T_2^{(c),A}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t) = T_2^{(c),A}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)}) \\ &= \bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1)\right). \end{aligned} \tag{8.29}$$

8.4.1 Situation: Stabile Verteilung

Im Falle einer stabilen Verteilung in Klasse c über die Zeit gilt (vgl. Seite 138): $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_1^{(c)} = \boldsymbol{\mu}_2^{(c)} = \ldots$.

Satz 18. Unter der Annahme einer stabilen Verteilung und dem Spezialfall aus Voraussetzung 4 (Seite 138) ist die bei der Erweiterung der Update-Methode *Online Diskriminanzanalyse mit adaptivem Vergessen* (Anagnostopoulos et al., 2012) verwendete Schätzfunktion $T_2^{(c),A}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})$ aus (8.4) bzw. (8.29) zum Zeitpunkt t erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c des kommenden Zeitpunktes $t+1$:

$$\mathrm{E}\left(T_2^{(c),A}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right) = \boldsymbol{\mu}^{(c)}.$$

Beweis. Unter den Voraussetzungen und Annahmen gilt für die Erwartungswerte von $\boldsymbol{Y}_i^{(c)}$ bzw. der ursprünglichen Schätzfunktion $T_1^{(c),A}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_i)$ des Zeitpunktes i und des Mittelwertes $\bar{\boldsymbol{Y}}^{(c)}$:

$$\begin{aligned} &\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) = \mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_i)\right) = \mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_i^{(c)})\right) \overset{(6.15)}{=} \boldsymbol{\mu}^{(c)}, \\ &\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) = \frac{n_{\text{trend}}^{(c)}\,\boldsymbol{\mu}^{(c)}}{n_{\text{trend}}^{(c)}} = \boldsymbol{\mu}^{(c)}. \end{aligned}$$

Da sowohl $\boldsymbol{Y}_i^{(c)}$ als auch $\bar{\boldsymbol{Y}}_t^{(c)}$ erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose sind, lässt sich die Erwartungstreue für $\boldsymbol{\mu}_{t+1}^{(c)}$ der gesamten Schätzfunktion ableiten:

$$\begin{aligned} &\mathrm{E}\left(T_2^{(c),A}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right) \\ &= \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1)\right) \right) \\ &= \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right)\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1)\right) \end{aligned}$$

$$= \boldsymbol{\mu}^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right) \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{\mu}^{(c)} \right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1) \right) = \boldsymbol{\mu}^{(c)} = \boldsymbol{\mu}_{t+1}^{(c)}.$$

□

8.4.2 Situation: Linearer Trend der Erwartungswertvektoren

Auch für die Situation unter concept drift bzw. bei Unterstellung eines linearen Trends der Erwartungswertvektoren in den Klassen (7.1) wird der Spezialfall aus Voraussetzung 4 betrachtet, dass bis zum Zeitpunkt t nur Beobachtungen in Klasse c realisiert werden.

Unter diesen Voraussetzungen kann gezeigt werden, dass die Schätzfunktion $T_2^{(c),A}$ zum Zeitpunkt t immer noch erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ aus Klasse c des folgenden Zeitpunktes $t+1$ ist. Dies drückt der folgende Satz aus.

Satz 19. Unter der Annahme eines linearen Trends (6.1)/(7.1) der Erwartungswertvektoren der Klassen ist die bei der Erweiterung der Update-Methode *Online Diskriminanzalyse mit adaptivem Vergessen* (Anagnostopoulos et al., 2012) verwendete Schätzfunktion $T_2^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)})$ aus (8.4) bzw. (8.29) zum Zeitpunkt t im Spezialfall aus Voraussetzung 4 (Seite 138) erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ von Klasse c des kommenden Zeitpunktes $t+1$:

$$\mathrm{E}\left(T_2^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)}) \right) = \boldsymbol{\mu}_{t+1}^{(c)}.$$

Beweis. Zunächst werden einige Vorüberlegungen bezüglich der einzelnen Komponenten der Schätzfunktion (8.29) unter den Voraussetzungen getroffen.

Die $\boldsymbol{X}_i^{(c)}$ sind bei Unterstellung eines linearen Trends nicht alle identisch wie $\boldsymbol{X}^{(c)}$ verteilt. Daher sehen die Erwartungswerte der Zufallsvektoren $\boldsymbol{Y}_i^{(c)}$ und $\bar{\boldsymbol{Y}}_t^{(c)}$ folgendermaßen aus:

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{Y}_i^{(c)} \right) &= \mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i) \right) = \mathrm{E}\left(T_1^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_i^{(c)} \right) \\
&\overset{(6.16)}{=} \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} \cdot \frac{1}{\sum_{l=1}^{i-1} \left(\prod_{k=l}^{i-1} \lambda_{(k)}^{(c)} \right) + 1} \left(\sum_{j=1}^{i-1} \left(\left(\prod_{k=j}^{i-1} \lambda_{(k)}^{(c)} \right) j \right) + i \right), \\
\mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)} \right) &= \mathrm{E}\left(\frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \boldsymbol{Y}_i^{(c)} \right) = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \mathrm{E}\left(\boldsymbol{Y}_i^{(c)} \right) \\
&= \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} \cdot \frac{\sum_{j=1}^{i-1} \left(\left(\prod_{k=j}^{i-1} \lambda_{(k)}^{(c)} \right) j \right) + i}{\sum_{l=1}^{i-1} \left(\prod_{k=l}^{i-1} \lambda_{(k)}^{(c)} \right) + 1} \right) \\
&= \boldsymbol{\beta}_0^{(c)} + \frac{\boldsymbol{\beta}_1^{(c)}}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{\sum_{j=1}^{i-1} \left(\left(\prod_{k=j}^{i-1} \lambda_{(k)}^{(c)} \right) j \right) + i}{\sum_{l=1}^{i-1} \left(\prod_{k=l}^{i-1} \lambda_{(k)}^{(c)} \right) + 1} \right).
\end{aligned}$$

In der Schätzfunktion wird die Differenz aus $z_i^{(c)}$ und $\bar{z}_t^{(c)}$ betrachtet (vgl. (8.27)/(8.28)):

$$z_i^{(c)} - \bar{z}_t^{(c)} = \frac{\sum\limits_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum\limits_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1} - \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{\sum\limits_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum\limits_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1}\right).$$

Ebenso basiert die Schätzfunktion auf der Differenz von dem Erwartungswert von $\boldsymbol{Y}_i^{(c)}$ und jenem des entsprechenden Mittelwertes $\bar{\boldsymbol{Y}}_t^{(c)}$ über alle Zeitpunkte $t-n_{\text{trend}}^{(c)}+1,\ldots,t$:

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) &= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} \cdot \frac{\sum_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1} \\
&\quad - \boldsymbol{\beta}_0^{(c)} - \frac{\boldsymbol{\beta}_1^{(c)}}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{\sum_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1}\right) \\
&= \boldsymbol{\beta}_1^{(c)}\left(\frac{\sum_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1}\right. \\
&\quad \left. - \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{\sum_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1}\right)\right).
\end{aligned}$$

Das Produkt aus beiden Differenzen lässt sich folgendermaßen vereinfachen:

$$\begin{aligned}
&\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right)\right) \\
&= \boldsymbol{\beta}_1^{(c)}\left(\frac{\sum\limits_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum\limits_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1} - \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{\sum\limits_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum\limits_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1}\right)\right)^2 \\
&= \boldsymbol{\beta}_1^{(c)}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2. \qquad (8.30)
\end{aligned}$$

Aus diesen Vorüberlegungen gilt nun für den Erwartungswert der Schätzfunktion (8.29):

$$\begin{aligned}
&\mathrm{E}\left(T_2^{(c),A}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t)\right) = \mathrm{E}\left(T_2^{(c),A}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)})\right) \\
&\quad = \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right) - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\mathrm{E}\left(\boldsymbol{Y}_i^{(c)}\right) - \mathrm{E}\left(\bar{\boldsymbol{Y}}_t^{(c)}\right)\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)} - (t+1)\right) \\
&\overset{(8.28)/(8.30)}{=} \boldsymbol{\beta}_0^{(c)} + \frac{\boldsymbol{\beta}_1^{(c)}}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{\sum_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1}\right) \\
&\qquad - \boldsymbol{\beta}_1^{(c)}\left(\frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{\sum_{j=1}^{i-1}\left(\left(\prod_{k=j}^{i-1}\lambda_{(k)}^{(c)}\right)j\right)+i}{\sum_{l=1}^{i-1}\left(\prod_{k=l}^{i-1}\lambda_{(k)}^{(c)}\right)+1}\right) - (t+1)\right)
\end{aligned}$$

$$= \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)}(t+1) \overset{(7.1)}{=} \boldsymbol{\mu}_{t+1}^{(c)}.$$

□

Die erweiterte Schätzfunktion $T_2^{(c),A}$ zum Zeitpunkt t ist demnach im betrachteten Spezialfall weiterhin erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Klasse c zum Zeitpunkt $t+1$, falls ein linearer Trend der Erwartungswertvektoren (7.1) vorliegt.

8.5 Zusammenfassung

Die Resultate der vorangegangenen vier Abschnitte 8.1–8.4 sowie Kapitel 6, also die Ergebnisse der Untersuchung der Erwartungstreue der verschiedenen nicht-erweiterten und erweiterten Schätzfunktionen für $\boldsymbol{\mu}_{t+1}^{(c)}$, sind in Tabelle 8.1 zusammengefasst.

In Kapitel 6 wurde die Erwartungstreue der Schätzfunktionen der nicht-erweiterten Methoden für Online Diskriminanzanalyse unter bestimmten Voraussetzungen untersucht. Die entsprechenden betrachteten drei Schätzfunktionen sind im Folgenden zum Vergleich noch einmal aufgeführt:

$$T_1^{(c),P}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) \overset{(6.3)}{=} \bar{\boldsymbol{X}}_t^{(c)} = \frac{1}{\sum_{j=1}^{t} \mathbb{1}_{\{\boldsymbol{X}_j \sim F_c\}}} \sum_{i=1}^{t} \left(\boldsymbol{X}_i \cdot \mathbb{1}_{\{\boldsymbol{X}_i \sim F_c\}} \right), \tag{8.31}$$

$$\begin{aligned} &T_1^{(c),K}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)}) \\ &\overset{(6.10)}{=} \frac{1}{(1-\lambda)(t-1)+\lambda} \cdot \\ &\left(\frac{(1-\lambda)^{t-1} \prod_{j=1}^{t-1} j}{\prod_{j=3}^{t} ((1-\lambda)(j-2)+\lambda)} \cdot \boldsymbol{X}_1^{(c)} + \sum_{j=2}^{t-1} \frac{(1-\lambda)^{t-j} \lambda \prod_{k=j}^{t-1} k}{\prod_{k=j+1}^{t} ((1-\lambda)(k-2)+\lambda)} \cdot \boldsymbol{X}_j^{(c)} + \lambda \boldsymbol{X}_t^{(c)} \right), \end{aligned} \tag{8.32}$$

$$T_1^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)}) \overset{(6.14)}{=} \frac{1}{\sum_{k=1}^{t-1} \left(\prod_{j=k}^{t-1} \lambda_{(j)}^{(c)} \right) + 1} \left(\sum_{i=1}^{t-1} \left(\prod_{j=i}^{t-1} \lambda_{(j)}^{(c)} \right) \boldsymbol{X}_i^{(c)} + \boldsymbol{X}_t^{(c)} \right). \tag{8.33}$$

Es stellte sich heraus, dass alle ursprünglichen Schätzfunktionen $T_1^{(c)}$ der drei betrachteten Methoden zum Zeitpunkt t im Datenstrom unter bestimmten Voraussetzungen erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ des kommenden Zeitpunktes sind, falls kein concept drift vorliegt und demnach eine stabile Verteilung über die Zeit unterstellt wird. Für $T_1^{(c),P}$ konnte die Erwartungstreue dabei im allgemeinen Fall bewiesen werden (vgl. Abschnitt 6.2.1). Aufgrund der recht komplexen Form der iterativen Variante des Schätzwertes für den Erwartungswertvektor und demnach der Schätzfunktion bei *OLDC* (vgl. Abschnitt 6.3.1) und der *Online Diskriminanzanalyse mit adaptivem Vergessen* (vgl. Abschnitt 6.4.1) wurde der Beweis der beiden letzteren Methoden auf den Spezialfall beschränkt, dass bis zum Zeitpunkt t im Datenstrom nur Beobachtungen in Klasse c realisiert

Tabelle 8.1: Zusammenfassung der Situationen, für die Erwartungstreue der verschiedenen Schätzfunktionen $T(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t)$ bewiesen (✓) oder widerlegt (x) wurde. Der Fall „allgemein“ schließt den vorangegangenen Spezialfall mit ein. In eckigen Klammern: Abschnitt des Beweises.

	Schätzfunktion					
Situation	$T_1^{(c),P}$	$T_2^{(c),P}$	$T_1^{(c),K}$	$T_2^{(c),K}$	$T_1^{(c),A}$	$T_2^{(c),A}$
Kein Drift						
Alle Beob. aus c			✓ [6.3.1]	✓ [8.3.1]	✓ [6.4.1]	✓ [8.4.1]
allgemein	✓ [6.2.1]	✓ [8.2.1]				
Linearer Trend						
Alle Beob. aus c				✓ [8.3.2]		✓ [8.4.2]
allgemein	x [6.2.2]	✓ [8.2.2]	x [6.3.2]		x [6.4.2]	

werden. In diesem Fall vereinfacht sich die Schätzfunktion zu (8.32) bzw. (8.33), da einige zufällige Komponenten als fest betrachtet werden können und in den linearen Regressionsmodellen nur nicht-stochastische Regressoren $z_i^{(c)}$ betrachtet werden. Es ist jedoch davon auszugehen, dass die Erwartungstreue auch allgemein ohne Einschränkung auf diesen Spezialfall gilt. Dies wird anhand der folgenden Simulationsstudie in Kapitel 9 untersucht.

Bei Unterstellung eines linearen Trends der Erwartungswertvektoren in den einzelnen Klassen kann die Erwartungstreue von keiner Schätzfunktion mehr erfüllt werden. In Abschnitt 6.2.2 wurde bewiesen, dass $T_1^{(c),P}$ nicht erwartungstreu für den Erwartungswertvektor der Prognose im Falle von identischen a-priori Klassenwahrscheinlichkeiten über die Zeit ist. In Abschnitt 6.3.2 konnte die Erwartungstreue von $T_1^{(c),K}$ für zwei Gegenbeispiele ($t = 3$ und $t = 4$) im Falle des Auftretens nur einer Klasse c widerlegt werden. In Abschnitt 6.4.2 wurde die Erwartungstreue von $T_1^{(c),A}$ für die Prognose des Erwartungswertvektors zum Zeitpunkt $t + 1$ ebenfalls für den Spezialfall, dass bis zum Zeitpunkt t nur Beobachtungen in Klasse c realisiert werden, widerlegt. Insgesamt ließ sich daraus schlussfolgern, dass die bisherige nicht-erweiterte Schätzfunktion $T_1^{(c)}$ im Falle eines linearen Trends der Erwartungswertvektoren in den Klassen in allen Methoden bezüglich des wahren Erwartungswertvektors $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose verzerrt ist.

Aufgrund dieses Ergebnisses wurde in Kapitel 7 ein allgemeiner erweiterter Schätzer entwickelt, welcher in allen Methoden für Online Diskriminanzanalyse Anwendung finden kann und die jeweiligen bisherigen Schätzer für die Erwartungswertvektoren der Klassen in der Klassifikationsregel der Diskriminanzanalyse ersetzen soll.

Die Verzerrung der ursprünglichen Schätzfunktion soll durch die entsprechende allgemeine Schätzfunktion (8.3) der erweiterten Methoden korrigiert werden:

$$T_2^{(c)}(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_t) = \bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i \in I} \left(Z_i^{(c)} - \bar{Z}_t^{(c)} \right)^2} \right) \left(\bar{Z}_t^{(c)} - (t+1) \right).$$

Durch die Erweiterung bzw. die Ersetzung der ursprünglichen Schätzer für die Erwartungswertvektoren der Klassen soll folglich die Prognosegüte der Verfahren für Online Diskriminanzanalyse bei Vorliegen von concept drift verbessert werden. Die speziellen erweiterten Schätzfunktionen für die drei betrachteten (und erweiterten) Methoden für Online Diskriminanzanalyse sind im Folgenden noch einmal vergleichend nebeneinandergestellt:

$$T_2^{(c),P}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_t) = \bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i\in I}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i\in I}\left(Z_i^{(c)} - \bar{Z}_t^{(c)}\right)^2}\right)\left(\bar{Z}_t^{(c)} - (t+1)\right)$$

mit $\boldsymbol{Y}_i^{(c)} \overset{(8.31)}{=} T_1^{(c),P}(\boldsymbol{X}_1,\ldots,\boldsymbol{X}_i)$,

$$\bar{\boldsymbol{Y}}_t^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i\in I}\left(\frac{1}{\sum_{k=1}^{i}\mathbb{1}_{\{\boldsymbol{X}_k\sim F_c\}}}\sum_{j=1}^{i}\left(\boldsymbol{X}_j\cdot\mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}}\right)\right),$$

$$Z_i^{(c)} \overset{(8.7)}{=} \frac{1}{\sum_{k=1}^{i}\mathbb{1}_{\{\boldsymbol{X}_k\sim F_c\}}}\sum_{j=1}^{i}\left(j\cdot\mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}}\right),$$

$$\bar{Z}_t^{(c)} \overset{(8.8)}{=} \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i\in I}\left(\frac{1}{\sum_{k=1}^{i}\mathbb{1}_{\{\boldsymbol{X}_k\sim F_c\}}}\sum_{j=1}^{i}\left(j\cdot\mathbb{1}_{\{\boldsymbol{X}_j\sim F_c\}}\right)\right),$$

bzw. im Spezialfall aus Voraussetzung 4 (Seite 138)

$$T_2^{(c),P}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)}) = \bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)} - (t+1)\right)$$

mit $\boldsymbol{Y}_i^{(c)} = T_1^{(c),P}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_i^{(c)}) = \frac{1}{i}\sum_{j=1}^{i}\boldsymbol{X}_j^{(c)}$,

$$\bar{\boldsymbol{Y}}_t^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\boldsymbol{Y}_i^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(\frac{1}{i}\sum_{j=1}^{i}\boldsymbol{X}_j^{(c)}\right),$$

$$z_i^{(c)} \overset{(8.11)}{=} \frac{i+1}{2},$$

$$\bar{z}_t^{(c)} \overset{(8.12)}{=} \frac{1}{2n_{\text{trend}}^{(c)}}\left(\frac{t(t+1)}{2} - \frac{(t-n_{\text{trend}}^{(c)})(t-n_{\text{trend}}^{(c)}+1)}{2}\right) + \frac{1}{2},$$

$$T_2^{(c),K}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_t^{(c)}) = \bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)\left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)}\right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\left(z_i^{(c)} - \bar{z}_t^{(c)}\right)^2}\right)\left(\bar{z}_t^{(c)} - (t+1)\right)$$

mit $\boldsymbol{Y}_i^{(c)} \overset{(8.32)}{=} T_1^{(c),K}(\boldsymbol{X}_1^{(c)},\ldots,\boldsymbol{X}_i^{(c)})$,

$$\bar{\boldsymbol{Y}}_t^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}}\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t}\boldsymbol{Y}_i^{(c)},$$

$$z_i^{(c)} \overset{(8.22)}{=} \frac{1}{(1-\lambda)(i-1)+\lambda}\cdot \left(\frac{(1-\lambda)^{i-1}\prod_{j=1}^{i-1}j}{\prod_{j=3}^{i}((1-\lambda)(j-2)+\lambda)} + \sum_{j=2}^{i-1}\frac{(1-\lambda)^{i-j}\lambda\prod_{k=j}^{i-1}k}{\prod_{k=j+1}^{i}((1-\lambda)(k-2)+\lambda)}\cdot j + \lambda i\right),$$

$$\bar{z}_t^{(c)} \overset{(8.23)}{=} \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} z_i^{(c)},$$

$$T_2^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_t^{(c)}) = \bar{\boldsymbol{Y}}_t^{(c)} - \left(\frac{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right) \left(\boldsymbol{Y}_i^{(c)} - \bar{\boldsymbol{Y}}_t^{(c)} \right)}{\sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(z_i^{(c)} - \bar{z}_t^{(c)} \right)^2} \right) \left(\bar{z}_t^{(c)} - (t+1) \right)$$

mit $\boldsymbol{Y}_i^{(c)} \overset{(8.33)}{=} T_1^{(c),A}(\boldsymbol{X}_1^{(c)}, \ldots, \boldsymbol{X}_i^{(c)})$,

$$\bar{\boldsymbol{Y}}_t^{(c)} = \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \boldsymbol{Y}_i^{(c)},$$

$$z_i^{(c)} \overset{(8.27)}{=} \frac{1}{\sum_{l=1}^{i-1} \left(\prod_{k=l}^{i-1} \lambda_{(k)}^{(c)} \right) + 1} \left(\sum_{j=1}^{i-1} \left(\left(\prod_{k=j}^{i-1} \lambda_{(k)}^{(c)} \right) j \right) + i \right),$$

$$\bar{z}_t^{(c)} \overset{(8.28)}{=} \frac{1}{n_{\text{trend}}^{(c)}} \sum_{i=t-n_{\text{trend}}^{(c)}+1}^{t} \left(\frac{1}{\sum_{l=1}^{i-1} \left(\prod_{k=l}^{i-1} \lambda_{(k)}^{(c)} \right) + 1} \left(\sum_{j=1}^{i-1} \left(\left(\prod_{k=j}^{i-1} \lambda_{(k)}^{(c)} \right) j \right) + i \right) \right).$$

In den vorangehenden vier Abschnitten konnte bewiesen werden, dass jede einzelne der drei Schätzfunktionen zum Zeitpunkt t im Spezialfall aus Voraussetzung 4, dass bis zum Zeitpunkt t nur Beobachtungen in Klasse c realisiert werden, weiterhin ebenso wie die ursprüngliche Schätzfunktion $T_1^{(c)}$ erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose zum Zeitpunkt $t+1$ ist, falls eine stabile Verteilung über die Zeit unterstellt wird. Für die Schätzfunktion der erweiterten Methode *Sequential ILDA* konnte die Erwartungstreue sogar ohne Beschränkung auf den Spezialfall bewiesen werden (vgl. Abschnitt 8.2.1).

Darüber hinaus ist hervorzuheben, dass die neuen Schätzfunktionen entgegen der bisherigen Schätzfunktionen in einigen Fällen immer noch erwartungstreu für den Erwartungswertvektor der Prognose sind, falls concept drift in Form eines linearen Trends der Erwartungswertvektoren der Klassen vorliegt (vgl. Tabelle 8.1). Für die Schätzfunktion $T_2^{(c),P}$ der erweiterten Methode *Sequential ILDA* kann die Erwartungstreue auch in der Situation von diesem concept drift im allgemeinen Fall (vgl. Abschnitt 8.2.2) bewiesen werden, für die Schätzfunktionen $T_2^{(c),K}$ und $T_2^{(c),A}$ der erweiterten Methoden *OLDC* (vgl. Abschnitt 8.3.2) und *Online Diskriminanzanalyse mit adaptivem Vergessen* (vgl. Abschnitt 8.4.2) mit Einschränkung auf den oben genannten Spezialfall aus Voraussetzung 4 (Seite 138).

Es ist allerdings davon auszugehen, dass die Schätzfunktionen der erweiterten Methoden allgemein bei concept drift in Form eines linearen Trends (7.1) der Erwartungswertvektoren der Klassen erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Verteilung der Prognose für Klasse c sind. Daher wird der allgemeine Fall für verschiedene Formen und Stärken von concept drift im folgenden Kapitel 9 mithilfe einer umfangreichen Simulationsstudie untersucht, um die theoretischen Ergebnisse zu untermauern und weitere Erkenntnisse zu gewinnen.

9 Vergleich der Methoden basierend auf Simulationsstudien

In diesem Kapitel werden die verschiedenen Methoden für Online Diskriminanzanalyse mit der vorgestellten Erweiterung zur Verbesserung der Prognosegüte verglichen. Dabei werden verschiedene Datensituationen betrachtet, die sich in Art und Stärke des concept drifts voneinander unterscheiden. Zudem werden alle Verfahren auch auf Datensituationen ohne concept drift miteinander verglichen, um zu untersuchen, ob die Erweiterung in diesen Situationen nachteilig ist bzw. um zu zeigen, dass dem nicht so ist.

Der Fokus beim Vergleich der Ergebnisse liegt zum einen auf der Betrachtung der Prognosegüte, da diese bei der Klassifizierung neuer Beobachtungen durch eine Klassifikationsmethode eine große Rolle spielt und in dieser Arbeit explizit für bestimmte Datensituationen verbessert werden sollte. Zum anderen wird der zeitliche Verlauf der (prognostizierten) Mittelwertvektoren als Schätzer für die Erwartungswertvektoren in den einzelnen Klassen beziehungsweise die Abweichung dieser zu den wahren Erwartungswertvektoren betrachtet. Dies soll die theoretischen Ergebnisse aus Kapitel 8 unterstützen beziehungsweise an den Stellen ergänzen, an denen nur Spezialfälle der Erwartungstreue bewiesen wurden.

Die Simulationsstudie konzentriert sich auf die sequentiellen Methoden für Online Diskriminanzanalyse und ihre jeweilige Erweiterung und nicht auf die (teilweise neu eingeführten) entsprechenden Chunk Varianten. Es werden demnach *Sequential ILDA* (Abschnitt 4.2), *OLDC* mit fester und adaptiver Lernrate (Abschnitt 4.3) sowie *LDA-AF* und *QDA-AF* (Abschnitt 4.4) betrachtet, bei denen eine Aktualisierung des Modells immer auf einer einzelnen neuen Beobachtung im Datenstrom basiert. Chunk Varianten machen in der Praxis Sinn, wenn der Durchfluss des Datenstroms sehr groß ist und nicht nach jeder neuen Beobachtung ein aktuelles Modell bzw. eine Prognose erforderlich ist. Oder wenn im Datenstrom neue Beobachtungen schubweise ankommen und zu jedem Zeitpunkt eine Reihe neuer Beobachtungen gleichzeitig für eine Aktualisierung zur Verfügung steht. In jedem Fall sind die vorgestellten Chunk Versionen (*Chunk Incremental LDA* aus Abschnitt 4.2 und Erweiterungen der anderen Methoden auf Chunks aus Kapitel 5) nur sinnvoll, wenn die Beobachtungen eines jeden Chunks jeweils aus derselben Verteilung stammen und nur für jeweils ganze Chunks ein Drift vorliegt. Vor diesem Hintergrund bietet die Untersuchung der Chunk Varianten keinen Mehrwert gegenüber den sequentiellen Update-Methoden. Es fließen lediglich mehr Beobachtungen aus derselben Verteilung bei einer Aktualisierung des Modells ein. Die Diskriminanzanalysemodelle nach einer Aktualisierung durch eine einzelne

oder durch mehrere Beobachtungen sind jedoch qualitativ ähnlich. Zur Konzentration auf das Wesentliche wird daher die Einbindung der Chunk Versionen und ihrer Erweiterungen durch Integration lokaler linearer Regressionsmodelle in die Simulationsstudie als Ausblick offen gelassen.

Zur Planung des Aufbaus der Simulationsstudie zum Vergleich der verschiedenen Methoden zur Online Diskriminanzanalyse und ihrer in dieser Arbeit entwickelten Erweiterungen müssen zum einen die betrachteten Parameterbereiche und -kombinationen der Methoden festgelegt werden. Zum anderen muss der „Raum“ der Datensätze mit und ohne vorliegendem concept drift, auf welchem die Methoden verglichen werden sollen, definiert und eingegrenzt werden. Im folgenden Abschnitt 9.1 werden zwei typische in verschiedenen Veröffentlichungen zum Umgang mit concept drift betrachtete Datensituationen beschrieben. In Abschnitt 9.2 werden der Raum der weiteren betrachteten Datensituationen und Arten von concept drift definiert und die Datensätze für die Simulationsstudie werden beschrieben. Abschnitt 9.3 erläutert die Durchführung der Simulationsstudie und Abschnitt 9.4 befasst sich mit der Wahl der einstellbaren Parameter der untersuchten Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen. Abschnitt 9.5 beschreibt die Implementierung und Durchführung der Simulationsstudie. Schlussendlich werden in Abschnitt 9.6 die Ergebnisse der Simulationsstudie vorgestellt und diskutiert.

9.1 Typische betrachtete Datensituationen

In der Literatur zu concept drift und verschiedenen Veröffentlichungen zum Umgang mit concept drift haben sich typische simulierte Datensituationen zur Validierung der Ergebnisse etabliert. Da unter anderem die Methoden *OLDC* (Kuncheva und Plumpton, 2008) und *Online Diskriminanzanalyse mit adaptivem Vergessen* (Anagnostopoulos et al., 2012) erweitert werden und in den genannten Veröffentlichungen die simulierten Datensätze *moving plane* (unten *moving hyperplane*) und *STAGGER* verwendet werden, werden die entwickelten Erweiterungen der Methoden auch auf diesen Datensituationen angewandt, um einen Vergleich zu gewährleisten.

Zu beachten sei hier, dass streng genommen die Annahmen der Linearen Diskriminanzanalyse auf den genannten Datensituationen verletzt sind, da im ersten Fall keine normalverteilten Daten betrachtet werden, im zweiten Fall sogar diskrete Einflussgrößen. Generell handelt es sich um sehr synthetische Datensätze, die wenig praktische Relevanz haben. Daher werden im folgenden Abschnitt 9.2 weitere Datensituationen beschrieben, die verschiedene Arten von concept drift charakterisieren und den „Raum“ der möglichen Datensituationen möglichst gut abdecken sollen.

Moving hyperplane Aus der Idee der *moving hyperplane* (z. B. erläutert von Narasimhamurthy und Kuncheva (2007, S. 386 f.) und Hulten et al. (2001, S. 102)) resultierende

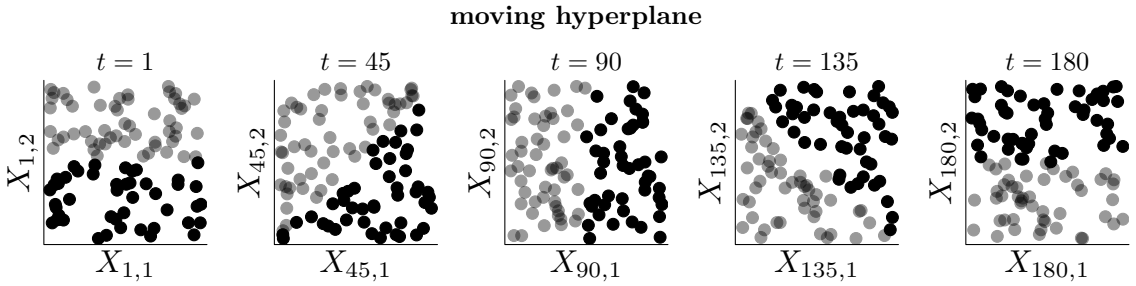


Abbildung 9.1: Veranschaulichung der zweidimensionalen Verteilungen durch *moving hyperplane* in den zwei Klassen (grau: Klasse 1, schwarz: Klasse 2) für fünf verschiedene Zeitpunkte. Die Klassengrenze rotiert jeden Zeitpunkt t um ein Grad um ihre eigene Achse bzw. den Nullpunkt des Einheitsquadrates.

künstliche Datensätze haben sich in der Literatur zu concept drift als gängige Testdatensätze zur Validierung der entwickelten Klassifikationsmethoden etabliert.

Zur Konstruktion werden die zweidimensionalen Beobachtungen zufällig aus dem Einheitsquadrat in $[-1, 1] \times [-1, 1]$ gezogen, d. h. beide Dimensionen können als Zufallszahlen aus $\mathcal{U}(-1, 1)$ gezogen werden. Die Zuordnung der Klassenausprägung y für eine so erzeugte Zufallszahl $\boldsymbol{x} = (x_1, x_2)^T$ erfolgt zum Zeitpunkt t anhand einer zeitabhängigen entgegen den Uhrzeigersinn rotierenden (1 Grad pro Zeitpunkt t) Hyperebene, beginnend bei einer Geraden durch den Ursprung und Null in zweiter Dimension (Horizontale) zum initialen Zeitpunkt $t = 0$:

$$y = \begin{cases} 1, & x_2 \cos\left(\frac{t\pi}{180}\right) - x_1 \sin\left(\frac{t\pi}{180}\right) > 0, \\ 2, & x_2 \cos\left(\frac{t\pi}{180}\right) - x_1 \sin\left(\frac{t\pi}{180}\right) \leq 0, \end{cases} \tag{9.1}$$

da die Gerade im Einheitskreis durch den Ursprung mit t° Steigung (für $t < 90$) beschrieben werden kann durch $x_2 = \frac{\sin\left(t^\circ \cdot \frac{\pi}{180}\right)}{\cos\left(t^\circ \cdot \frac{\pi}{180}\right)} \cdot x_1$.

Die zweidimensionale Verteilung der Beobachtungen aus den beiden Klassen ist in Abbildung 9.1 für verschiedene Zeitpunkte veranschaulicht.

Für den Datenstrom werden 360 zweidimensionale Beobachtungen simuliert. Die Zuordnung der entsprechenden Klassenausprägung erfolgt mithilfe von (9.1) für jeden Zeitpunkt $t = 1, \ldots, 360$. Es wird somit eine Datensituation mit einem incremental drift erzeugt, da sich die zugrunde liegende Verteilung für jeden Zeitpunkt durch Rotation um ein Grad nur leicht ändert. Zu beachten sei hier, dass trotz der stetigen Einflussgrößen beide Klassen perfekt trennbar sind. Zusätzlich unterliegt auch die Kovarianzmatrix der einzelnen Klassen einem Drift über die Zeit und die Verteilungen beider Klassen charakterisieren aufgrund der Einschränkung auf das Einheitsquadrat keine üblichen Verteilungen.

All dies widerspricht den Annahmen der Linearen Diskriminanzanalyse und spiegelt keine Datensituation aus der Praxis wider. Daher werden im Folgenden (Abschnitt 9.2) noch realistischere Datensituationen für die Simulationsstudie eingeführt.

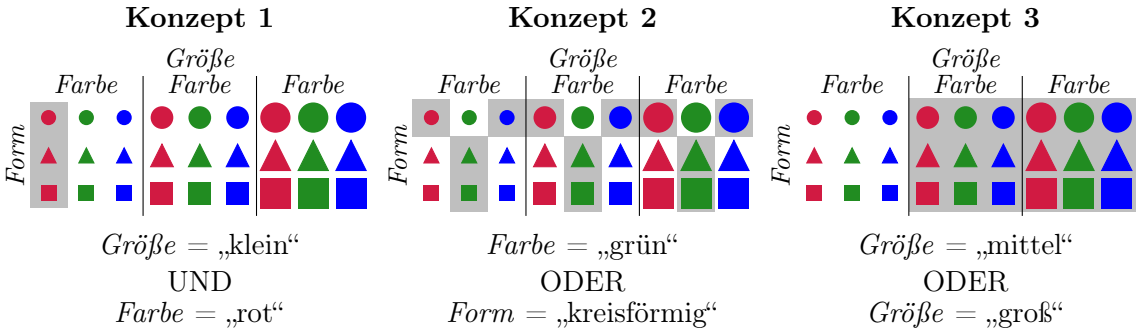


Abbildung 9.2: Schematische Darstellung der drei Konzepte von *STAGGER*. Grau hinterlegt sind jeweils die Objekte aus Klasse 1 (Kombinationen, für die logische Verknüpfung der Ausprägungen mit „JA“ beantwortet wird).

STAGGER Das *STAGGER* concept wurde von Schlimmer und Granger (1986) eingeführt und seitdem im Rahmen von Arbeiten zu concept drift immer wieder herangezogen. Narasimhamurthy und Kuncheva (2007, S. 386 f.) beschreiben die Konstruktion eines Datensatzes basierend auf dem *STAGGER* concept anschaulich. Diese wurde auch bereits zuvor von Widmer und Kubat (1996, S. 78) in kurzer Form beschrieben. Objekte werden durch drei diskrete Einflussgrößen mit jeweils drei möglichen Ausprägungen beschrieben und anhand dieser Ausprägungen in zwei Klassen gruppiert:

- *Größe* $\in$ {„klein“, „mittel“, „groß“}
- *Farbe* $\in$ {„rot“, „grün“, „blau“}
- *Form* $\in$ {„viereckig“, „kreisförmig“, „dreieckig“}

Demnach gibt es 27 Kombinationen, bezüglich derer sich die Objekte unterscheiden können.

Es werden nun drei verschiedene Konzepte betrachtet, anhand derer die Einteilung in zwei Gruppen y_1 und y_2 erfolgt. Dies ist in Abbildung 9.2 veranschaulicht:

- Konzept 1: Klasse 1: „klein“ und „rot“ (3 Kombinationen);
  Klasse 2: 24 Kombinationen
- Konzept 2: Klasse 1: „grün“ oder „kreisförmig“ (12 Kombinationen);
  Klasse 2: 15 Kombinationen
- Konzept 3: Klasse 1: „mittel“ oder „groß“ (18 Kombinationen);
  Klasse 2: 9 Kombinationen

Wie von Narasimhamurthy und Kuncheva (2007, S. 386 f.) beschrieben wird der Datensatz folgendermaßen konstruiert: Es wird ein Datenstrom aus 120 Beobachtungen betrachtet, bei dem jeweils 40 aufeinander folgende Beobachtungen einem Konzept entstammen, bevor ein plötzlicher Wechsel zwischen den Konzepten und demnach ein sudden drift erfolgt. Die 120 Beobachtungen werden unabhängig diskret gleichverteilt aus dem gesamten Wertebereich gezogen, d. h. jede Kombination mit Wahrscheinlichkeit $\frac{1}{27}$. Die Zuordnung der Ausprägung der Klassenvariable erfolgt dann durch die logischen Verknüpfungen von

Konzept 1 (Beobachtungen 1–40), Konzept 2 (Beobachtungen 41–80) und Konzept 3 (Beobachtungen 81–120). Die Klassenzugehörigkeit ändert sich zu den Zeitpunkten $t = 40$ und $t = 80$ abrupt:

$$P_q(Y = y_c \,|\, \boldsymbol{X} = \boldsymbol{x}) \neq P_r(Y = y_c \,|\, \boldsymbol{X} = \boldsymbol{x}) \neq P_s(Y = y_c \,|\, \boldsymbol{X} = \boldsymbol{x})$$
$$\text{für } q \in \{1, \ldots, 40\},\ r \in \{41, \ldots, 80\},\ s \in \{81, \ldots, 120\} \text{ und } c = 1, 2.$$

Folglich ändern sich aufgrund des Zusammenhangs, welcher durch den Satz von Bayes beschrieben wird, auch die a-priori Wahrscheinlichkeiten in beiden Klassen. Diese sehen für die beiden Klassen demnach in Abhängigkeit des Zeitpunktes t folgendermaßen aus:

$$p_t^{(1)} = \begin{cases} \frac{3}{27}, & 1 \leq t \leq 40, \\ \frac{12}{27}, & 41 \leq t \leq 80, \\ \frac{18}{27}, & 81 \leq t \leq 120, \end{cases} \qquad p_t^{(2)} = \begin{cases} \frac{24}{27}, & 1 \leq t \leq 40, \\ \frac{15}{27}, & 41 \leq t \leq 80, \\ \frac{9}{27}, & 81 \leq t \leq 120. \end{cases}$$

Diese Datensituation ist daher ein Beispiel für einen möglichen Auslöser eines Drifts, welche in Abschnitt 2.2.2 vorgestellt wurden (vgl. Abbildung 2.3 auf Seite 17).

Für die Klassifikation durch eine Lineare Diskriminanzanalyse werden jeweils zwei Dummy-Variablen für jede der drei diskreten Variablen mit ihren drei Merkmalsausprägungen eingefügt, wobei die Ausprägungen „groß“, „blau“ und „dreieckig“ als Referenzkategorie dienen. Damit erhält man einen Datensatz bestehend aus sechs Einflussvariablen (Dummy-Variablen) und der Zielvariablen.

Auch hierbei handelt es sich um ein recht künstliches Datenbeispiel, bei welchem zusätzlich aufgrund von diskreten Einflussvariablen die Annahmen der Linearen Diskriminanzanalyse verletzt sind. Im folgenden Abschnitt werden daher zusätzlich verschiedene weitere Datensituationen eingeführt, die den Raum der möglichen concept drift Situationen unter der Annahme variierender Erwartungswerte und fester Kovarianzen möglichst gut abdecken sollen.

9.2 Raum der betrachteten Datensituationen und Arten von Concept Drift

Der „Raum“ aller möglichen Datensituationen mit vorliegendem concept drift bzw. der „Raum“ möglicher concept drift Situationen ist unendlich groß. Dies wird bereits durch den Abschnitt 2.2.2 deutlich, in welchem concept drift aus verschiedenen Richtungen qualitativ und quantitativ beleuchtet wird. Es wird u. a. bereits unterschieden zwischen *Auslöser* und *Art* des Drifts sowie der in der Literatur häufig unterschiedene real und virtual drift diskutiert. Zudem werden neben den qualitativen Differenzierungen zusätzlich quantitative

Definitionen von Webb et al. (2016) vorgestellt, die unter anderem auf Distanzfunktionen basieren.

Zusätzlich zur Differenzierung verschiedener Arten und dem Auslöser des Drifts kann auch noch der Raum der betrachteten Parameter des Datensatzes abgetastet werden. Dazu zählen die Anzahl der Klassen M, die Anzahl betrachteter Parameter p sowie die spezielle Verteilung der Beobachtungen innerhalb jeder Klasse bzw. deren Parametrisierung. Bei Betrachtung multivariater Normalverteilungen umfasst dies die Erwartungswertvektoren (Lage im p-dimensionalen Raum) und die (gemeinsame) Kovarianzmatrix (Streuung im p-dimensionalen Raum). Hinsichtlich eines Klassifikationsproblems ist es nämlich relevant wie stark sich die Verteilungen der einzelnen Klassen überlappen und in welche Richtung der Drift jeder Klasse erfolgt.

Da ein Drift für die Verteilung jeder einzelnen Klasse getrennt betrachtet werden kann, jedoch die gemeinsame Verteilung für die Klassifikation eine Rolle spielt, ergeben sich unendlich viele Möglichkeiten an relevanten Datensätzen und somit ein unendlich großer Raum an möglichen concept drift Situationen.

Wie bereits in Abschnitt 2.2.2 erwähnt, wird in dieser Arbeit jegliche Änderung des concepts als concept drift bezeichnet. Daher wird nicht zwischen verschiedenen Auslösern unterschieden, da nur interessant ist, ob die a-posteriori Wahrscheinlichkeit der Klassenzugehörigkeit von einem Drift betroffen ist, unabhängig davon, ob dies direkt oder als Konsequenz der Veränderung einer anderen Größe der Fall ist. Der Fokus dieser Arbeit liegt auf der Betrachtung eines linearen Trends der Erwartungswertvektoren $\boldsymbol{\mu}^{(c)}$ über die Zeit. Der Drift wird daher direkt bezüglich der Erwartungswertvektoren modelliert, indem diese über die Zeit in den einzelnen Klassen variieren. Bezüglich der eingeführten Unterteilung verschiedener Auslöser eines Drifts von Moreno-Torres et al. (2012), welche ebenfalls in Kapitel 2.2.2 diskutiert wird, wird somit ein *covariate drift* (vgl. Abbildung 2.1 auf Seite 16) in der folgenden Simulationsstudie betrachtet.

Es werden Datensituationen mit verschiedener Ausprägung eines *incremental drifts* erzeugt (vgl. Abbildung 2.6 auf Seite 27), da die vorgestellte Erweiterung der Methoden für Online Diskriminanzanalyse für diese Art von concept drift entwickelt wurde und aufgrund der theoretischen Eigenschaften insbesondere in solchen Situationen die Prognosegüte verbessern können soll.

Zusätzlich werden jeweils ein Datenstrom mit einem *sudden drift* und einer ohne Drift, also mit stabiler Verteilung über die Zeit, sowie zwei Datensätze mit einem *gradual drift* konstruiert. Obwohl die vorgestellte Erweiterung der Methoden insbesondere für einen *incremental drift* entwickelt wurde, ist es interessant zu untersuchen, ob sie auch mit den drei genannten Situationen umgehen kann, da in praktischen Anwendungen die Form des zugrunde liegenden Drifts in den allermeisten Fällen nicht bekannt ist.

Reoccurring concepts (vgl. Abbildung 2.6 auf Seite 27) werden nicht beachtet, da die in dieser Arbeit entwickelte Erweiterung nicht für solche wiederkehrenden Muster gedacht ist. Es werden keine vollständigen Informationen über beliebig lange vergangene Verteilungen gespeichert, die dann wieder aufgerufen werden. Vielmehr werden solche wiederkehrenden Muster wie ein *incremental* oder *sudden drift* interpretiert, je nachdem wie stark die Veränderungen zwischen den einzelnen Verteilungen sind und wie schnell diese sich mit der Zeit verändern.

In allen Datensituationen werden $M = 2$ Klassen betrachtet, da Strukturen zur Analyse der Problemstellung vermutlich bereits bei zwei Klassen deutlich werden und gleichzeitig das Klassifikationsproblem einfach gehalten wird. Als Anzahl möglicher Variablen werden in der Simulationsstudie $p \in \{2, 3, 10\}$ betrachtet. Bei $p = 2$ Dimensionen sind die Datensituationen gut zu veranschaulichen. Dies ist auch im dreidimensionalen Raum noch eingeschränkt möglich. Um zu evaluieren, ob die Methoden und insbesondere die entwickelten Erweiterungen auch auf mehr Dimensionen noch gut funktionieren, werden bis zu $p = 10$ Dimensionen betrachtet.

Für beide Klassen werden in allen Datensituationen jeweils multivariate Normalverteilungen betrachtet, da dies die Annahme der Kanonischen Diskriminanzanalyse ist (vgl. Abschnitt 3.2). Die Kovarianzmatrix ist dabei sowohl in beiden Klassen identisch als auch über die Zeit invariant. In den simulierten Datensätzen wird durchgehend

$$\boldsymbol{\Sigma}_{p\times p} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

herangezogen. Es werden also unkorrelierte Variablen mit einheitlicher Varianz 2 über alle Dimensionen betrachtet, um das Datenproblem bezüglich der Streuung einfach zu halten. Die Konzentration auf eine einzelne Kovarianzmatrix für alle Datensituationen liegt darin begründet, dass der Fokus der entwickelten Erweiterung auf den Erwartungswertvektoren der Klassen liegt. Für die Streuung innerhalb der Klassen werden daher für jeden Zeitpunkt die theoretischen Voraussetzungen der LDA beachtet. Korrelierte Variablen würden das Problem in Hinblick auf die Betrachtung der Regressionsmodelle bei der Erweiterung kaum verändern, da der Trend aller Dimensionen einzeln modelliert wird. Es würde sich lediglich eine andere Klassifikationsgrenze bei der Diskriminanzanalyse ergeben.

Die Erwartungswertvektoren $\boldsymbol{\mu}_i^{(c)}$ beider Klassen $c \in \{1, 2\}$ unterscheiden sich und verändern sich jeweils mit der Zeit i, um unterschiedlichste Formen von concept drift zu repräsentieren. Da insbesondere das Zusammenspiel bzw. die gleichzeitige Betrachtung der Verteilungen innerhalb der beiden Klassen für die Klassifikation eine Rolle spielt, werden die Bewegungen der Erwartungswertvektoren im p-dimensionalen Raum in den einzelnen Klassen nicht unabhängig voneinander modelliert, sondern es werden die folgenden sechs

Situationen betrachtet, um eine Bandbreite verschiedener concept drift Situationen (und stabile Verteilungen) abzubilden:

- Incremental Drift durch „Kreisen“ der Erwartungswerte,
- Incremental Drift durch „Kreuzen“ der Erwartungswerte,
- Incremental Drift durch „Vorbeilaufen“ der Erwartungswerte,
- Incremental Drift durch „Vorbeilaufen“ (gerade) der Erwartungswerte,
- Gradual Drift mit „Kreuzen“ der Erwartungswerte,
- Gradual Drift mit „Austausch“ der Erwartungswerte,
- Sudden Drift,
- Kein Drift bzw. Betrachtung einer stabilen Verteilung.

Für jede dieser Datensituationen wird ein p-dimensionaler Datensatz mit 4000 Beobachtungen und jeweiligen Klassenlabels erzeugt, welcher als Datenstrom mit einer auftretenden Beobachtung zu jedem Zeitpunkt interpretiert werden kann. Die a-priori Wahrscheinlichkeiten sind konstant $p^{(1)} = p^{(2)} = 0.5$, sodass zu jedem der 4000 Zeitpunkte zufällig eine Beobachtung aus der jeweiligen Verteilung einer der beiden Klassen erzeugt wird. Eine Visualisierung der Verteilungen beider Klassen und ihrer Veränderung über die Zeit bei Betrachtung von $p = 2$ Variablen ist in Abbildung 9.3 für die ersten vier Situationen und die Situation mit Sudden Drift für verschiedene Zeitpunkte beispielhaft dargestellt. Abbildung 9.4 enthält eine Veranschaulichung des Trends der Erwartungswertvektoren in den beiden Klassen bei Betrachtung von $p = 3$ Variablen für die ersten vier Situationen.

In den folgenden Abschnitten wird die Simulation der acht Datensituationen – insbesondere die Modellierung des jeweils betrachteten Trends der Erwartungswertvektoren – im Detail erläutert. Bei den ersten vier Datensituationen a)–d) wird ein incremental drift modelliert, in dem Sinne, dass sich die den Beobachtungen zugrunde liegenden Verteilungen beider Klassen zu jedem einzelnen Zeitpunkt (schrittweise) verändern.

a) „Kreisen“ der Erwartungswerte Die Erwartungswertvektoren liegen im p-dimensionalen Raum zu jedem Zeitpunkt auf einem (zweidimensionalen) Kreis mit Radius 2 auf gegenüberliegenden Seiten (vgl. Abbildungen 9.3 (a) und 9.4 (a)) beginnend bei

$$\boldsymbol{\mu}_1^{(1)} = \begin{pmatrix} 2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_1^{(2)} = \begin{pmatrix} -2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Bei mehr als $p = 2$ betrachteten Variablen werden die $p - 2$ letzten Elemente der Erwartungswertvektoren beider Klassen über alle Zeitpunkte konstant auf 0 gesetzt, es wird nur eine Bewegung (auf dem Kreis) bezüglich der ersten beiden Dimensionen betrachtet. Zu jedem Zeitpunkt erfolgt eine Rotation der Verteilungen beider Klassen auf diesem Kreis

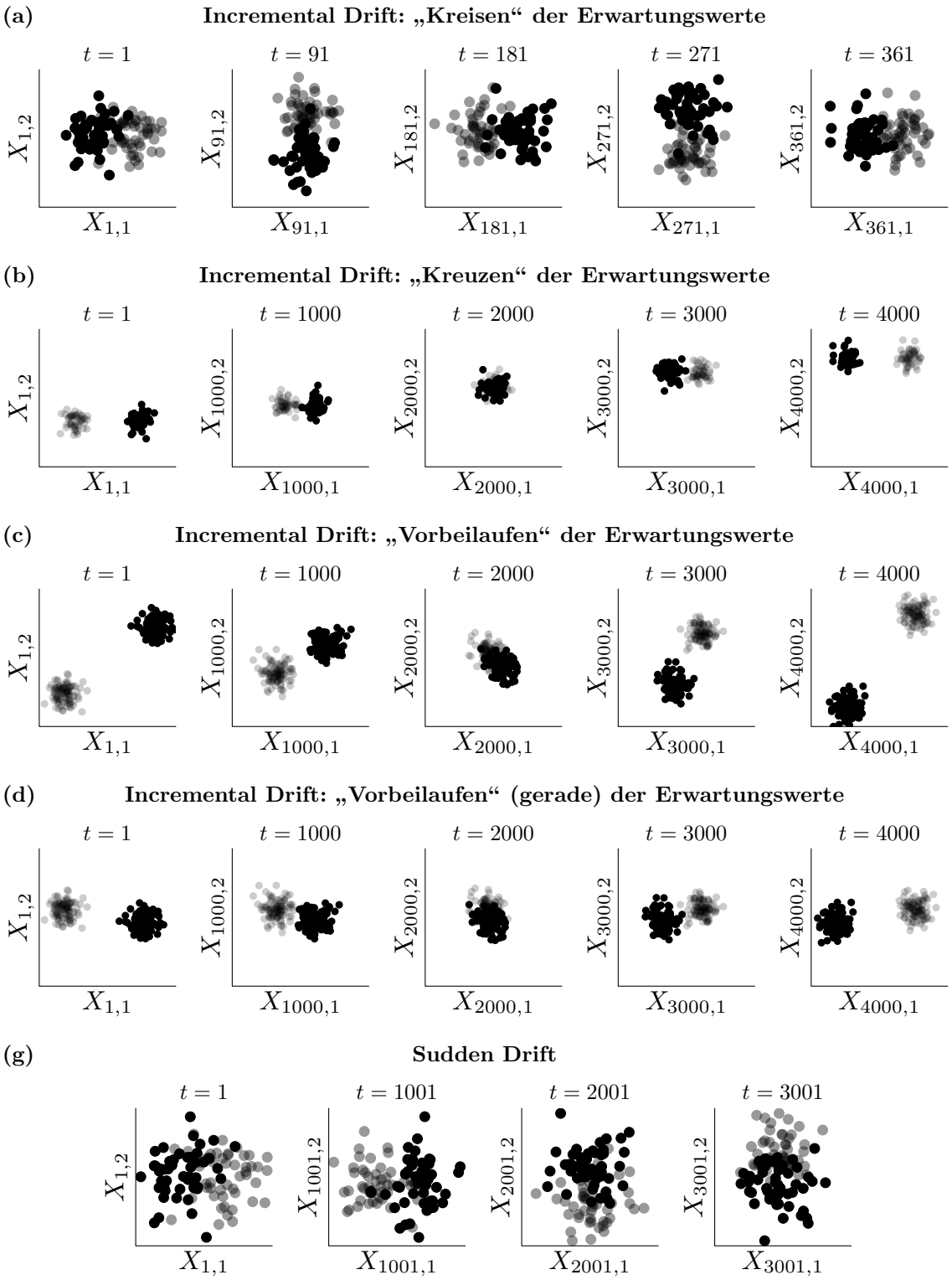


Abbildung 9.3: Veranschaulichung der zweidimensionalen Verteilungen in den zwei Klassen (grau: Klasse 1, schwarz: Klasse 2) der verschiedenen betrachteten Datensituationen (Zeilen) für jeweils verschiedene Zeitpunkte (Spalten).

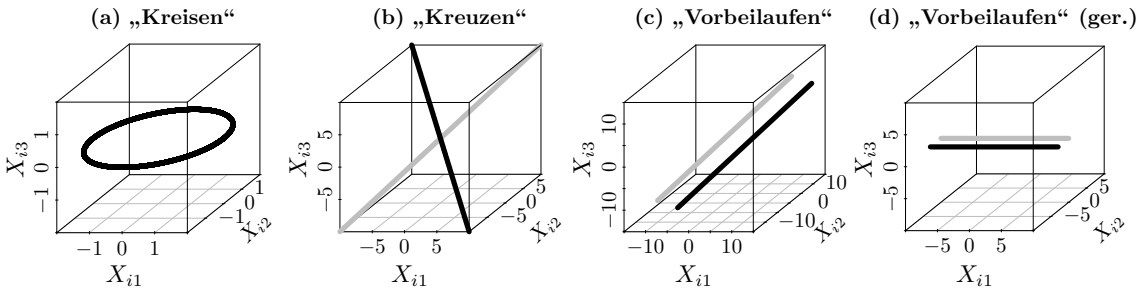


Abbildung 9.4: Veranschaulichung der Funktionen, auf denen sich die Erwartungswertvektoren der zwei Klassen (grau: Klasse 1, schwarz: Klasse 2) im Falle von $p = 3$ Dimensionen für die Zeitpunkte $i = 1, \ldots, 4000$ beim incremental drift bewegen (Grafik erstellt mithilfe des `R`-Paketes `scatterplot3d` (Ligges und Mächler, 2003)).

um jeweils ein Grad in dieselbe Richtung, sodass nach jeweils 360 Zeitpunkten eine volle Rotation erreicht und die Verteilung des Ausgangspunktes wieder angenommen wird.

Die Annahme eines globalen linearen Trends der Erwartungswerte ist hier nicht erfüllt. Diese Situation ist jedoch interessant, da bei der Erweiterung der Methoden ein *lokaler* linearer Trend betrachtet wird. In der Simulationsstudie kann somit untersucht werden, ob beliebige (nicht-lineare) Trends der Erwartungswertvektoren durch einen linearen Trend approximiert werden können und auch in beliebigen concept drift Situationen eine Verbesserung des Prognosefehlers erzielt werden kann.

b) „Kreuzen“ der Erwartungswerte In dieser Datensituation ist die Annahme eines linearen Trends der Erwartungswerte global erfüllt. Die beiden Erwartungswertvektoren bewegen sich jeweils mit konstanter und in beiden Klassen identischer Geschwindigkeit linear im p-dimensionalen Raum (vgl. Abbildungen 9.3 (b) und 9.4 (b)). Die Geraden „schneiden“ sich in der Mitte des Datenstroms (also nach 2000 bis 2001 Beobachtungen), die Verteilungen der beiden Klassen nähern sich mit der Zeit also immer mehr an, bis sie sich fast komplett überschneiden, bevor sie sich wieder voneinander entfernen.

Die Erwartungswertvektoren für die Zeitpunkte $i = 1, \ldots, 4000$ folgen den folgenden linearen Trends:

$$\begin{aligned}
\boldsymbol{\mu}_i^{(1)} &= \boldsymbol{\beta}_0^{(1)} + \boldsymbol{\beta}_1^{(1)} i = -10\, \tilde{\boldsymbol{\beta}}_0^{(1)} + 0.005\, \tilde{\boldsymbol{\beta}}_1^{(1)} i, \\
\boldsymbol{\mu}_i^{(2)} &= \boldsymbol{\beta}_0^{(2)} + \boldsymbol{\beta}_1^{(2)} i = -10\, \tilde{\boldsymbol{\beta}}_0^{(2)} + 0.005\, \tilde{\boldsymbol{\beta}}_1^{(2)} i
\end{aligned}$$

mit $\tilde{\boldsymbol{\beta}}_0^{(1)} = \mathbf{1}_p$, $\tilde{\boldsymbol{\beta}}_1^{(1)} = \mathbf{1}_p$ jeweils p-dimensionale Einsenvektoren für Klasse 1 und

$$\tilde{\boldsymbol{\beta}}_0^{(2)} = \tilde{\boldsymbol{\beta}}_1^{(2)} = \begin{cases} (-1, 1)^T, & p = 2, \\ (-1, 1, 1)^T, & p = 3, \\ (-1, -1, -1, -1, -1, 1, 1, 1, 1, 1)^T, & p = 10. \end{cases}$$

Zu Beginn des Datenstroms wird die Lage der Verteilungen charakterisiert durch

$$\boldsymbol{\mu}_1^{(1)} = -9.995 \cdot \mathbf{1}_p, \quad \boldsymbol{\mu}_1^{(2)} = \begin{cases} (9.995, -9.995)^T, & p = 2, \\ (9.995, -9.995, -9.995)^T, & p = 3, \\ (\underbrace{9.995, \ldots, 9.995}_{5 \text{ Elemente}}, \underbrace{-9.995, \ldots, -9.995}_{5 \text{ Elemente}})^T, & p = 10. \end{cases}$$

Zum Zeitpunkt $i = 4000$ haben sich die Erwartungswerte folgendermaßen verschoben und ausgehend vom Ursprung $\mathbf{0}_p$ in allen Dimensionen die Seiten beinahe vertauscht:

$$\boldsymbol{\mu}_{4000}^{(1)} = 10 \cdot \mathbf{1}_p, \quad \boldsymbol{\mu}_{4000}^{(2)} = \begin{cases} (-10, 10)^T, & p = 2, \\ (-10, 10, 10)^T, & p = 3, \\ (\underbrace{-10, \ldots, -10}_{5 \text{ Elemente}}, \underbrace{10, \ldots, 10}_{5 \text{ Elemente}})^T, & p = 10. \end{cases}$$

c) „Vorbeilaufen" der Erwartungswerte In dieser Situation wird die Verteilung der Klasse 1 zu jedem Zeitpunkt analog zu jener aus Datensituation b) simuliert. Der Erwartungswertvektor von Klasse 2 bewegt sich aus entgegengesetzter Richtung in gleicher Geschwindigkeit auf einer parallelen Geraden an jenem von Klasse 1 vorbei. Dadurch laufen die beiden Verteilungen im Laufe der Zeit auf parallelen Geraden im p-dimensionalen Raum aneinander vorbei (vgl. Abbildungen 9.3 (c) und 9.4 (c)). Diese parallelen Geraden liegen im p-dimensionalen Raum in jeder Dimension absolut um den Wert 3 voneinander entfernt. Im Speziellen wird der folgende lineare Trend der Erwartungswertvektoren für $i = 1, \ldots, 4000$ modelliert:

$$\begin{aligned} \boldsymbol{\mu}_i^{(1)} &= \boldsymbol{\beta}_0^{(1)} + \boldsymbol{\beta}_1^{(1)} i = -10\, \tilde{\boldsymbol{\beta}}_0^{(1)} + 0.005\, \tilde{\boldsymbol{\beta}}_1^{(1)} i, \\ \boldsymbol{\mu}_i^{(2)} &= \boldsymbol{\beta}_0^{(2)} + \boldsymbol{\beta}_1^{(2)} i = \boldsymbol{\beta}_0^{(2)} - 0.005\, \tilde{\boldsymbol{\beta}}_1^{(2)} i \end{aligned}$$

mit $\tilde{\boldsymbol{\beta}}_0^{(1)} = \mathbf{1}_p$, $\tilde{\boldsymbol{\beta}}_1^{(1)} = \mathbf{1}_p$, $\tilde{\boldsymbol{\beta}}_1^{(2)} = \mathbf{1}_p$ jeweils p-dimensionale Einsenvektoren und

$$\boldsymbol{\beta}_0^{(2)} = \begin{cases} (13.005, 7.005)^T, & p = 2, \\ (13.005, 13.005, 7.005)^T, & p = 3, \\ (13.005, 13.005, 13.005, 13.005, 13.005, 7.005, 7.005, 7.005, 7.005, 7.005)^T, & p = 10 \end{cases}$$

der Intercept für Klasse 2.

Zum Startzeitpunkt $i = 1$ sind die Erwartungswertvektoren gegeben durch

$$\boldsymbol{\mu}_1^{(1)} = -9.995 \cdot \mathbf{1}_p, \quad \boldsymbol{\mu}_1^{(2)} = \begin{cases} (13, 7)^T, & p = 2, \\ (13, 13, 7)^T, & p = 3, \\ (\underbrace{13, \ldots, 13}_{5 \text{ Elemente}}, \underbrace{7, \ldots, 7}_{5 \text{ Elemente}})^T, & p = 10 \end{cases}$$

und zum Endzeitpunkt $i = 4000$ des Datenstroms sehen sie folgendermaßen aus:

$$\boldsymbol{\mu}_{4000}^{(1)} = 10 \cdot \mathbf{1}_p, \quad \boldsymbol{\mu}_{4000}^{(2)} = \begin{cases} (-6.995, -12.995)^T, & p = 2, \\ (-6.995, -6.995, -12.995)^T, & p = 3, \\ (\underbrace{-6.995, \ldots, -6.995}_{5 \text{ Elemente}}, \underbrace{-12.995, \ldots, -12.995}_{5 \text{ Elemente}})^T, & p = 10. \end{cases}$$

Zu den Zeitpunkten $i = 2000$ und $i = 2001$ haben die Erwartungswertvektoren den kleinsten euklidischen Abstand zueinander. Da symmetrische Kovarianzmatrizen betrachtet werden, überlappen sich zu diesem Zeitpunkt die Klassen also verhältnismäßig am stärksten. Der Erwartungswert der Klasse 1 liegt zum Zeitpunkt $i = 2000$ im Ursprung $\boldsymbol{\mu}_{2000}^{(1)} = \mathbf{0}_p$. Daraus ergibt sich der Abstand $\sqrt{(-3.005)^2 + (2.995)^2} \approx 4.2426$ zum Erwartungswert von Klasse 2 bei $p = 2$, $\sqrt{(-3.005)^2 \cdot 2 + (2.995)^2} \approx 5.1990$ bei $p = 3$ und $\sqrt{(-3.005)^2 \cdot 5 + (2.995)^2 \cdot 5} \approx 9.4868$ bei $p = 10$ Dimensionen. Die Daten sind so konstruiert, dass genau in der Mitte des Datenstroms (Zeitpunkt $i = 2000.5$, welcher nicht angenommen wird) der euklidische Abstand beider Erwartungswertvektoren minimal $\sqrt{(-3)^2 p} = 3\sqrt{p}$ wäre.

d) „Vorbeilaufen" (gerade) der Erwartungswerte Auch in dieser Datensituation werden die Verteilungen so konstruiert, dass sich die Erwartungswertvektoren beider Klassen auf zwei parallelen Geraden in entgegengesetzter Richtung aufeinander zubewegen. Allerdings wird hier nur die Veränderung in der ersten Dimension betrachtet (vgl. Abbildungen 9.3 (d) und 9.4 (d)), die anderen $p-1$ Dimensionen der Erwartungswertvektoren sind zeitinvariant. Die so konstruierten Geraden liegen in der zweiten Dimension des p-dimensionalen Raumes um den Wert 3 voneinander entfernt. Die letzten $p-2$ Dimensionen sind zu jedem Zeitpunkt i in beiden Klassen konstant 0.

Formal werden die Geraden für die Klassen $c \in \{1, 2\}$, auf welchen sich die Erwartungswertvektoren bewegen, folgendermaßen definiert:

$$\begin{aligned} \boldsymbol{\mu}_i^{(1)} &= \boldsymbol{\beta}_0^{(1)} + \boldsymbol{\beta}_1^{(1)} i = -10\, \tilde{\boldsymbol{\beta}}_0^{(1)} + 0.005\, \tilde{\boldsymbol{\beta}}_1^{(1)} i, \\ \boldsymbol{\mu}_i^{(2)} &= \boldsymbol{\beta}_0^{(2)} + \boldsymbol{\beta}_1^{(2)} i = \boldsymbol{\beta}_0^{(2)} - 0.005\, \tilde{\boldsymbol{\beta}}_1^{(2)} i \end{aligned}$$

mit $\tilde{\boldsymbol{\beta}}_0^{(1)} = \boldsymbol{e}_1$, $\tilde{\boldsymbol{\beta}}_1^{(1)} = \boldsymbol{e}_1$, $\tilde{\boldsymbol{\beta}}_1^{(2)} = \boldsymbol{e}_1$ jeweils p-dimensionale Einheitsvektoren mit einer 1 in erster Dimension und

$$\boldsymbol{\beta}_0^{(2)} = \begin{cases} (10.005, -3)^T, & p = 2, \\ (10.005, -3, 0)^T, & p = 3, \\ (10.005, -3, 0, 0, 0, 0, 0, 0, 0, 0)^T, & p = 10 \end{cases}$$

der Intercept für Klasse 2.

Zum Zeitpunkt $i = 1$ sind die Erwartungswertvektoren gleich

$$\boldsymbol{\mu}_1^{(1)} = -9.995\,\boldsymbol{e}_1, \quad \boldsymbol{\mu}_1^{(2)} = (10, -3, \underbrace{0, \ldots, 0}_{p-2 \text{ Elemente}})^T$$

und zum Endzeitpunkt $i = 4000$ des Datenstroms sind sie durch

$$\boldsymbol{\mu}_{4000}^{(1)} = 10\,\boldsymbol{e}_1, \quad \boldsymbol{\mu}_{4000}^{(2)} = (-9.995, -3, \underbrace{0, \ldots, 0}_{p-2 \text{ Elemente}})^T$$

gegeben.

Wie bei der Datensituation c) liegt der Erwartungswertvektor von Klasse 1 zum Zeitpunkt $i = 2000$ im Ursprung $\boldsymbol{\mu}_{2000}^{(1)} = \mathbf{0}_p$. Zu diesem sowie dem folgenden Zeitpunkt $i = 2001$ ist der Erwartungswertvektor der Klasse 2 in erster Dimension absolut nur um 0.005 verschoben, sodass der euklidische Abstand zwischen den Zentren beider Verteilungen mit $\sqrt{(-0.005)^2 + 3^2} \approx 3$ hier am kleinsten ist. Zum nicht angenommenen Zeitpunkt $i = 2000.5$ wäre der euklidische Abstand zwischen beiden Klassen mit $\sqrt{3^2} = 3$ minimal. In dieser Datensituation ist dies unabhängig von der betrachteten Anzahl an Variablen bzw. der Dimension p, da die letzten $p - 2$ Dimensionen konstant 0 gesetzt werden.

e) Gradual Drift mit „Kreuzen“ der Erwartungswerte Beim gradual drift wird eine datengenerierende Funktion nach und nach durch die andere ersetzt, das heißt es erfolgt ein „weicher“ Übergang von einer zur anderen Verteilung in jeder Klasse. In diesem Fall von $\mathcal{N}\left(\boldsymbol{\mu}_1^{(1)}, \boldsymbol{\Sigma}\right)$ zu $\mathcal{N}\left(\boldsymbol{\mu}_{4000}^{(1)}, \boldsymbol{\Sigma}\right)$ bzw. von $\mathcal{N}\left(\boldsymbol{\mu}_1^{(2)}, \boldsymbol{\Sigma}\right)$ zu $\mathcal{N}\left(\boldsymbol{\mu}_{4000}^{(2)}, \boldsymbol{\Sigma}\right)$, wobei die Erwartungswertvektoren für die zwei Verteilungen in jeder Klasse in Anlehnung an den ersten sowie den letzten Zeitpunkt der Situation b) gewählt werden:

$$\boldsymbol{\mu}_1^{(1)} = -10 \cdot \mathbf{1}_p, \quad \boldsymbol{\mu}_1^{(2)} = \begin{cases} 10 \cdot (1, -1)^T, & p = 2, \\ 10 \cdot (1, -1, -1)^T, & p = 3, \\ 10 \cdot (\underbrace{1, \ldots, 1}_{5 \text{ Elemente}}, \underbrace{-1, \ldots, -1}_{5 \text{ Elemente}})^T, & p = 10, \end{cases}$$

$$\boldsymbol{\mu}_{4000}^{(1)} = 10 \cdot \mathbf{1}_p, \quad \boldsymbol{\mu}_{4000}^{(2)} = \begin{cases} 10 \cdot (-1, 1)^T, & p = 2, \\ 10 \cdot (-1, 1, 1)^T, & p = 3, \\ 10 \cdot (\underbrace{-1, \ldots, -1}_{5 \text{ Elemente}}, \underbrace{1, \ldots, 1}_{5 \text{ Elemente}})^T, & p = 10. \end{cases}$$

Eine Beobachtung aus Klasse c resultiert zum Zeitpunkt i mit Wahrscheinlichkeit $\frac{n-i}{n-1}$ aus Konzept 1, also der Verteilung $\mathcal{N}\left(\boldsymbol{\mu}_1^{(c)}, \boldsymbol{\Sigma}\right)$ und mit Wahrscheinlichkeit $1 - \frac{n-i}{n-1} = \frac{i-1}{n-1}$ aus Konzept 2, also der Verteilung $\mathcal{N}\left(\boldsymbol{\mu}_{4000}^{(c)}, \boldsymbol{\Sigma}\right)$.

f) Gradual Drift mit „Austausch“ der Erwartungswerte In dieser Datensituation werden die Daten ähnlich erzeugt wie in e).

Jedoch werden die einzelnen Erwartungswertvektoren, zwischen denen ein langsamer Übergang erfolgt, anders gewählt:

$$\boldsymbol{\mu}_1^{(1)} = \begin{pmatrix} -2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_{4000}^{(1)} = \begin{pmatrix} 2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_1^{(2)} = \begin{pmatrix} 2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \boldsymbol{\mu}_{4000}^{(2)} = \begin{pmatrix} -2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Dadurch werden allmählich die beiden Klassen „vertauscht“, da die Ausgangsverteilung der Klasse 1 identisch zu jener von Klasse 2 zum Zeitpunkt $i = 4000$ ist und andersrum.

g) Sudden Drift Zur Repräsentation eines sudden drifts wird als Ausgangssituation die Idee der Bewegung der Erwartungswertvektoren auf einem Kreis mit Radius 2 im p-dimensionalen Raum wie in a) herangezogen. Es wird jedoch eine andere Geschwindigkeit und Stärke des Drifts betrachtet. Die Erwartungswertvektoren bewegen sich nicht in konstanter Geschwindigkeit immer um ein Grad pro Zeitpunkt in gleicher Richtung, sondern es werden drei abrupte Verschiebungen der Verteilungen um jeweils 180, 90 und 180 Grad zu den Zeitpunkten $t = 1001, 2001, 3001$ modelliert (vgl. Abbildung 9.3 (g)). Es werden also vier verschiedene Verteilungen der Klassen über die Zeit betrachtet, wobei sich die Verteilungen bezüglich der Lage jeweils plötzlich ändern:

$$\boldsymbol{\mu}_i^{(1)} = \begin{cases} (2, 0, \ldots, 0)^T, & i = 1, \ldots, 1000, \\ (-2, 0, \ldots, 0)^T, & i = 1001, \ldots, 2000, \\ (0, -2, \ldots, 0)^T, & i = 2001, \ldots, 3000, \\ (0, 2, \ldots, 0)^T, & i = 3001, \ldots, 4000, \end{cases}$$

$$\boldsymbol{\mu}_i^{(2)} = \begin{cases} (-2, 0, \ldots, 0)^T, & i = 1, \ldots, 1000, \\ (2, 0, \ldots, 0)^T, & i = 1001, \ldots, 2000, \\ (0, 2, \ldots, 0)^T, & i = 2001, \ldots, 3000, \\ (0, -2, \ldots, 0)^T, & i = 3001, \ldots, 4000. \end{cases}$$

Es sei darauf hingewiesen, dass eine Bewegung um jeweils 180 Grad eine Vertauschung der Klassenzuordnung bedeutet, da die Erwartungswerte der zwei Klassen ausgetauscht werden.

h) ohne Drift Zuletzt wird zur Untersuchung des Verhaltens aller Methoden inklusive der entwickelten Erweiterung auf Datensätzen mit stabiler Verteilung auch ein Datenstrom ohne concept drift, also mit einer stabilen Verteilung über die Zeit konstruiert. Für diesen

Datenstrom wird dabei für jeden Zeitpunkt zufällig ($p^{(1)} = p^{(2)} = 0.5$) eine Beobachtung aus einer der beiden folgenden Verteilungen gezogen:

$$\boldsymbol{X}|Y = y_1 \sim \mathcal{N}\left(\begin{pmatrix}2\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}2 & 0 & 0\\0 & \ddots & 0\\0 & 0 & 2\end{pmatrix}\right), \quad \boldsymbol{X}|Y = y_2 \sim \mathcal{N}\left(\begin{pmatrix}-2\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}2 & 0 & 0\\0 & \ddots & 0\\0 & 0 & 2\end{pmatrix}\right).$$

9.3 Durchführung der Simulationsstudie

Zusätzlich zu den 4000 Beobachtungen für den Datenstrom wird zu jedem Zeitpunkt ein weiterer Testdatensatz aus 100 Beobachtungen simuliert, wobei die Beobachtungen ebenfalls den jeweiligen Verteilungen dieses Zeitpunktes folgen. Dieser Testdatensatz dient im Folgenden dazu die Prognosegüte der einzelnen Methoden auf den Datensätzen zu bestimmen und zu vergleichen.

Darüber hinaus wird die gesamte Simulation jeder einzelnen Datensituation jeweils 100 Mal wiederholt, um bei der Auswertung für jeden einzelnen Zeitpunkt einen mittleren Prognosefehler und die Varianz des Prognosefehlers bestimmen zu können, insgesamt also eine empirische Verteilung des Prognosefehlers zu jedem Zeitpunkt des Datenstroms erfassen zu können. Zudem wird durch Betrachtung des Mittelwertes die Varianz des Schätzers für den Prognosefehler reduziert. Alle betrachteten Methoden und ihre jeweiligen Erweiterungen werden auf allen in Abschnitt 9.1 und 9.2 beschriebenen Datensituationen angewandt.

9.4 Wahl der Parametereinstellungen für Methoden

Zusätzlich zu dem Raum der betrachteten Datensituationen und den Arten von concept drift (Abschnitt 9.1 und 9.2) müssen für die Simulationsstudie die Parametereinstellungen der betrachteten Methoden festgelegt werden. Alle betrachteten Methoden wurden in Kapitel 4 und die Erweiterungen in Kapitel 7 theoretisch beschrieben. Tabelle 9.1 umfasst eine Erklärung der jeweiligen einstellbaren Parameter aller betrachteten Methoden und Erweiterungen sowie ihre (möglichen) Wertebereiche.

Mittels des Parameters n_{init} kann gesteuert werden, anhand wie vieler Beobachtungen die jeweilige Methode für Online Diskriminanzanalyse initialisiert wird. In kleineren Vorsimulationen zeigte sich jedoch, dass die Initialisierung kaum einen Einfluss auf den Verlauf des Prognosefehlers hat. Aufgrund verschiedener Anzahlen an Beobachtungen zur Initialisierung der Methoden variiert lediglich die Prognosegüte der ersten paar Beobachtungen im Datenstrom sehr stark. Daher sollte sowieso stets eine Einpendelphase betrachtet werden, bis das Modell durch eine genügend hohe Anzahl an Beobachtungen nach der Initialisierung aktualisiert wurde. In der Simulationsstudie werden daher alle Methoden auf der Basis von

Tabelle 9.1: Erklärung der Parameter der Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen.

Parameter	**Wertebereich**	**Erklärung/Inhaltliche Beschreibung**
n_{init}	$\mathbb{N}$	Anzahl der Beobachtungen für Initialisierung des Modells (DA).
λ_-	$0 \leq \lambda_- < \lambda_+ \leq 1$	Untere Grenze für Faktoren λ bei der *Online Diskriminanzanalyse mit adaptivem Vergessen.*
λ_+	$0 \leq \lambda_- < \lambda_+ \leq 1$	Obere Grenze für Faktoren λ bei der *Online Diskriminanzanalyse mit adaptivem Vergessen.*
α_{min}	$0 < \alpha_{\text{min}} < \alpha_{\text{max}}$ klein	Unterer Schwellenwert für die Schrittweite α bei der *Online Diskriminanzanalyse mit adaptivem Vergessen.*
α_{max}	$0 < \alpha_{\text{min}} < \alpha_{\text{max}}$ klein	Oberer Schwellenwert für die Schrittweite α bei der *Online Diskriminanzanalyse mit adaptivem Vergessen.*
$\tilde{\boldsymbol{\theta}}_0^{(0)}$	S. 82 ff. und S. 227	Startwerte für den *M-AF* Algorithmus bei der *Online Diskriminanzanalyse mit adaptivem Vergessen.*
$\tilde{\boldsymbol{\theta}}_0^{(c)},\ c = 1, \ldots, M$	S. 77 ff. und S. 227	Startwerte für den *G-AF* Algorithmus für jede Klasse bei der *Online Diskriminanzanalyse mit adaptivem Vergessen.*
$\tilde{\boldsymbol{\theta}}_0^{(P)}$	S. 86 ff. und S. 227	Startwerte für den Durchlauf des *G-AF* Algorithmus zur Aktualisierung der gepoolten Kovarianzmatrix bei der *Online Diskriminanzanalyse* (bei LDA) *mit adaptivem Vergessen* (*LDA-AF*).
λ	$(0, 1)$	Lernrate bei *OLDC*.
λ_{start}	$(0, 1)$	Initiale Lernrate für die ersten $2L$ Updates, falls eine adaptive Lernrate bei *OLDC* betrachtet wird.
L	$\mathbb{N}$	Fensterbreite für Update der adaptiven Lernrate bei *OLDC*.
N_{trend}	$\mathbb{N}$ (vgl. (7.2) auf S. 162)	Breite des Intervalls zur Anpassung eines lokalen linearen Regressionsmodells.
n_μ	$\mathbb{N}$	Anzahl der Mittelwerte für Initialisierung des linearen Trendmodells.

$n_{\text{init}} = 10$ Beobachtungen initialisiert, wenn der Datenstrom $p \in \{2, 3\}$ Dimensionen aufweist und auf Basis von $n_{\text{init}} = 20$ Beobachtungen, wenn der Datenstrom 10-dimensional ist. Hierbei ist zu beachten, dass *QDA-AF* und *LDA-AF* (vgl. Abschnitt 4.4) streng genommen keine Beobachtungen zur Initialisierung benötigen, sondern direkt datenunabhängig initialisiert werden können, weshalb bei diesen Methoden $n_{\text{init}} = 1$ betrachtet wird.

Die Parameter λ_- und λ_+ sind die Schwellenwerte für den Faktor λ bei der Online Diskriminanzanalyse mit adaptivem Vergessen (vgl. Formel (4.49) auf Seite 81). Wie in Anagnostopoulos et al. (2012, S. 146) werden die Werte auf $\lambda_- := 0.7$ und $\lambda_+ := 0.999$ festgesetzt. Analog sind α_{min} und α_{max} Schwellenwerte für die Schrittweite α (vgl. Formel (4.51) auf Seite 82) beim Gradientenabstiegsverfahren. Anagnostopoulos et al. (2012, S. 146) schlagen $\alpha_{\text{min}} := 10^{-8}$ und $\alpha_{\text{max}} := 10^{-6}$ vor.

Die Startwerte $\tilde{\boldsymbol{\theta}}_0^{(0)}$, $\tilde{\boldsymbol{\theta}}_0^{(c)}$, $c = 1, \ldots, M$, und $\tilde{\boldsymbol{\theta}}_0^{(P)}$ für die Parameter der Online QDA bzw. Online LDA mit adaptivem Vergessen werden ebenfalls analog zu den Vorschlägen von Anagnostopoulos et al. (2012) festgesetzt (vgl. auch Abschnitt 4.4). Die Startwerte $\tilde{\boldsymbol{\theta}}_0^{(0)}$ für den *M-AF* Algorithmus (Algorithmus 2, Seite 83) sehen dabei folgendermaßen aus:

- $\tilde{P}_0^{(1)} = \frac{1}{M}, \ldots, \tilde{P}_0^{(M)} = \frac{1}{M}$,
- $N_0^{(0)} = 1$,
- $\alpha_0^{(0)}$ zufällig aus 100 gleichmäßigen Gitterpunkten aus $[\alpha_{\min}, \alpha_{\max}] = [10^{-8}, 10^{-6}]$,
- $\lambda_0^{(0)}$ zufällig aus 100 gleichmäßigen Gitterpunkten aus $[\lambda_-, \lambda_+] = [0.7, 0.999]$,
- $\left(\tilde{P}_0^{(1)}\right)' = 0, \ldots, \left(\tilde{P}_0^{(M)}\right)' = 0$,
- $\left(N_0^{(0)}\right)' = 0$,
- $\left(J_0^{(0)}\right)' = 0$.

Die Startwerte $\tilde{\boldsymbol{\theta}}_0^{(P)}$ und $\tilde{\boldsymbol{\theta}}_0^{(c)}$, $c = 1, \ldots, M$, für den *G-AF* Algorithmus (Algorithmus 1, Seite 82) werden folgendermaßen gewählt:

- $\tilde{\boldsymbol{m}}_{n_0}^{(P)} = \mathbf{0}_p$,
- $d_0^{(P)} = -1000$,
- $\boldsymbol{G}_0^{(P)} = 1000\, \boldsymbol{I}_{p\times p}$,
- $\tilde{\boldsymbol{\Sigma}}_0^{(P)} = \mathbf{0}_{p\times p}$,
- $\tilde{\boldsymbol{\Pi}}_0^{(P)} = \mathbf{0}_{p\times p}$,
- $\left(\tilde{\boldsymbol{m}}_{n_0}^{(P)}\right)' = \mathbf{0}_p$,
- $\left(d_0^{(P)}\right)' = 0$,
- $\left(\boldsymbol{G}_0^{(P)}\right)' = \mathbf{0}_{p\times p}$,
- $\left(\tilde{\boldsymbol{\Sigma}}_0^{(P)}\right)' = \mathbf{0}_{p\times p}$,
- $\left(\tilde{\boldsymbol{\Pi}}_0^{(P)}\right)' = \mathbf{0}_{p\times p}$,
- $\lambda_0^{(P)}$ zufällig aus 100 gleichmäßigen Gitterpunkten des Intervalls $[\lambda_-, \lambda_+] = [0.7, 0.999]$,
- $N_0^{(P)} = 0$,
- $\left(N_0^{(P)}\right)' = 0$,
- $\left(J_0^{(P)}\right)' = 0$,
- $\alpha_0^{(P)}$ zufällig aus 100 gleichmäßigen Gitterpunkten des Intervalls $[\alpha_{\min}, \alpha_{\max}] = [10^{-8}, 10^{-6}]$.

- $\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)} = \mathbf{0}_p$,
- $d_0^{(c)} = -1000$,
- $\boldsymbol{G}_0^{(c)} = 1000\, \boldsymbol{I}_{p\times p}$,
- $\tilde{\boldsymbol{\Sigma}}_0^{(c)} = \mathbf{0}_{p\times p}$,
- $\tilde{\boldsymbol{\Pi}}_0^{(c)} = \mathbf{0}_{p\times p}$,
- $\left(\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)}\right)' = \mathbf{0}_p$,
- $\left(d_0^{(c)}\right)' = 0$,
- $\left(\boldsymbol{G}_0^{(c)}\right)' = \mathbf{0}_{p\times p}$,
- $\left(\tilde{\boldsymbol{\Sigma}}_0^{(c)}\right)' = \mathbf{0}_{p\times p}$,
- $\left(\tilde{\boldsymbol{\Pi}}_0^{(c)}\right)' = \mathbf{0}_{p\times p}$,
- $\lambda_{(n_0^{(c)})}^{(c)}$ zufällig aus 100 gleichmäßigen Gitterpunkten des Intervalls $[\lambda_-, \lambda_+] = [0.7, 0.999]$,
- $N_0^{(c)} = 0$,
- $\left(N_0^{(c)}\right)' = 0$,
- $\left(J_0^{(c)}\right)' = 0$,
- $\alpha_0^{(c)}$ zufällig aus 100 gleichmäßigen Gitterpunkten des Intervalls $[\alpha_{\min}, \alpha_{\max}] = [10^{-8}, 10^{-6}]$.

Bei der Methode *OLDC* (vgl. Abschnitt 4.3) wird eine Lernrate λ benötigt. Um einen möglichst großen Bereich des Definitionsbereichs $\lambda \in (0, 1)$ gleichmäßig abzudecken, werden bei der Simulationsstudie die Werte $\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ betrachtet. Bei *OLDC* mit adaptiver Lernrate wird der eingeschränkte Bereich der Startwerte $\lambda_{\text{start}} \in \{0.1, 0.5, 0.9\}$

gewählt, da alle Kombinationen mit den Werten für die Fensterbreite L zur Adaption der Lernrate betrachtet werden müssen und die Anzahl an Parameterkombinationen somit schnell steigt. Zudem werden durch den gewählten Bereich kleine, mittlere und hohe Lernraten als Startwerte betrachtet. Es kann somit analysiert werden, wie gut die Adaption der Lernrate bei unterschiedlichen Bereichen des Startwertes funktioniert. Der Parameter L definiert die Fenstergröße, d. h. die Anzahl der L vorangegangenen Beobachtungen im Datenstrom, die zur Optimierung und schrittweisen Anpassung der Lernrate herangezogen werden (vgl. Seite 75). Hier werden in der Simulationsstudie die Werte $L \in \{5, 20, 50\}$ betrachtet, um verschieden große Zeitfenster abzudecken und zu analysieren.

Bei den Erweiterungen der Methoden durch zusätzliche Integration und regelmäßiger Aktualisierung eines lokalen linearen Trendmodells zur Modellierung und Prognose eines möglichen Drifts der Erwartungswerte der Klassen werden zusätzliche Parameter benötigt (vgl. Abschnitt 7.3). Die Anpassung der lokalen linearen Regressionsmodelle erfolgt immer jeweils anhand der letzten $n_{\text{trend}}^{(c)}$ aktualisierten Mittelwerte des Fensters der Breite N_{trend} (vgl. (7.2) auf Seite 162) im Datenstrom. Es wird also ein gleitendes Fenster betrachtet. Dadurch sollen auch höher dimensionale Trends der Erwartungswertvektoren der Klassen linear approximiert werden können. In der Simulationsstudie werden für die meisten Datensituationen Fenster von $N_{\text{trend}} \in \{10, 20, 50, 100, 200, 300\}$ Aktualisierungen des Diskriminanzanalysemodells betrachtet. Da die Datenströme *STAGGER* und *moving hyperplane* (vgl. Abschnitt 9.1) aus weniger Beobachtungen bestehen, wird bei diesen zwischen den Werten $N_{\text{trend}} \in \{2, 3, 5, 10, 20, 50, 100\}$ variiert.

Im Datenstrom wird somit auch bei den erweiterten Methoden zunächst nur das Diskriminanzanalysemodell mit jeder neuen Beobachtung aktualisiert. Die Initialisierung der lokalen linearen Regressionsmodelle erfolgt im Datenstrom erst nach n_μ Aktualisierungen und demnach auf Basis der aktualisierten Mittelwerte der Klassen aus diesen ersten n_μ Modell-Updates. Dabei wird $n_\mu = N_{\text{trend}}$ gewählt.

Insgesamt resultieren aus dieser Wahl der Werte für die freien Parameter folgende Anzahlen an Parameterkombinationen für die Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen, die in der Simulationsstudie für jeden Datensatz betrachtet werden:

Methode	**Anzahl Parameterkombinationen für Situationen aus** Abschnitt 9.1	9.2
Sequential Incremental LDA	1	1
OLDC mit fester Lernrate	5	5
OLDC mit adaptiver Lernrate	9	9
Online QDA mit adaptivem Vergessen	1	1
Online LDA mit adaptivem Vergessen	1	1
Erweiterung: Sequential Incremental LDA mit linearem Modell	7	6
Erweiterung: OLDC mit fester Lernrate mit linearem Modell	35	30
Erweiterung: OLDC mit adaptiver Lernrate mit linearem Modell	63	54
Erweiterung: Online QDA mit adaptivem Vergessen mit linearem Modell	7	6
Erweiterung: Online LDA mit adaptivem Vergessen mit linearem Modell	7	6

In folgender Tabelle 9.2 auf der nächsten Seite sind alle einstellbaren Parameter für die betrachteten Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen zusammengefasst. Die grau markierten Parameter werden in der Simulationsstudie nicht variiert, sondern wie oben beschrieben auf einen festen oder zufällig gewählten Wert gesetzt.

9.5 Implementierung

Die gesamte Simulationsstudie – inklusive Erzeugung der Daten – wurde in der Statistiksoftware **R** (R Core Team, 2019) durchgeführt. Bei der Erzeugung der Datensätze wurde dabei das **R**-Paket `mvtnorm` (Genz und Bretz, 2009; Genz et al., 2019) verwendet, um Zufallszahlen aus der multivariaten Normalverteilung zu generieren. Die Bayesfehler der Datensituationen wurden mithilfe des **R**-Paketes `cubature` (Narasimhan et al., 2018) durch numerische Integration der Dichtefunktionen bestimmt. Weiterhin wurden Funktionen der **R**-Pakete `gtools` (Warnes et al., 2018), `biglm` (Lumley, 2013) und `MASS` (Venables und Ripley, 2002) innerhalb der Funktionen zur Erstellung der Datensituationen, der Anwendung der Methoden für Online Diskriminanzanalyse oder ihrer Erweiterungen sowie der Ausführung der Berechnung der Ergebnisse verwendet.

Für die parallele Berechnung auf dem HPC Rechencluster der Fakultät Statistik sowie des ITMC (LIDO2 und LIDO3) wurde Funktionalität des **R**-Paketes `BatchJobs` (Bischl et al., 2015) herangezogen. Die meisten Grafiken wurden in **R** erstellt und mithilfe des **R**-Paketes `tikzDevice` (Sharpsteen und Bracken, 2018) in TikZ Code umgewandelt, bevor sie in das LaTeX-Dokument eingebettet wurden.

Tabelle 9.2: Einstellbare Parameter der Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen. Grau sind jene Parameter, die nicht variiert werden.

Methode	**Parameter der Methode**	
Sequential Incremental LDA	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
OLDC mit fester Lernrate	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
	λ	($\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$)
OLDC mit adaptiver Lernrate	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
	λ_{start}	($\lambda_{\text{start}} \in \{0.1, 0.5, 0.9\}$)
	L	($L \in \{5, 20, 50\}$)
Online QDA mit adaptivem Vergessen	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
	λ_-, λ_+	($\lambda_- = 0.7$, $\lambda_+ = 0.999$)
	$\alpha_{\min}$, $\alpha_{\max}$	($\alpha_{\min} = 10^{-8}$, $\alpha_{\max} = 10^{-6}$)
	$\tilde{\boldsymbol{\theta}}_0^{(0)} :=$	$\left\{ \tilde{P}_0^{(1)}, \ldots, \tilde{P}_0^{(M)}, \left(\tilde{P}_0^{(1)}\right)', \ldots, \left(\tilde{P}_0^{(M)}\right)', \lambda_0^{(0)}, N_0^{(0)}, \left(N_0^{(0)}\right)' J_0^{(0)}, \left(J_0^{(0)}\right)', \alpha_0^{(0)} \right\}$
	$\tilde{\boldsymbol{\theta}}_0^{(c)} :=$	$\left\{ \tilde{\boldsymbol{m}}^{(c)}_{n_0^{(c)}}, d_0^{(c)}, \boldsymbol{G}_0^{(c)}, \tilde{\boldsymbol{\Sigma}}_0^{(c)}, \tilde{\boldsymbol{\Pi}}_0^{(c)}, \left(\tilde{\boldsymbol{m}}^{(c)}_{n_0^{(c)}}\right)', \left(d_0^{(c)}\right)', \left(\boldsymbol{G}_0^{(c)}\right)', \left(\tilde{\boldsymbol{\Sigma}}_0^{(c)}\right)', \left(\tilde{\boldsymbol{\Pi}}_0^{(c)}\right)', \lambda^{(c)}_{(n_0^{(c)})}, N_0^{(c)}, \left(N_0^{(c)}\right)', J_0^{(c)}, \left(J_0^{(c)}\right)', \alpha_0^{(c)} \right\}, \quad c = 1, \ldots, M$
Online LDA mit adaptivem Vergessen	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
	λ_-, λ_+	($\lambda_- = 0.7$, $\lambda_+ = 0.999$)
	$\alpha_{\min}$, $\alpha_{\max}$	($\alpha_{\min} = 10^{-8}$, $\alpha_{\max} = 10^{-6}$)
	$\tilde{\boldsymbol{\theta}}_0^{(0)} :=$	$\left\{ \tilde{P}_0^{(1)}, \ldots, \tilde{P}_0^{(M)}, \left(\tilde{P}_0^{(1)}\right)', \ldots, \left(\tilde{P}_0^{(M)}\right)', \lambda_0^{(0)}, N_0^{(0)}, \left(N_0^{(0)}\right)' J_0^{(0)}, \left(J_0^{(0)}\right)', \alpha_0^{(0)} \right\}$
	$\tilde{\boldsymbol{\theta}}_0^{(c)} :=$	$\left\{ \tilde{\boldsymbol{m}}^{(c)}_{n_0^{(c)}}, d_0^{(c)}, \boldsymbol{G}_0^{(c)}, \tilde{\boldsymbol{\Sigma}}_0^{(c)}, \tilde{\boldsymbol{\Pi}}_0^{(c)}, \left(\tilde{\boldsymbol{m}}^{(c)}_{n_0^{(c)}}\right)', \left(d_0^{(c)}\right)', \left(\boldsymbol{G}_0^{(c)}\right)', \left(\tilde{\boldsymbol{\Sigma}}_0^{(c)}\right)', \left(\tilde{\boldsymbol{\Pi}}_0^{(c)}\right)', \lambda^{(c)}_{(n_0^{(c)})}, N_0^{(c)}, \left(N_0^{(c)}\right)', J_0^{(c)}, \left(J_0^{(c)}\right)', \alpha_0^{(c)} \right\}, \quad c = 1, \ldots, M$
	$\tilde{\boldsymbol{\theta}}_0^{(P)} :=$	$\left\{ \tilde{\boldsymbol{m}}_{n_0}^{(P)}, d_0^{(P)}, \boldsymbol{G}_0^{(P)}, \tilde{\boldsymbol{\Sigma}}_0^{(P)}, \tilde{\boldsymbol{\Pi}}_0^{(P)}, \left(\tilde{\boldsymbol{m}}_{n_0}^{(P)}\right)', \left(d_0^{(P)}\right)', \left(\boldsymbol{G}_0^{(P)}\right)', \left(\tilde{\boldsymbol{\Sigma}}_0^{(P)}\right)', \left(\tilde{\boldsymbol{\Pi}}_0^{(P)}\right)', \lambda_0^{(P)}, N_0^{(P)}, \left(N_0^{(P)}\right)', J_0^{(P)}, \left(J_0^{(P)}\right)', \alpha_0^{(P)} \right\}$
Erweiterung: Sequential Incremental LDA mit linearem Modell	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
	N_{trend}	($N_{\text{trend}} \in \{10, 20, 50, 100, 200, 300\}$ bzw. $N_{\text{trend}} \in \{2, 3, 5, 10, 20, 50, 100\}$)
	n_μ	($n_\mu = N_{\text{trend}}$)

Fortsetzung auf der nächsten Seite

Methode	**Parameter der Methode**	
Erweiterung: OLDC mit fester Lernrate mit linearem Modell	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
	λ	($\lambda \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$)
	N_{trend}	($N_{\text{trend}} \in \{10, 20, 50, 100, 200, 300\}$ bzw. $N_{\text{trend}} \in \{2, 3, 5, 10, 20, 50, 100\}$)
	n_μ	($n_\mu = N_{\text{trend}}$)
Erweiterung: OLDC mit adaptiver Lernrate mit linearem Modell	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
	λ_{start}	($\lambda_{\text{start}} \in \{0.1, 0.5, 0.9\}$)
	L	($L \in \{5, 20, 50\}$)
	N_{trend}	($N_{\text{trend}} \in \{10, 20, 50, 100, 200, 300\}$ bzw. $N_{\text{trend}} \in \{2, 3, 5, 10, 20, 50, 100\}$)
	n_μ	($n_\mu = N_{\text{trend}}$)
Erweiterung: Online QDA mit adaptivem Vergessen mit linearem Modell	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
	λ_-, λ_+	($\lambda_- = 0.7$, $\lambda_+ = 0.999$)
	$\alpha_{\min}$, $\alpha_{\max}$	($\alpha_{\min} = 10^{-8}$, $\alpha_{\max} = 10^{-6}$)
	$\tilde{\boldsymbol{\theta}}_0^{(0)}$	$:= \Big\{\tilde{P}_0^{(1)}, \ldots, \tilde{P}_0^{(M)}, \left(\tilde{P}_0^{(1)}\right)', \ldots, \left(\tilde{P}_0^{(M)}\right)', \lambda_0^{(0)}, N_0^{(0)}, \left(N_0^{(0)}\right)' J_0^{(0)}, \left(J_0^{(0)}\right)', \alpha_0^{(0)}\Big\}$
	$\tilde{\boldsymbol{\theta}}_0^{(c)}$	$:= \Big\{\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)}, d_0^{(c)}, \boldsymbol{G}_0^{(c)}, \tilde{\boldsymbol{\Sigma}}_0^{(c)}, \tilde{\boldsymbol{\Pi}}_0^{(c)}, \left(\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)}\right)', \left(d_0^{(c)}\right)', \left(\boldsymbol{G}_0^{(c)}\right)', \left(\tilde{\boldsymbol{\Sigma}}_0^{(c)}\right)', \left(\tilde{\boldsymbol{\Pi}}_0^{(c)}\right)', \lambda_{(n_0^{(c)})}^{(c)}, N_0^{(c)}, \left(N_0^{(c)}\right)', J_0^{(c)}, \left(J_0^{(c)}\right)', \alpha_0^{(c)}\Big\}, \quad c = 1, \ldots, M$
	N_{trend}	($N_{\text{trend}} \in \{10, 20, 50, 100, 200, 300\}$ bzw. $N_{\text{trend}} \in \{2, 3, 5, 10, 20, 50, 100\}$)
	n_μ	($n_\mu = N_{\text{trend}}$)
Erweiterung: Online LDA mit adaptivem Vergessen mit linearem Modell	n_{init}	($n_{\text{init}} = 10$ (bei $p = 2$ und $p = 3$) bzw. $n_{\text{init}} = 20$ (bei $p = 10$))
	λ_-, λ_+	($\lambda_- = 0.7$, $\lambda_+ = 0.999$)
	$\alpha_{\min}$, $\alpha_{\max}$	($\alpha_{\min} = 10^{-8}$, $\alpha_{\max} = 10^{-6}$)
	$\tilde{\boldsymbol{\theta}}_0^{(0)}$	$:= \Big\{\tilde{P}_0^{(1)}, \ldots, \tilde{P}_0^{(M)}, \left(\tilde{P}_0^{(1)}\right)', \ldots, \left(\tilde{P}_0^{(M)}\right)', \lambda_0^{(0)}, N_0^{(0)}, \left(N_0^{(0)}\right)' J_0^{(0)}, \left(J_0^{(0)}\right)', \alpha_0^{(0)}\Big\}$
	$\tilde{\boldsymbol{\theta}}_0^{(c)}$	$:= \Big\{\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)}, d_0^{(c)}, \boldsymbol{G}_0^{(c)}, \tilde{\boldsymbol{\Sigma}}_0^{(c)}, \tilde{\boldsymbol{\Pi}}_0^{(c)}, \left(\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)}\right)', \left(d_0^{(c)}\right)', \left(\boldsymbol{G}_0^{(c)}\right)', \left(\tilde{\boldsymbol{\Sigma}}_0^{(c)}\right)', \left(\tilde{\boldsymbol{\Pi}}_0^{(c)}\right)', \lambda_{(n_0^{(c)})}^{(c)}, N_0^{(c)}, \left(N_0^{(c)}\right)', J_0^{(c)}, \left(J_0^{(c)}\right)', \alpha_0^{(c)}\Big\}, \quad c = 1, \ldots, M$
	$\tilde{\boldsymbol{\theta}}_0^{(P)}$	$:= \Big\{\tilde{\boldsymbol{m}}_{n_0}^{(P)}, d_0^{(P)}, \boldsymbol{G}_0^{(P)}, \tilde{\boldsymbol{\Sigma}}_0^{(P)}, \tilde{\boldsymbol{\Pi}}_0^{(P)}, \left(\tilde{\boldsymbol{m}}_{n_0}^{(P)}\right)', \left(d_0^{(P)}\right)', \left(\boldsymbol{G}_0^{(P)}\right)', \left(\tilde{\boldsymbol{\Sigma}}_0^{(P)}\right)', \left(\tilde{\boldsymbol{\Pi}}_0^{(P)}\right)', \lambda_0^{(P)}, N_t^{(P)}, \left(N_t^{(P)}\right)', J_t^{(P)}, \left(J_t^{(P)}\right)', \alpha_0^{(P)}\Big\}$
	N_{trend}	($N_{\text{trend}} \in \{10, 20, 50, 100, 200, 300\}$ bzw. $N_{\text{trend}} \in \{2, 3, 5, 10, 20, 50, 100\}$)
	n_μ	($n_\mu = N_{\text{trend}}$)

9.6 Ergebnisse der Simulationsstudie

Im Folgenden werden die Ergebnisse der Simulationsstudie für alle betrachteten Datensituationen aus den Abschnitten 9.1 und 9.2 analysiert. Die Update-Methoden für Online Diskriminanzanalyse wurden erweitert mit dem Ziel der Verbesserung der Prognosegüte der Klassifikatoren zu jedem Zeitpunkt im Falle von concept drift.

Prognosegüte lässt sich dabei anhand verschiedener Maßzahlen charakterisieren bzw. messen. Bei Klassifikation liegt der Fokus in der Praxis meist auf Minimierung des Klassifikationsfehlers. Im Zusammenhang mit Datenströmen und Klassifikationsmethoden wie der Diskriminanzanalyse spielt dabei auch die Klassifikation neuer, „zukünftiger" Beobachtungen eine Rolle. Daher steht als Erstes im Fokus den *Prognosefehler* der aktualisierten Klassifikatoren zu jedem Zeitpunkt im Datenstrom zu verbessern bzw. durch die entwickelte Erweiterung im Vergleich zu den ursprünglichen Methoden zu verringern.

Zu dieser Analyse wird der mittlere Prognosefehler (aus 100 Simulationswiederholungen) zu jedem Zeitpunkt des Datenstroms für jede Datensituation für jede Methode und ihre Erweiterung gemeinsam grafisch dargestellt, sodass der Verlauf des mittleren Prognosefehlers über die Zeit verglichen werden kann. Der Prognosefehler wird dabei durch die relative Häufigkeit fehlklassifizierter Beobachtungen des Testdatensatzes des kommenden Zeitpunktes geschätzt, wenn die Beobachtungen dieses Testdatensatzes mit dem aktuellen Klassifikator im Datenstrom klassifiziert werden. Als ein einzelnes reduziertes Maß für einen quantitativen Vergleich aller Methoden und ihrer Erweiterungen auf jeder Datensituation dient der *durchschnittliche mittlere Prognosefehler* über den gesamten Zeitraum:

$$\frac{1}{n_{\text{Situation}} - n_{\text{init}}} \sum_{i=1+n_{\text{init}}}^{n_{\text{Situation}}} \Bigg(\underbrace{\frac{1}{100} \sum_{s=1}^{100} \frac{\#(\text{Fehlklassifikationen in Testdatensatz}_i^s)}{100}}_{\text{mittlerer Prognosefehler}} \Bigg),$$

wobei Testdatensatz_i^s den Testdatensatz aus 100 Beobachtungen der Verteilung zum Zeitpunkt i in Simulationsdurchlauf s beschreibt, $n_{\text{Situation}}$ die Anzahl der Beobachtungen (bzw. Zeitpunkte) in der jeweiligen Datensituation und n_{init} die Anzahl der Beobachtungen des Datenstroms zur Initialisierung der jeweiligen Update-Methode für die Diskriminanzanalyse.

Zur Beurteilung der Streuung wird zusätzlich die Varianz des Prognosefehlers (aus 100 Simulationswiederholungen) zu jedem Zeitpunkt des Datenstroms bestimmt und die *durchschnittliche Varianz des Prognosefehlers* über den gesamten Datenstrom ermittelt.

Die Methoden *ILDA* und *OLDC* mit fester und adaptiver Lernrate sowie ihre Erweiterungen werden dabei bei allen Datensituationen mit $p = 2$ und $p = 3$ sowie „moving hyperplane" und „STAGGER" auf den $n_{\text{init}} = 10$ ersten Beobachtungen initialisiert, bei den Datensituationen mit $p = 10$ auf den ersten $n_{\text{init}} = 20$ Beobachtungen, damit mehr

Beobachtungen als Variablen (Dimensionen) für die initiale Schätzung der Kovarianzmatrizen zur Verfügung stehen. Die Initialisierung von *QDA-AF* und *LDA-AF* funktioniert hingegen datenunabhängig, sodass direkt die Aktualisierungsschritte erfolgen können.

Die *Prognosegüte* lässt sich auf der anderen Seite auch direkt auf die Erwartungswertvektoren beziehen, deren Trend mittels der lokalen linearen Regressionsmodelle in der Erweiterung der Update-Methoden für die Diskriminanzanalyse modelliert wird. Dadurch können die Erwartungswertvektoren des kommenden Zeitpunktes mittels dieser Regressionsmodelle prognostiziert werden. In Kapitel 8 wurde die *Erwartungstreue* der verwendeten Schätzer für die Erwartungswertvektoren der einzelnen aktualisierten Klassifikatoren in den erweiterten Methoden zu jedem Zeitpunkt für einige Spezialfälle unter bestimmten Annahmen des concept drifts theoretisch bewiesen.

Die Ergebnisse der Simulationsstudie sollen diese theoretischen Ergebnisse der Spezialfälle untermauern sowie auch im allgemeinen Fall demonstrieren, dass die Erwartungswertvektoren durch die Erweiterungen der Methoden besser geschätzt werden können, sodass verbesserte Schätzer in die Klassifikatoren einfließen können. Als Maß zur Bewertung der Erwartungstreue wird dabei der euklidische Abstand d zwischen geschätzten und wahren Erwartungswertvektoren herangezogen. Hier ist zu beachten, dass der euklidische Abstand per Definition immer positiv ist. Daher wäre der mittlere euklidische Abstand über die 100 Simulationsdurchläufe „verzerrt“. Im Falle einer erwartungstreuen Schätzung der Erwartungswertvektoren der Klassen ist davon auszugehen, dass die einzelnen Schätzer annähernd symmetrisch um den wahren Erwartungswertvektor streuen. Dies kann vom euklidischen Abstand jedoch nicht abgebildet werden. Positive und negative Abweichungen würden sich im Mittel nicht aufheben, sondern aufsummieren, wodurch ein großer mittlerer euklidischer Abstand über alle 100 Simulationsdurchläufe resultieren würde. Der Mittelwert würde somit nicht die „Lage“ der wahren Verteilung repräsentieren.

Um dieses Problem zu minimieren, werden nicht zu jedem Zeitpunkt im Datenstrom die einzelnen euklidischen Abstände aus jeweils 100 Simulationsdurchläufen gemittelt. Stattdessen wird der euklidische Abstand d zwischen den mittleren aktualisierten Mittelwertvektoren über alle 100 Simulationsdurchläufe $\bar{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}$ (bei ursprünglichen Methoden) bzw. den mittleren prognostizierten Erwartungswertvektoren aus den lokalen linearen Regressionsmodellen $\bar{\hat{\boldsymbol{m}}}^{(c)}_{n_{t+1}^{(c)}}$ (bei erweiterten Methoden) und den wahren Erwartungswertvektoren $\boldsymbol{\mu}^{(c)}_{t+1}$ (in jeder der p Dimensionen) des folgenden Zeitpunktes $t+1$ als Distanzmaß herangezogen:

$$d(\bar{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}, \boldsymbol{\mu}^{(c)}_{t+1}) = \left|\left|\left(\bar{\boldsymbol{m}}^{(c)}_{n_t^{(c)}} - \boldsymbol{\mu}^{(c)}_{t+1}\right)\right|\right|_2 = \sqrt{\sum_{j=1}^{p}\left(\left[\bar{\boldsymbol{m}}^{(c)}_{n_t^{(c)}}\right]_j - \left[\boldsymbol{\mu}^{(c)}_{t+1}\right]_j\right)^2}$$
$$= \sqrt{\left(\bar{\boldsymbol{m}}^{(c)}_{n_t^{(c)}} - \boldsymbol{\mu}^{(c)}_{t+1}\right)^T\left(\bar{\boldsymbol{m}}^{(c)}_{n_t^{(c)}} - \boldsymbol{\mu}^{(c)}_{t+1}\right)} \qquad (9.2)$$

bzw.

$$d(\bar{\hat{\boldsymbol{m}}}^{(c)}_{n^{(c)}_{t+1}}, \boldsymbol{\mu}^{(c)}_{t+1}) = \left|\left|\left(\bar{\hat{\boldsymbol{m}}}^{(c)}_{n^{(c)}_{t+1}} - \boldsymbol{\mu}^{(c)}_{t+1}\right)\right|\right|_2 = \sqrt{\sum_{j=1}^{p}\left(\left[\bar{\hat{\boldsymbol{m}}}^{(c)}_{n^{(c)}_{t+1}}\right]_j - \left[\boldsymbol{\mu}^{(c)}_{t+1}\right]_j\right)^2}$$
$$= \sqrt{\left(\bar{\hat{\boldsymbol{m}}}^{(c)}_{n^{(c)}_{t+1}} - \boldsymbol{\mu}^{(c)}_{t+1}\right)^T\left(\bar{\hat{\boldsymbol{m}}}^{(c)}_{n^{(c)}_{t+1}} - \boldsymbol{\mu}^{(c)}_{t+1}\right)}. \quad (9.3)$$

Für eine Beurteilung der Streuung wird erneut die Varianz betrachtet. Hierzu sind jedoch wiederholte Beobachtungen des euklidischen Abstandes nötig, weshalb die Varianz der einzelnen euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren der einzelnen 100 Simulationsdurchläufe var $\left(d(\boldsymbol{m}^{(c)}_{n^{(c)}_t}, \boldsymbol{\mu}^{(c)}_{t+1})\right)$ bzw. var $\left(d(\hat{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t+1}}, \boldsymbol{\mu}^{(c)}_{t+1})\right)$ betrachtet wird.

Zum Vergleich aller Methoden sowie ihrer Erweiterungen wird dabei als weitere einzelne Maßzahl neben dem durchschnittlichen mittleren Prognosefehler über die Zeit der *durchschnittliche euklidische Abstand* zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit betrachtet. Dies ist in allen Datensituationen möglich, bei denen die Beobachtungen zu jedem Zeitpunkt aus einer festen und bekannten Verteilung simuliert werden und somit die wahren Erwartungswertvektoren bekannt sind. Ausgenommen davon sind daher „moving hyperplane“ und „STAGGER“ sowie die Datensituationen Gradual Drift mit „Kreuzen“ und Gradual Drift mit „Austausch“. In diesen Fällen konzentriert sich die Analyse der Ergebnisse auf die Auswertung der Prognosefehler.

Anders als beim Vergleich der durchschnittlichen mittleren Prognosefehler über die Zeit werden für die Bestimmung des durchschnittlichen euklidischen Abstandes alle Methoden auf einer Datensituation jeweils auf derselben Anzahl an Beobachtungen initialisiert. Alle Methoden werden daher auf den 2- und 3-dimensionalen Datensätzen auf den ersten $n_{\text{init}} = 10$ Beobachtungen initialisiert, auf den 10-dimensionalen Datensätzen auf den ersten $n_{\text{init}} = 20$ Beobachtungen. Dies erfolgt für einen fairen Vergleich der Ergebnisse für alle Methoden und ihre Erweiterungen. Bei *QDA-AF* und *LDA-AF* ist der euklidische Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren in den ersten Aktualisierungsschritten in manchen Datensituationen recht groß. Dies liegt daran, dass die Initialisierung datenunabhängig erfolgt. Liegen die wahren Erwartungswertvektoren somit weit entfernt von den „initialen“ Werten, dauert es ein paar Aktualisierungsschritte, bis die Erwartungswerte gut geschätzt werden. Werden nun anders als bei allen anderen Methoden von Beginn des Datenstroms Aktualisierungsschritte bei *QDA-AF* und *LDA-AF* betrachtet, fließen diese eventuell großen euklidischen Abstände in den Mittelwert ein und verhindern einen „fairen“ Vergleich mit den anderen Methoden. Daher werden hier für alle Methoden gleich viele Aktualisierungsschritte betrachtet.

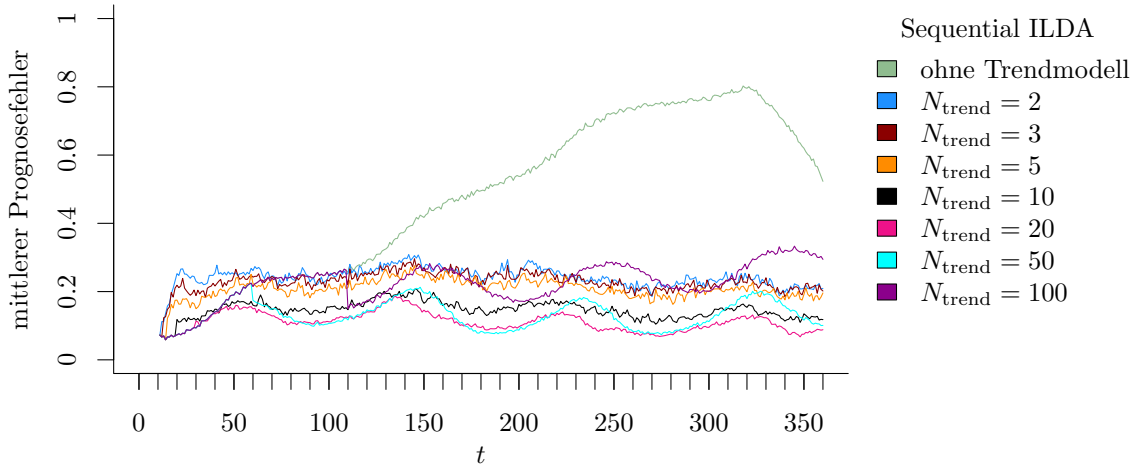
9.6.1 Moving hyperplane und STAGGER

Zunächst werden die Resultate der verschiedenen Methoden und ihrer Erweiterungen auf den Datensituationen **moving hyperplane** und **STAGGER** (vgl. Abschnitt 9.1) verglichen. In den folgenden Tabellen 9.3 und 9.4 sind die durchschnittlichen mittleren Prognosefehler über die Zeit inklusive der durchschnittlichen Varianz des Prognosefehlers über die Zeit für alle betrachteten Methoden für die beiden Datensituationen zusammengefasst. Die durchschnittliche Varianz charakterisiert dabei die Streuung des Prognosefehlers über die Zeit.

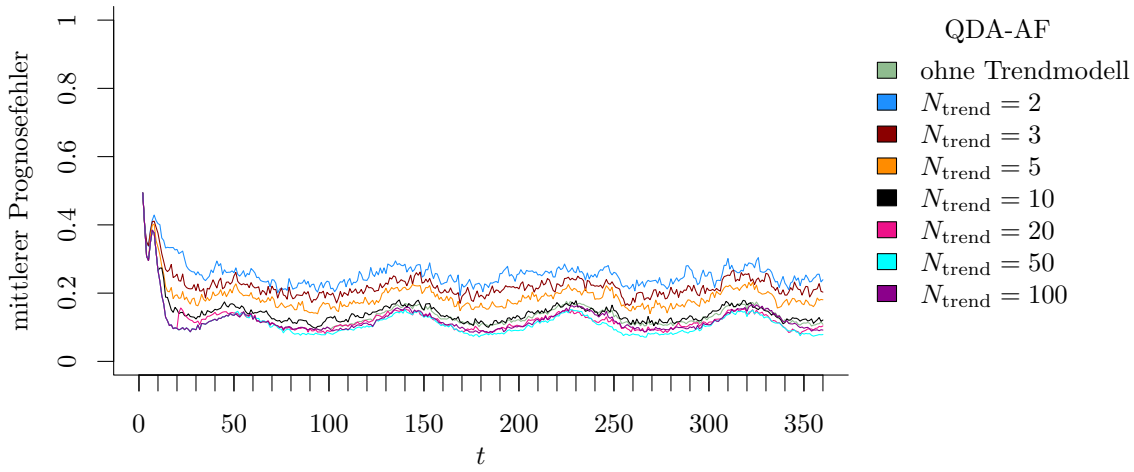
Die einzelnen mittleren Prognosefehler bzw. der Verlauf über die Zeit für den gesamten Datenstrom ist für alle betrachteten Methoden für die beiden Datensituationen in den folgenden Abbildungen 9.5–9.9 und 9.12–9.16 veranschaulicht. Dabei ist für jede Methode der Verlauf des Prognosefehlers innerhalb einer Grafik vergleichend dargestellt mit den Prognosefehlern bei Einbezug lokaler linearer Regressionsmodelle zur Modellierung und Prognose eines Trends der Erwartungswerte basierend auf einer verschiedenen Anzahl von $n_{\text{trend}}^{(c)}$ aktualisierten Mittelwerten, charakterisiert durch Fenster der Breite N_{trend} (vgl. Formel (7.2)).

Moving hyperplane Der erste Block (erste „Zeile“) in Tabelle 9.3 auf Seite 242 umfasst die Ergebnisse für die betrachteten Methoden ohne Erweiterungen durch ein Regressionsmodell. Hier ist bereits zu sehen, dass jene Methoden, welche für den Umgang mit concept drift entwickelt wurden (*QDA-AF*, *LDA-AF*, *OLDC fix* mit $\lambda > 0.5$ und *OLDC adaptive*) der standardmäßigen (adaptiven) Linearen Diskriminanzanalyse nach Fisher (*ILDA*) und der Kanonischen LDA (*OLDC fix* mit $\lambda = 0.5$) überlegen sind. Ohne jegliche Art von Gewichtung der aktualisierten Größen bei der Diskriminanzanalyse (*OLDC*) oder Gewichtung der Likelihood Terme (*QDA-AF* und *LDA-AF*) beträgt der durchschnittliche mittlere Prognosefehler hier fast 0.5. Dies wird auch anhand der Abbildungen 9.5 (a) und 9.6 (b) deutlich. Der Prognosefehler (grüne Kurve) steigt im Laufe der Zeit auf etwa 0.8 an, was bei zwei Klassen deutlich schlechter als der Prognosefehler bei zufälliger Klassenzuordnung ist.

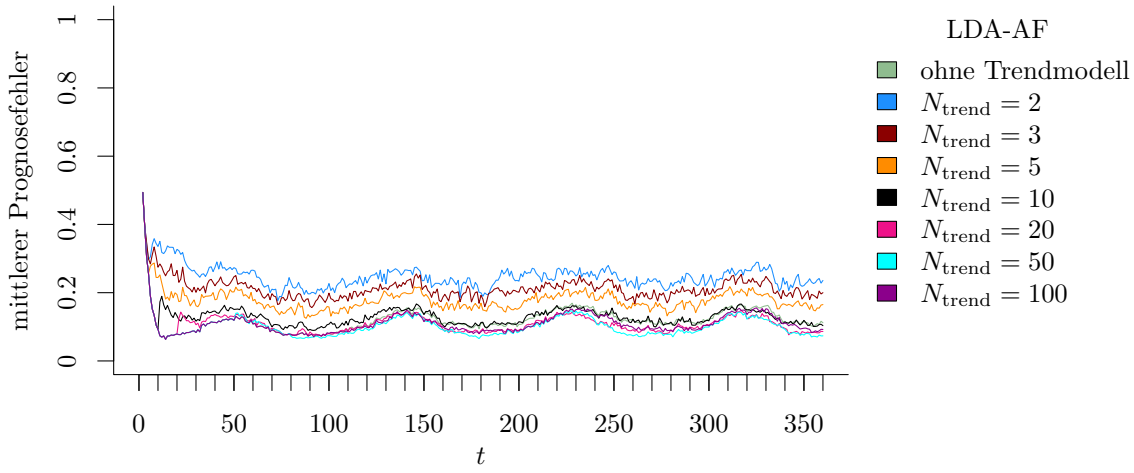
Bei den anderen Methoden (*QDA-AF*, *LDA-AF*, *OLDC fix* mit $\lambda = 0.9$ und *OLDC* mit adaptiver Lernrate) setzt sich die (grüne) Kurve nicht so deutlich von den anderen Kurven ab (vgl. Abbildungen 9.5–9.9). Anhand der Ergebnisse in Tabelle 9.3 wird jedoch deutlich, dass in fast allen Fällen (Spalten) der durchschnittliche mittlere Prognosefehler über die Zeit durch Einführung lokaler linearer Regressionsmodelle zur Modellierung und Prognose des Trends der Erwartungswerte bei geeigneter Wahl von N_{trend} deutlich reduziert werden kann. Mit wachsendem N_{trend} sinkt der durchschnittliche mittlere Prognosefehler für alle Methoden bis zu einem Wert von $N_{\text{trend}} = 20$ bzw. $N_{\text{trend}} = 50$, bevor er für $N_{\text{trend}} = 100$ wieder ansteigt. Die Prognosefehler der ursprünglichen Methoden werden jedoch bis $N_{\text{trend}} = 300$ nicht wieder überschritten.



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

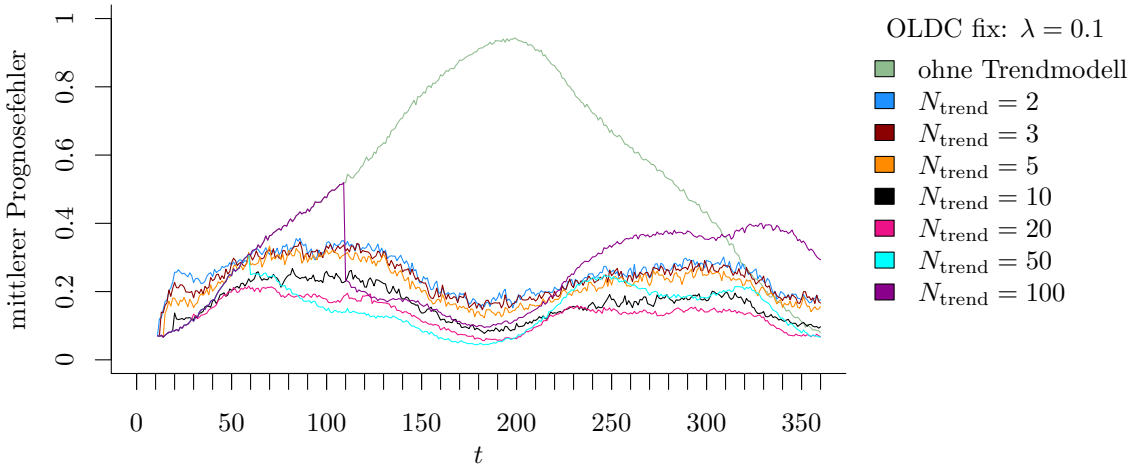


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

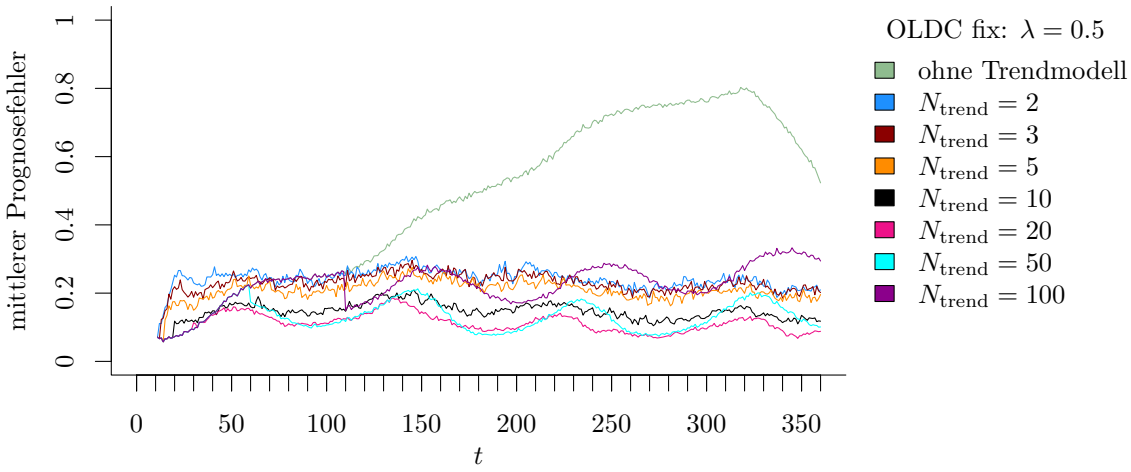


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

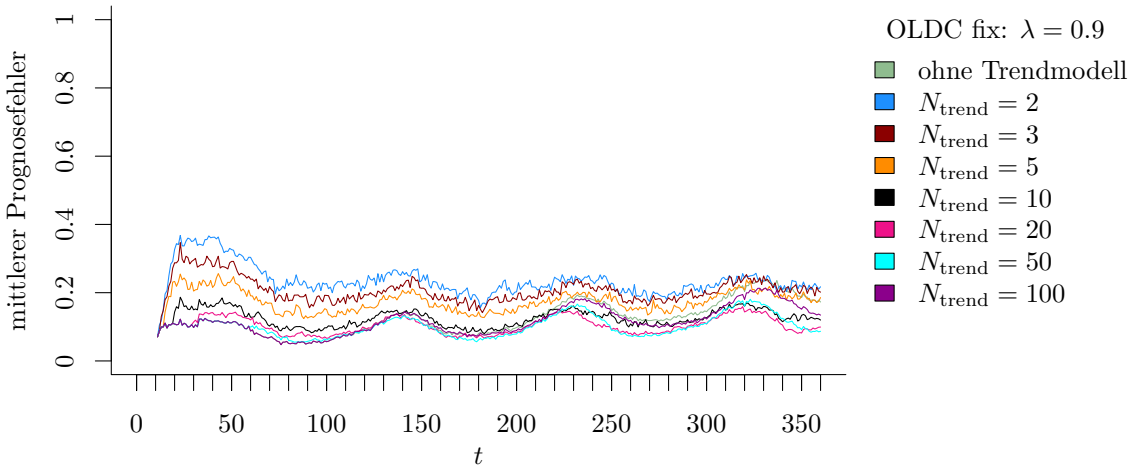
Abbildung 9.5: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **moving hyperplane**.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.

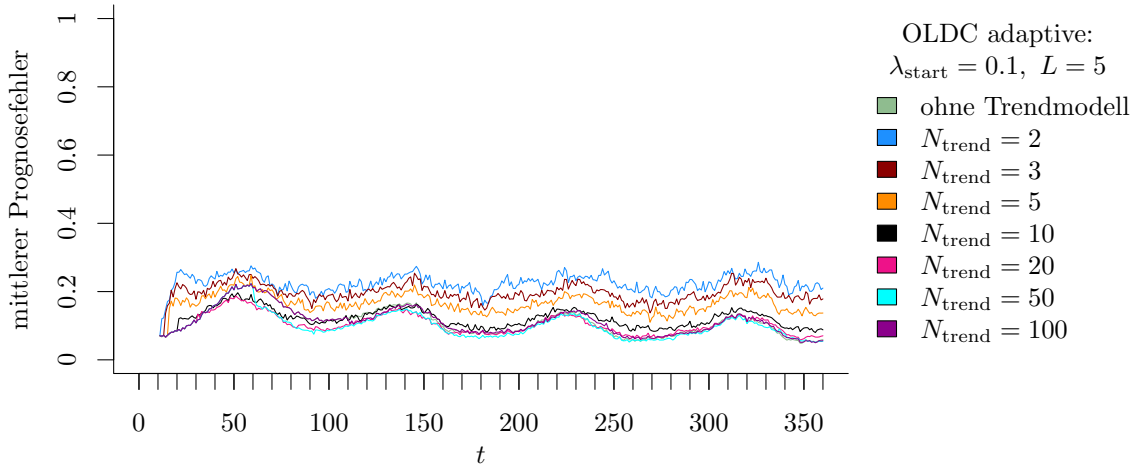


(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.

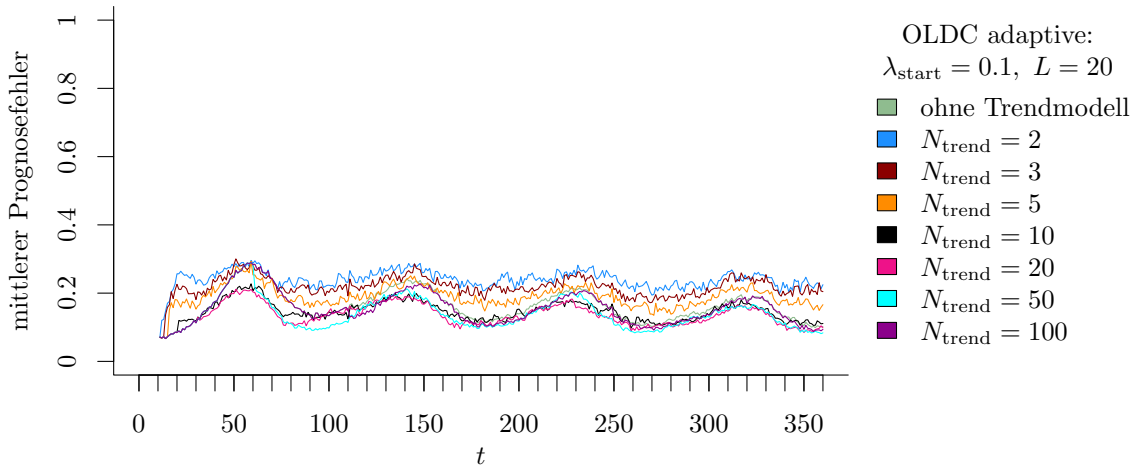


(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

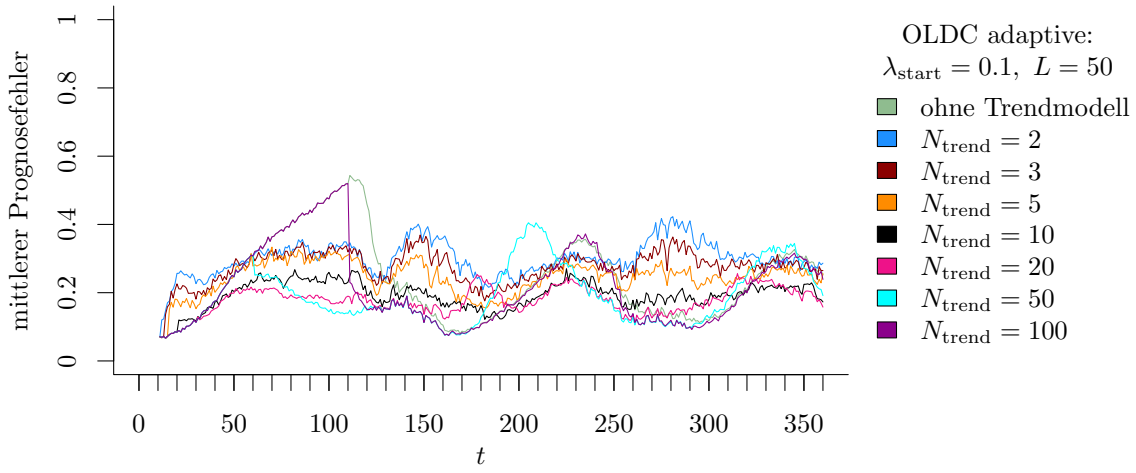
Abbildung 9.6: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **moving hyperplane**.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

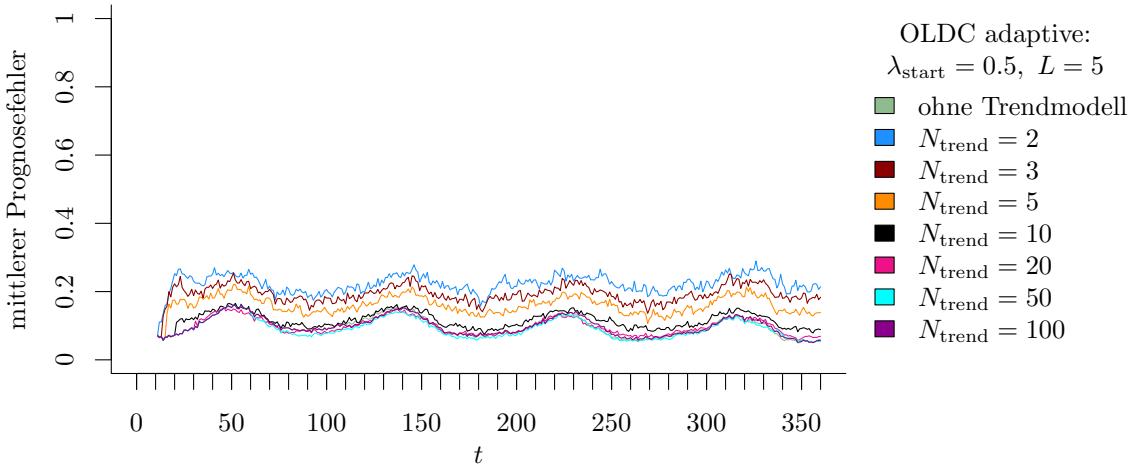


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

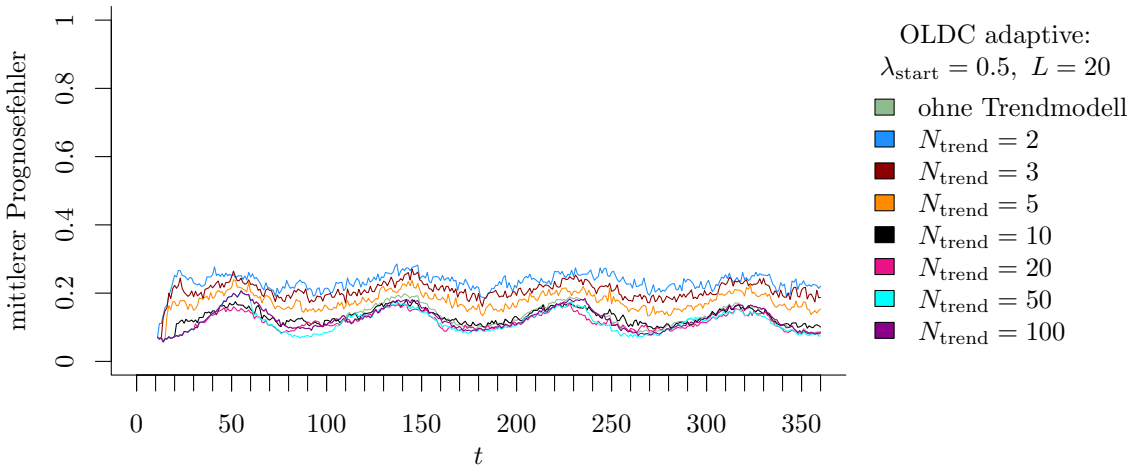


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

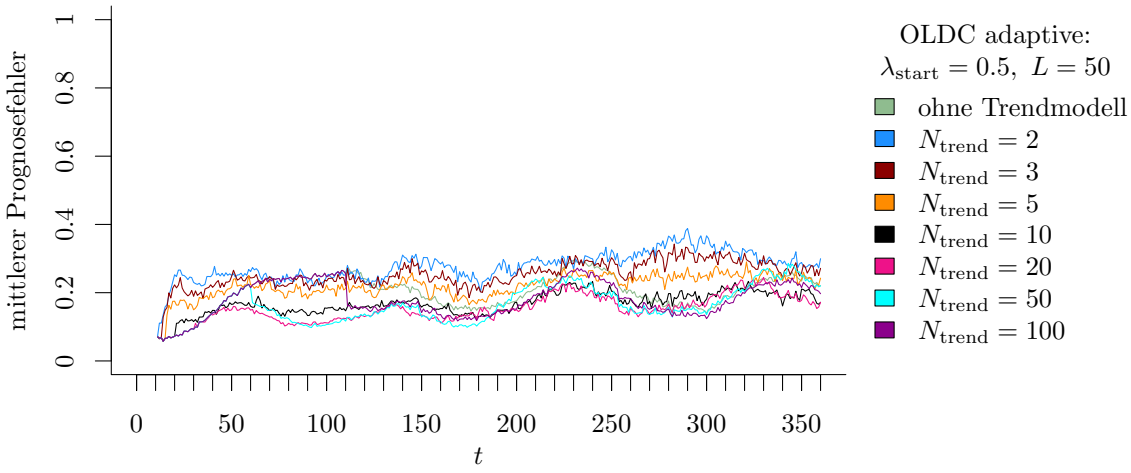
Abbildung 9.7: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **moving hyperplane**.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

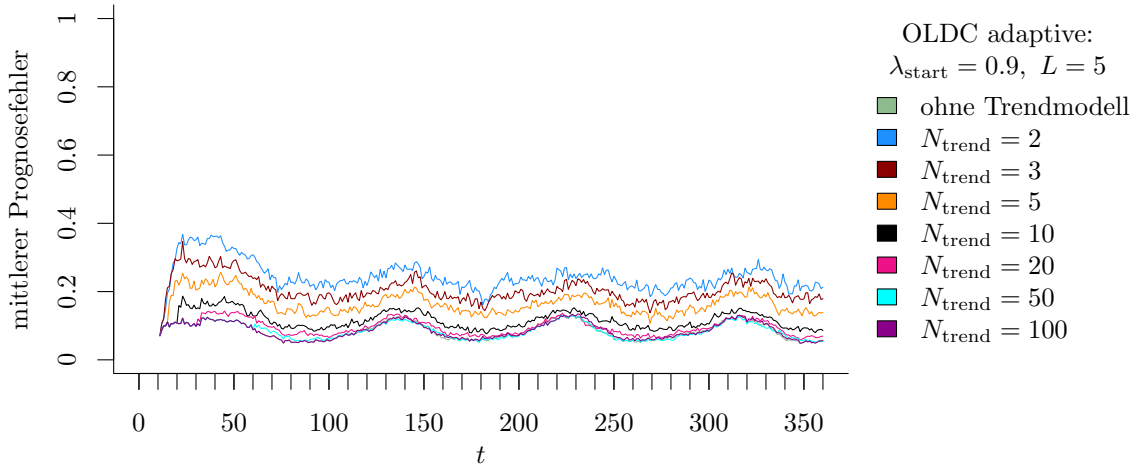


(b) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

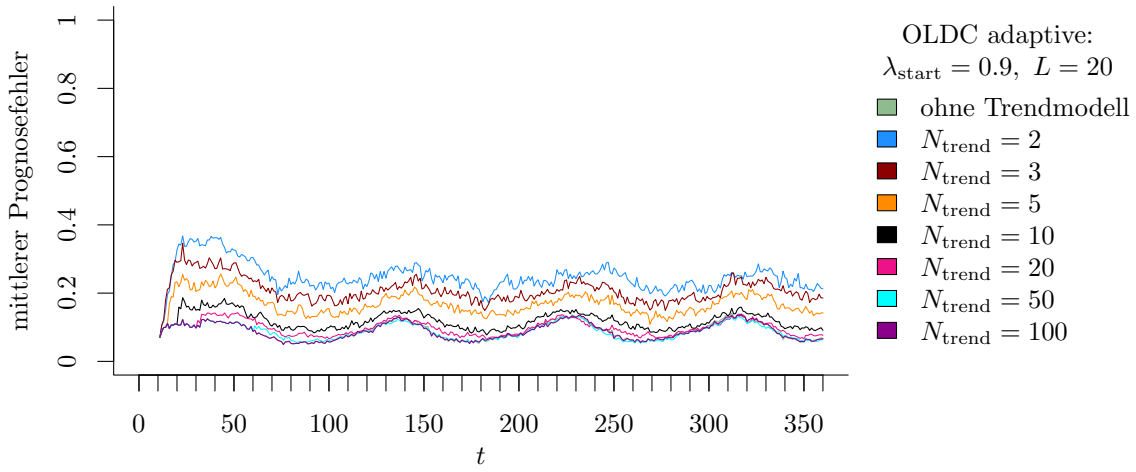


(c) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

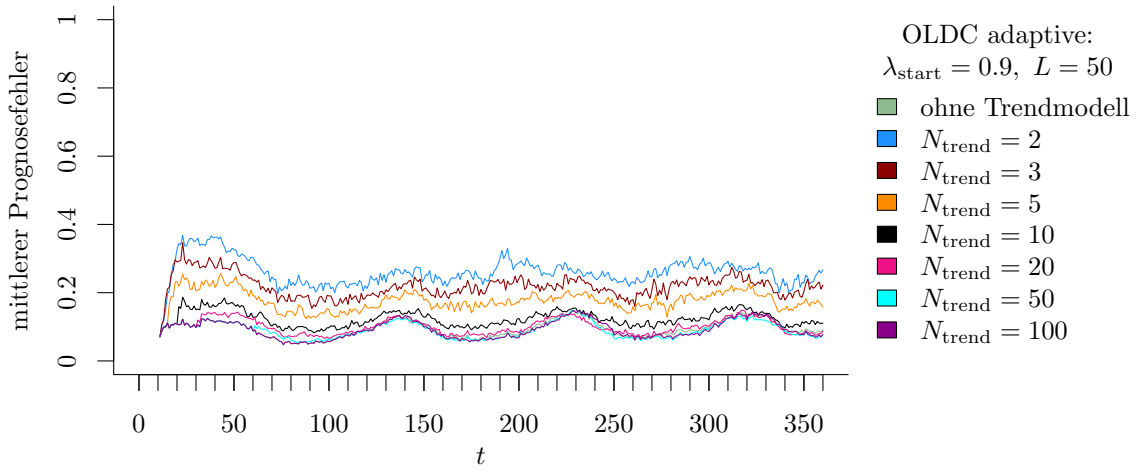
Abbildung 9.8: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **moving hyperplane**.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.9: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **moving hyperplane**.

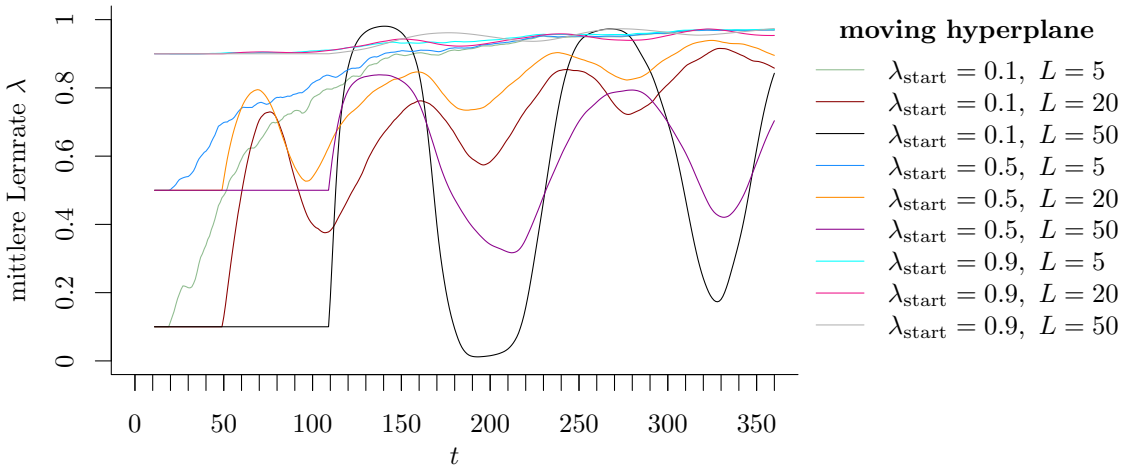


Abbildung 9.10: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **moving hyperplane**.

Die Methode *OLDC* mit adaptiver Lernrate bildet hier eine Ausnahme (vgl. Spalte „OLDC adaptive“). Der durchschnittliche Prognosefehler sinkt für die meisten Kombinationen aus λ_{start} und L mit steigendem N_{trend} bis zu einem gewissen Wert von N_{trend}. Dabei muss jedoch N_{trend} umso größer sein, je schmaler das Fenster L zur Adaption der Lernrate und je größer der Startwert λ_{start} ist, damit die Prognosefehler der ursprünglichen Methode ohne Erweiterung verringert werden können. Für viele Parameterkombinationen ist dies somit auch möglich. Die Methode *OLDC adaptive* ist bereits in ihrer ursprünglichen Variante den anderen Methoden überlegen. Allerdings ist die Wahl von λ_{start} und L sehr entscheidend.

Abbildung 9.10 veranschaulicht die Veränderung der adaptiven Lernrate bei *OLDC* im Laufe des Datenstroms für alle betrachteten Parameterkombinationen aus λ_{start} und L. Bei hohem Startwert $\lambda_{\text{start}} = 0.9$ verändert sich die Lernrate im Laufe der Zeit unabhängig von L kaum, sie wird allenfalls noch leicht größer. Bei kleineren Startwerten hat der Parameter L der Größe des Fensters zur Adaption einen stärkeren Einfluss. Bei breitem Fenster L sind starke Schwankungen zu erkennen. Es lässt sich daher schließen, dass ein breites Fenster hier ungeeignet ist, da aufgrund der stetigen Veränderung der Verteilungen im Datenstrom die Lernrate immer verhältnismäßig hoch sein sollte, damit aktuelle Beobachtungen beim Update der Klassifikationsregel stärker gewichtet werden. Je kleiner das Fenster L gewählt wird, desto schwächer sind die Schwankungen und die Lernrate konvergiert im Laufe der Zeit langsam gegen einen hohen Wert.

Diese Ergebnisse wirken sich direkt auf den Prognosefehler aus (vgl. Abbildungen 9.7–9.9). Die größten Schwankungen des Prognosefehlers resultieren jeweils bei $L = 50$. Der Verlauf des Prognosefehlers sieht bei $\lambda_{\text{start}} = 0.9$ für alle L relativ ähnlich aus. Bei Integration lokaler linearer Regressionsmodelle auf genügend großem N_{trend} steigt der durchschnittliche Prognosefehler über die Zeit mit wachsender Fenstergröße L und mit sinkendem Startwert λ_{start} (vgl. Tabelle 9.3). Insbesondere bei $\lambda_{\text{start}} = 0.9$ schwanken die durchschnittlichen Prognosefehler für die verschiedenen Fensterbreiten L nur leicht.

Tabelle 9.3: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **moving hyperplane** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend}); **grau fettgedruckt**: minimaler Wert pro „Spalte“ (pro Methode); ***schwarz kursiv und fettgedruckt***: minimaler mittlerer Prognosefehler insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4821 (0.007)	0.1326 (0.007)	0.1199 (0.008)	0.1	0.5175 (0.005)	5	0.1127 (0.004)
						20	0.1634 (0.007)
						50	0.2407 (0.009)
				0.3	0.5260 (0.006)		
				0.5	0.4822 (0.007)	5	0.0983 (0.003)
						20	0.1312 (0.005)
						50	0.2018 (0.011)
				0.7	0.2860 (0.003)		
				0.9	0.1251 (0.003)	5	***0.0850*** (0.003)
						20	0.0893 (0.003)
						50	0.0942 (0.003)
2	0.2424 (0.016)	0.2556 (0.023)	0.2473 (0.025)	0.1	0.2525 (0.017)	5	0.2250 (0.021)
						20	0.2406 (0.021)
						50	0.3001 (0.024)
				0.3	0.2433 (0.016)		
				0.5	0.2424 (0.016)	5	*0.2229* (0.021)
						20	0.2326 (0.021)
						50	0.2775 (0.023)
				0.7	0.2418 (0.016)		
				0.9	0.2356 (0.020)	5	0.2440 (0.025)
						20	0.2478 (0.025)
						50	0.2606 (0.028)
3	0.2297 (0.014)	0.2210 (0.017)	0.2122 (0.019)	0.1	0.2404 (0.015)	5	0.1965 (0.015)
						20	0.2193 (0.017)
						50	0.2756 (0.022)
				0.3	0.2316 (0.015)		
				0.5	0.2297 (0.014)	5	*0.1928* (0.015)
						20	0.2072 (0.016)
						50	0.2538 (0.021)
				0.7	0.2268 (0.014)		
				0.9	0.2074 (0.015)	5	0.2046 (0.018)
						20	0.2083 (0.018)
						50	0.2185 (0.021)
5	0.2082 (0.012)	0.1887 (0.012)	0.1780 (0.013)	0.1	0.2201 (0.013)	5	0.1655 (0.009)
						20	0.1904 (0.012)
						50	0.2441 (0.019)
				0.3	0.2105 (0.012)		
				0.5	0.2081 (0.012)	5	*0.1604* (0.009)
						20	0.1761 (0.011)
						50	0.2201 (0.017)
				0.7	0.2032 (0.011)		
				0.9	0.1743 (0.010)	5	0.1651 (0.011)
						20	0.1684 (0.011)
						50	0.1766 (0.013)
10	0.1467 (0.007)	0.1441 (0.007)	0.1300 (0.007)	0.1	0.1642 (0.009)	5	0.1222 (0.005)
						20	0.1460 (0.007)
						50	0.1930 (0.013)
				0.3	0.1496 (0.008)		
				0.5	0.1466 (0.007)	5	*0.1159* (0.005)
						20	0.1305 (0.006)
						50	0.1684 (0.011)
				0.7	0.1444 (0.007)		
				0.9	0.1253 (0.005)	5	0.1167 (0.005)
						20	0.1193 (0.006)
						50	0.1239 (0.006)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
20	**0.1118** (0.005)	0.1223 (0.005)	0.1089 (0.005)	0.1	0.1335 (0.007)	5	0.1050 (0.004)
						20	0.1354 (0.007)
						50	0.1771 (0.012)
				0.3	0.1144 (0.005)		
				0.5	0.1117 (0.005)	5	0.0983 (0.003)
						20	0.1166 (0.005)
						50	0.1525 (0.010)
				0.7	0.1140 (0.005)		
				0.9	0.1039 (0.004)	5	*0.0957* (0.003)
						20	0.0984 (0.004)
						50	0.1008 (0.004)
50	0.1315 (0.004)	**0.1129** (0.004)	**0.1016** (0.004)	0.1	0.1528 (0.006)	5	0.1017 (0.003)
						20	0.1387 (0.007)
						50	0.1915 (0.013)
				0.3	0.1378 (0.004)		
				0.5	0.1315 (0.004)	5	0.0920 (0.003)
						20	0.1164 (0.005)
						50	0.1632 (0.013)
				0.7	0.1286 (0.004)		
				0.9	**0.1021** (0.003)	5	*0.0853* (0.003)
						20	0.0883 (0.003)
						50	0.0909 (0.003)
100	0.2243 (0.003)	0.1225 (0.004)	0.1096 (0.005)	0.1	0.2681 (0.006)	5	0.1132 (0.003)
						20	0.1551 (0.006)
						50	0.2173 (0.008)
				0.3	0.2496 (0.004)		
				0.5	0.2242 (0.003)	5	0.0988 (0.003)
						20	0.1253 (0.004)
						50	0.1805 (0.009)
				0.7	0.1873 (0.003)		
				0.9	0.1164 (0.002)	5	*0.0860* (0.002)
						20	0.0885 (0.003)
						50	0.0920 (0.003)

Für *QDA-AF* und *LDA-AF* wird erst durch ein Fenster der Breite $N_{\text{trend}} = 20$ für die lokalen linearen Regressionsmodelle ein geringerer durchschnittlicher Prognosefehler über die Zeit erreicht als bei der jeweils ursprünglichen Methode ohne Regressionsmodell (vgl. erste „Zeile“ in Tabelle 9.3). Diese beiden Methoden sind ebenfalls bereits in ihrer ursprünglichen Variante sehr gut.

Die Tatsache, dass der Prognosefehler für fast alle Methoden durch Einführung lokaler linearer Regressionsmodelle zunächst mit wachsendem N_{trend} verringert werden kann, sich ab einem gewissen Schwellenwert von $N_{\text{trend}} > 50$ die Prognosegüte jedoch wieder verschlechtert, lässt sich dadurch erklären, dass die Voraussetzungen eines linearen Trends der Erwartungswerte in dieser Datensituation nicht erfüllt sind. Durch die Simulation der Daten im Einheitsquadrat mit einer rotierenden Klassengrenze bewegen sich die Erwartungswertvektoren mit der Zeit ungefähr auf einem Kreis (vgl. Abbildung 9.1 auf Seite 213). Da die aktualisierten Mittelwerte aus der LDA zur Anpassung des linearen Regressionsmodells herangezogen werden und diese zeitlich verzögerte Erwartungswerte repräsentieren (vgl. Abschnitt 7.3), kann die nicht-lineare Bewegung der Erwartungswerte bis zu einem gewissen Punkt durch ein lokales lineares Regressionsmodell approximiert werden.

Bei Betrachtung der Grafiken in Abbildungen 9.5–9.9 fällt wie bei den durchschnittlichen mittleren Prognosefehlern in Tabelle 9.3 auf, dass hier eine Wahl von $N_{\text{trend}} = 20$ bzw.

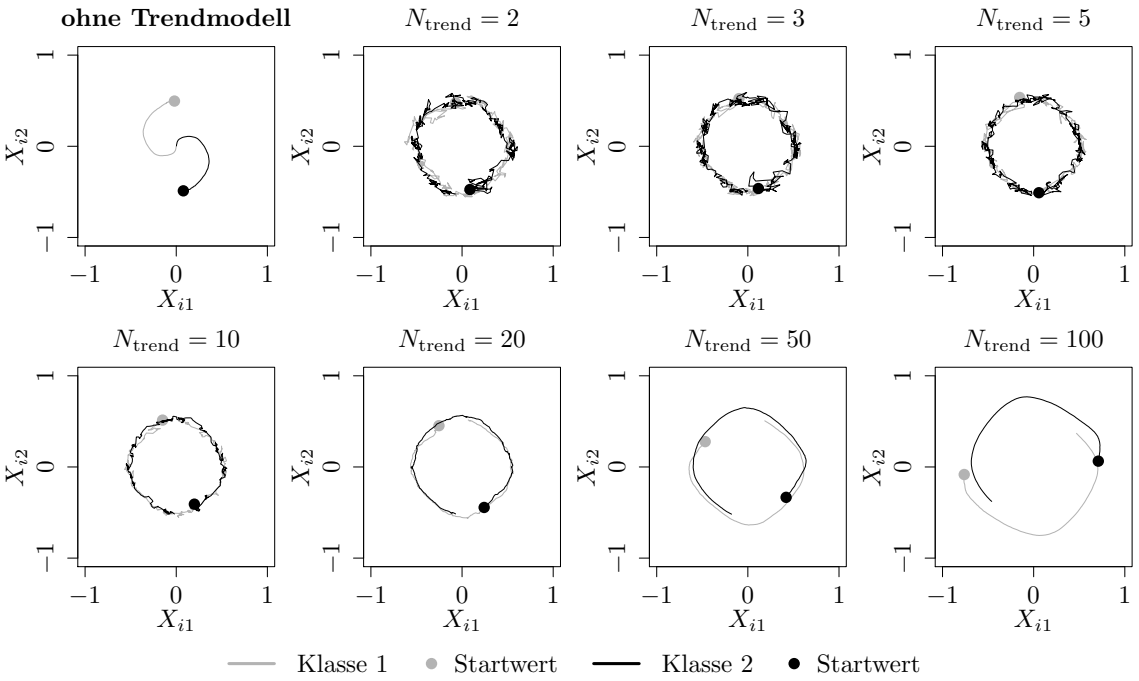


Abbildung 9.11: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren auf der Datensituation **moving hyperplane** für den gesamten Datenstrom. Ergebnisse aus *ILDA* und Erweiterung mit verschiedenen N_{trend}.

$N_{\text{trend}} = 50$ bei allen Methoden am besten erscheint. Die Kurven (pink und türkis) des Prognosefehlers liegen in den meisten Fällen zum Großteil der Zeitpunkte sichtbar unterhalb von den anderen Kurven, weshalb die Prognosegüte zu beinahe jedem Zeitpunkt im Datenstrom besser ist.

Die durchschnittliche Varianz des Prognosefehlers steigt zunächst bei Integration lokaler linearer Regressionsmodelle an, sinkt jedoch mit steigendem N_{trend} wieder, sodass die Werte der ursprünglichen Methoden wieder erreicht oder sogar unterboten werden (vgl. Tabelle 9.3).

In Abbildung 9.11 ist der Verlauf der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren für die Update-Methode *ILDA* der Fisher LDA sowie für die Erweiterung durch lokale lineare Regressionsmodelle auf Fenstern N_{trend} verschiedener Größe dargestellt. Die dicken Punkte markieren dabei den jeweils ersten prognostizierten Erwartungswertvektor in beiden Klassen im Datenstrom. Aufgrund der Rotation der Beobachtungen im Einheitsquadrat (vgl. Abbildung 9.1 auf Seite 213) bewegen sich die Erwartungswertvektoren beider Klassen im Laufe der Zeit approximativ auf einem Kreis. Anhand der grafischen Veranschaulichung wird deutlich, dass die Erwartungswertvektoren im Laufe des Datenstroms mit der ursprünglichen Methode nicht gut geschätzt werden (vgl. obere linke Grafik in Abbildung 9.11). Durch Erweiterung der Methode rotieren die Schätzer approximativ auf einem Kreis und repräsentieren deutlich stärker die wahren

Erwartungswertvektoren (die nicht bekannt sind) der zugrunde liegenden Verteilung der Daten zu jedem Zeitpunkt. Bei kleinem N_{trend} ist der Verlauf noch „zackig“. Die Schätzungen sind noch vergleichsweise unsicher und die Varianz ist verhältnismäßig hoch, da wenig Beobachtungen in die einzelnen Regressionsmodelle einfließen. Ab $N_{\text{trend}} = 20$ sieht der Verlauf stabil aus. Für $N_{\text{trend}} = 100$ wird der Radius des Kreises zu groß, da die Fenster für die Regressionsmodelle zu groß sind und die lineare Approximation des Trends der Erwartungswertvektoren für den hier vorliegenden nicht-linearen Trend nicht mehr greift.

Insgesamt wird anhand aller Ergebnisse deutlich, dass unter den betrachteten Fenstergrößen $N_{\text{trend}} \in \{20, 50\}$ am besten ist. Die Erwartungswerte werden so am besten geschätzt.

Die wichtigsten Ergebnisse zusammengefasst sind die Folgenden:

- Bereits für sehr kleines N_{trend} bei Integration lokaler linearer Regressionsmodelle zur Erweiterung der Methoden kann der Prognosefehler im Gegensatz zur ursprünglichen Methode für *ILDA* und *OLDC fix* mit kleinen Lernraten $\lambda < 0.9$ teilweise deutlich verringert werden.
- Bis zu einem gewissen Wert von N_{trend} sinkt der durchschnittliche mittlere Prognosefehler mit steigendem N_{trend} für alle Methoden.
- Unabhängig von der Erweiterung ist bei *OLDC fix* eine hohe Lernrate besser, der durchschnittliche mittlere Prognosefehler über die Zeit sinkt mit steigendem λ.
- Die besten Ergebnisse liefern *QDA-AF*, *LDA-AF* und *OLDC adaptive* mit geeigneter Wahl von λ_{start} und L. Hier kann jedoch ebenfalls die Einführung lokaler linearer Regressionsmodelle den Prognosefehler bei genügend großer Wahl von N_{trend} ($N_{\text{trend}} \geq 20$ bei *QDA-AF* und *LDA-AF*, bei *OLDC* abhängig von λ_{start} und L) noch verbessern.
- Bei *OLDC* mit adaptiver Lernrate ist in dieser Datensituation ein hoher Startwert $\lambda_{\text{start}} = 0.9$ geeignet. Dieser ist für die Prognosegüte jedoch weniger wichtig als die Wahl der Fensterbreite L zur Adaption der Lernrate. Für $L = 5$ resultieren die geringsten Prognosefehler.
- Den geringsten durchschnittlichen mittleren Prognosefehler über die Zeit von 0.0850 liefert *OLDC adaptive* mit $\lambda_{\text{start}} = 0.9$ und $L = 5$.

Fazit: Trotz nicht erfüllter Annahme eines linearen Trends der Erwartungswertvektoren in den Klassen sowie keiner Normalverteilung bzw. gar keiner typischen Verteilung durch die synthetische Generierung der Beobachtungen im Einheitsquadrat kann somit die Prognosegüte durch die Einführung lokaler linearer Regressionsmodelle zur Modellierung und Prognose des Drifts der Erwartungswerte für alle Methoden verbessert werden. Dies liegt daran, dass der nicht-lineare Trend lokal linear approximiert werden kann. Es stellt sich heraus, dass dabei ein Wert von $N_{\text{trend}} = 20$ bzw. $N_{\text{trend}} = 50$ zu guten Resultaten führt. Am besten können sich *OLDC adaptive*, *QDA-AF* und *LDA-AF* an den concept drift adaptieren, wobei trotzdem durch die Erweiterung der Methoden noch eine Verbesserung hinsichtlich des Prognosefehlers möglich ist.

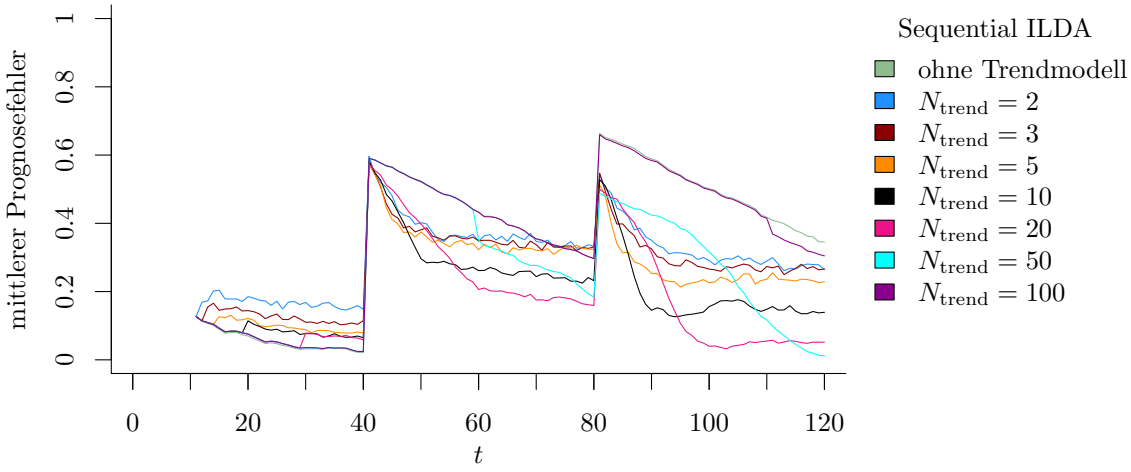
STAGGER Bei der Datensituation **STAGGER** gibt es zwei plötzliche Drifts, nach 40 und nach 80 Beobachtungen (vgl. Seite 214 f.). Diese Strukturbrüche sind im zeitlichen Verlauf des Prognosefehlers aller Methoden in den Abbildungen 9.12–9.16 zu erkennen. Dies liegt daran, dass alle Methoden sowie auch die Erweiterung nicht für sudden drifts bzw. Strukturbrücke entwickelt wurden und bei solchen erst nach einer gewissen Zeit eine Adaption an den Drift erfolgen kann.

Die Erweiterung der Methoden kann trotzdem eine Verbesserung der Prognosegüte liefern. Bei den Methoden *ILDA* und *OLDC fix* ist dies an den Kurven des mittleren Prognosefehlers in den Abbildungen 9.12 (a) bzw. 9.13 (a) und (b) zu erkennen. Ab dem ersten Drift (Zeitpunkt 40) liegen alle Kurven der Erweiterung unterhalb jener der nicht erweiterten Methode (grüne Kurve), sobald das erste Mal ein Regressionsmodell an die letzten $n_{\text{trend}}^{(c)}$ Mittelwertvektoren angepasst wird. Lediglich bis zum ersten Drift verschlechtert die Erweiterung zunächst die Prognosegüte. Da bis zu diesem Zeitpunkt eine stabile Verteilung vorliegt, kann diese mit den ursprünglichen Update-Methoden gut angepasst werden.

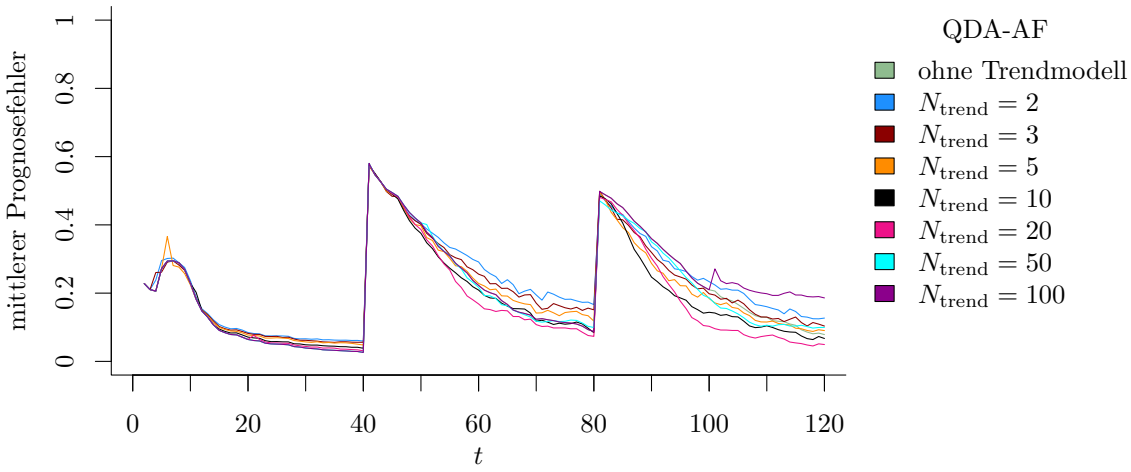
Bei Betrachtung der Ergebnisse auf dem gesamten Datenstrom mittelt sich dieser Nachteil der erweiterten Methoden vor dem ersten Drift heraus bzw. er hat weniger Einfluss als der Vorteil nach dem ersten Drift. Daher ist der durchschnittliche mittlere Prognosefehler über die Zeit für die Erweiterung der Methoden bei allen betrachteten Werten für N_{trend} deutlich geringer als ohne Betrachtung zusätzlicher Regressionsmodelle zur Schätzung und Prognose des Trends der Erwartungswerte (vgl. Spalten „ILDA“ und „OLDC fix“ in Tabelle 9.4).

Für *ILDA* und *OLDC fix* mit $\lambda \in \{0.1, 0.3, 0.5\}$ sinkt der durchschnittliche mittlere Prognosefehler mit steigendem N_{trend} bis zu einem Wert von $N_{\text{trend}} = 20$, er liegt jedoch auch für $N_{\text{trend}} \in \{50, 100\}$ noch unterhalb des Fehlers der ursprünglichen Methode. Für $\lambda \in \{0.7, 0.9\}$ ist der Fehler bereits bei *OLDC fix* ohne Erweiterung geringer als bei kleineren Lernraten. Bei Erweiterung des Modells mit lokalen linearen Regressionsmodellen sinkt zwar auch hier der durchschnittliche Prognosefehler mit wachsendem N_{trend} bis $N_{\text{trend}} = 20$, der ursprüngliche Fehler wird bei Betrachtung von $\lambda = 0.7$ jedoch erst mit $N_{\text{trend}} = 5$ und bei $\lambda = 0.9$ erst mit $N_{\text{trend}} = 10$ unterschritten. Generell sinkt der Fehler mit steigendem λ, d. h. größerer Lernrate (vgl. „Zeilen“ in Spalte „OLDC fix“).

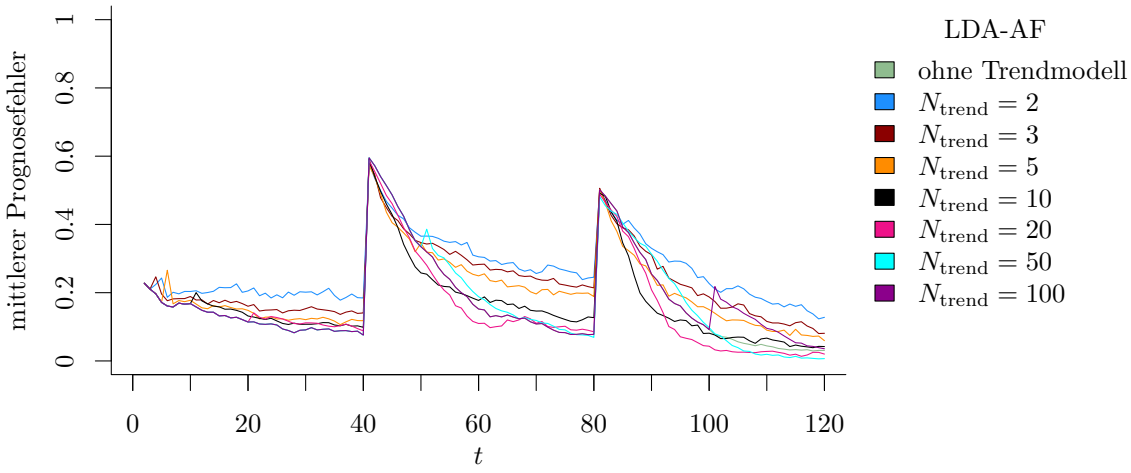
Bei *OLDC* mit adaptiver Lernrate sind die durchschnittlichen mittleren Prognosefehler bereits bei der ursprünglichen Methode vergleichsweise gering, außer bei kleinem Startwert λ_{trend} der Lernrate und großem Fenster L zur Adaption der Lernrate (vgl. Spalte „OLDC adaptive“ und Abbildungen 9.14–9.16). Trotzdem können die durchschnittlichen Prognosefehler durch die Erweiterung noch vereinzelt verringert werden. Tendenziell muss bei kleinerem Fenster L das betrachtete Fenster N_{trend} für die Regressionsmodelle breiter sein, damit der ursprüngliche Prognosefehler verringert werden kann. Der durchschnittliche mittlere Prognosefehler bei Integration lokaler linearer Regressionsmodelle sinkt bis zu einem Wert von $N_{\text{trend}} = 20$. Die Fehler werden durchgehend mit wachsendem Fenster L größer. In den meisten Fällen sinkt der Fehler zudem mit wachsendem Startwert λ_{start}.



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

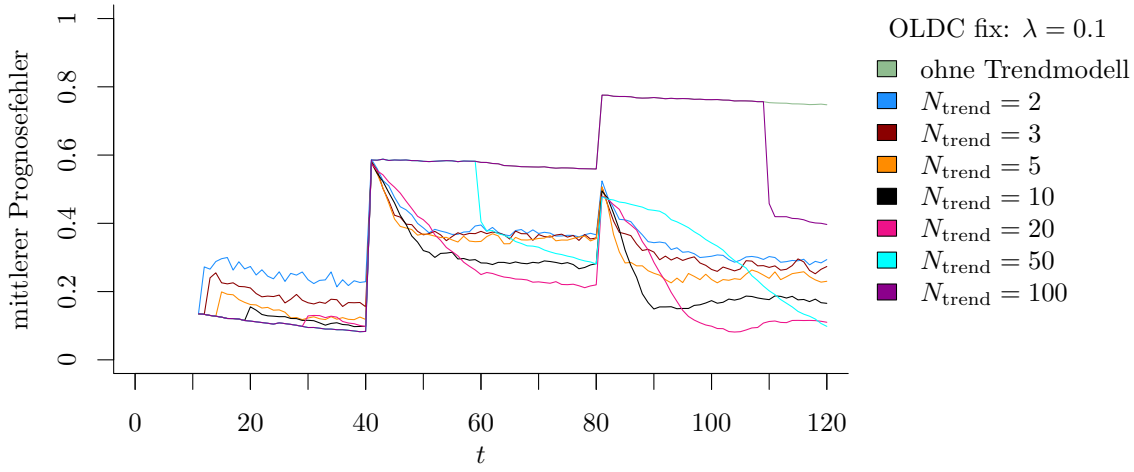


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

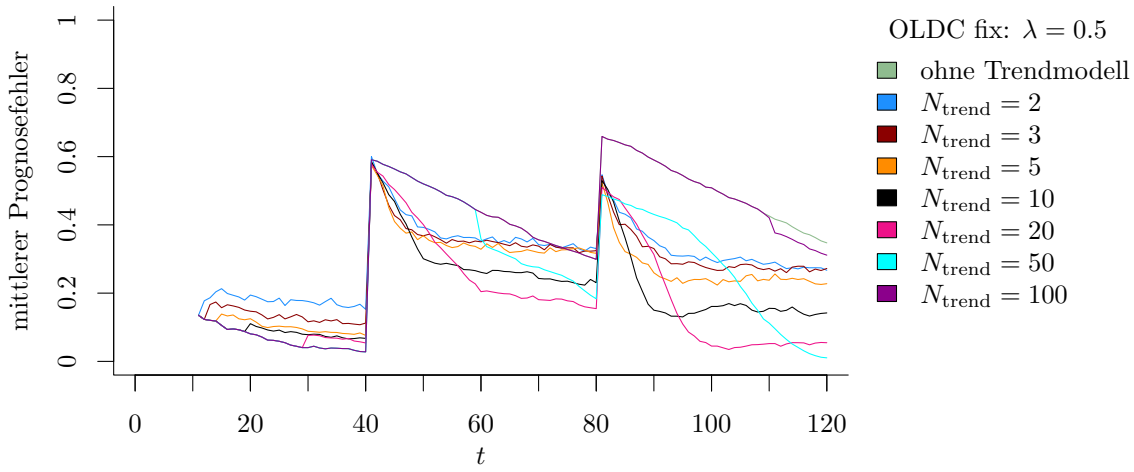


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

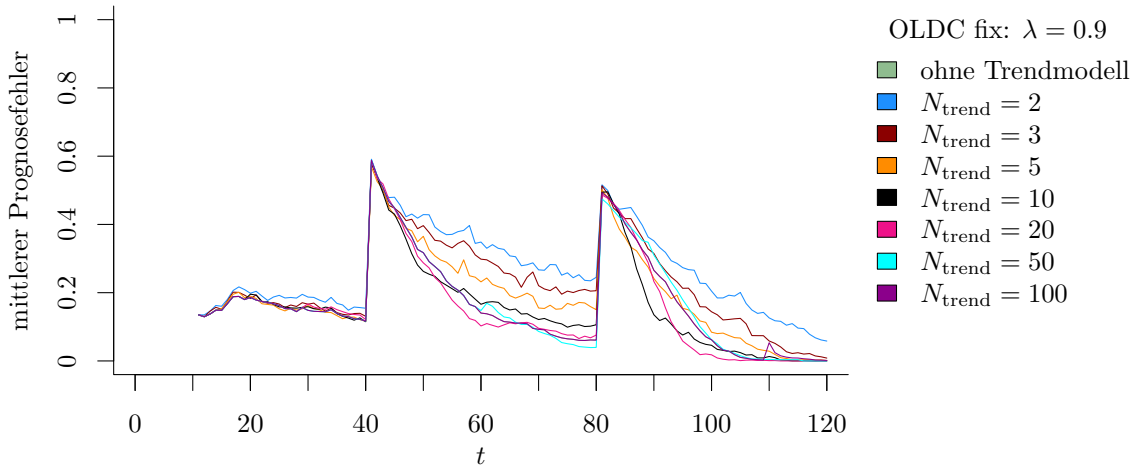
Abbildung 9.12: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **STAGGER**.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.

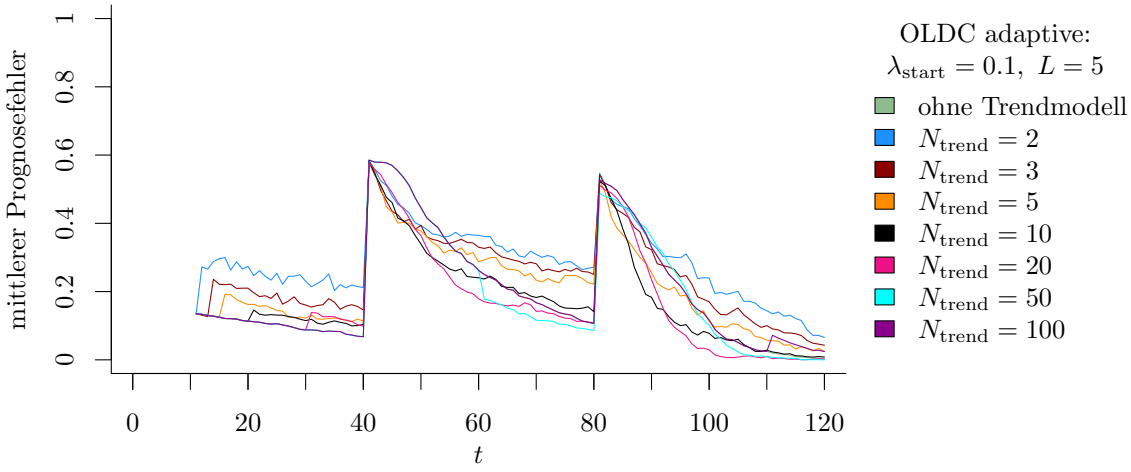


(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.

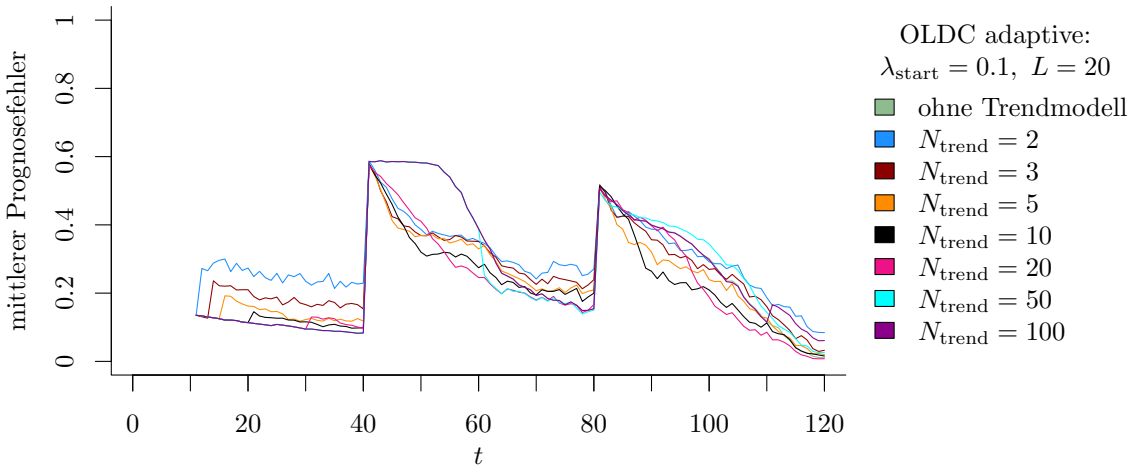


(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

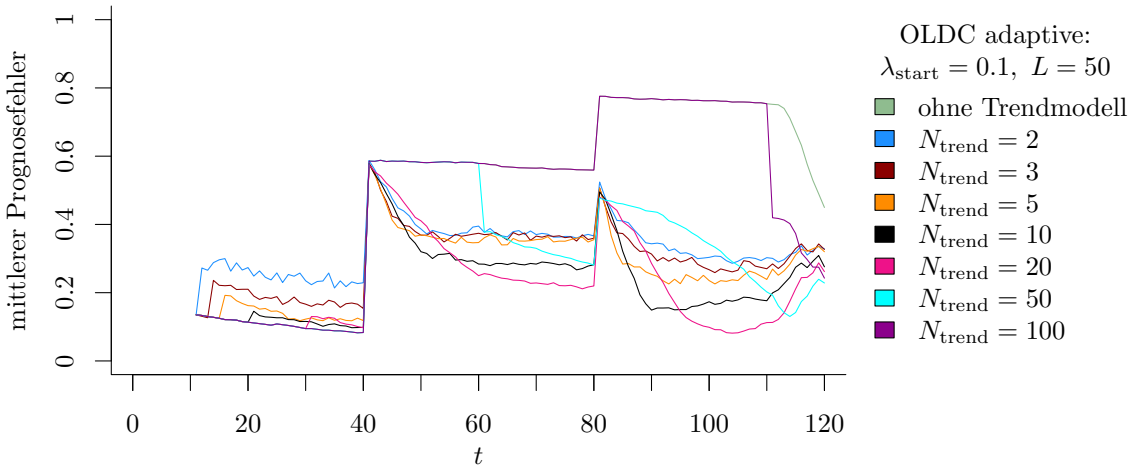
Abbildung 9.13: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **STAGGER**.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

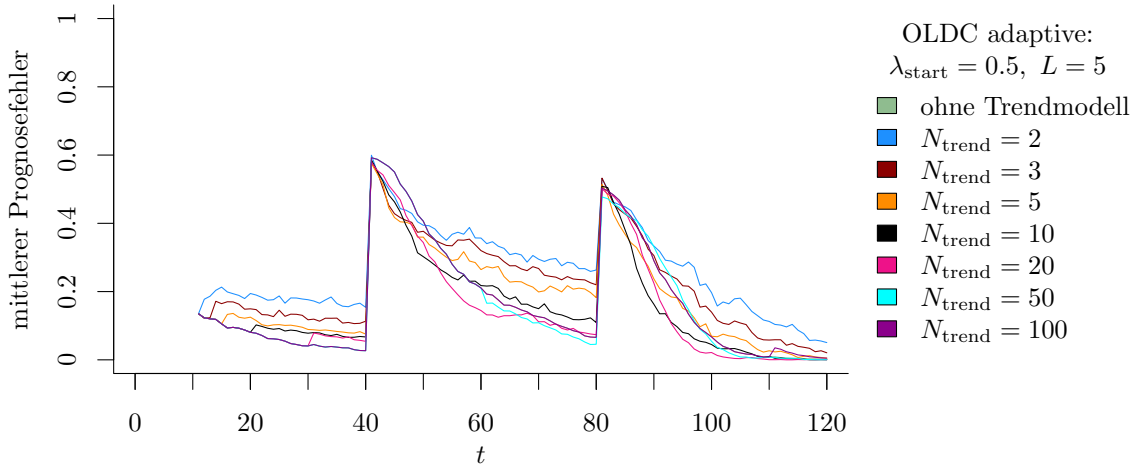


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

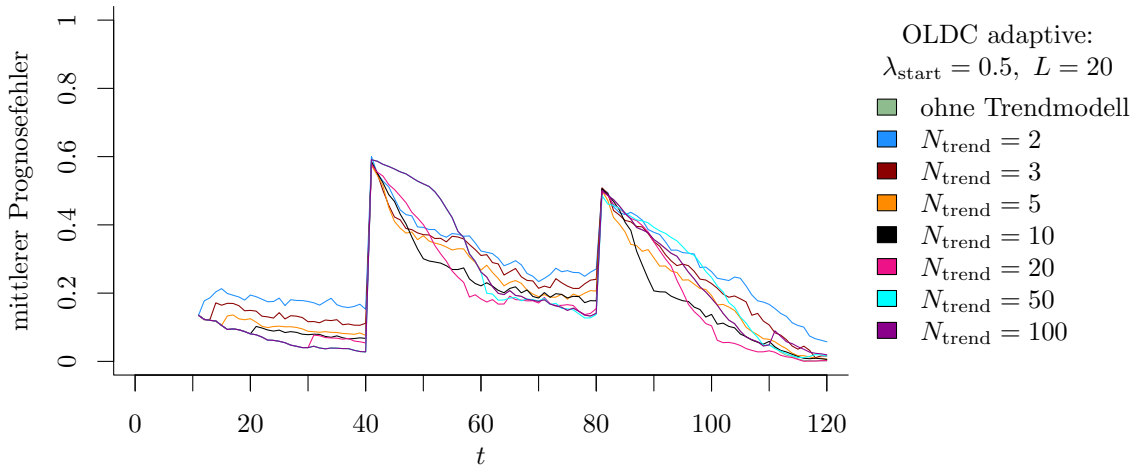


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

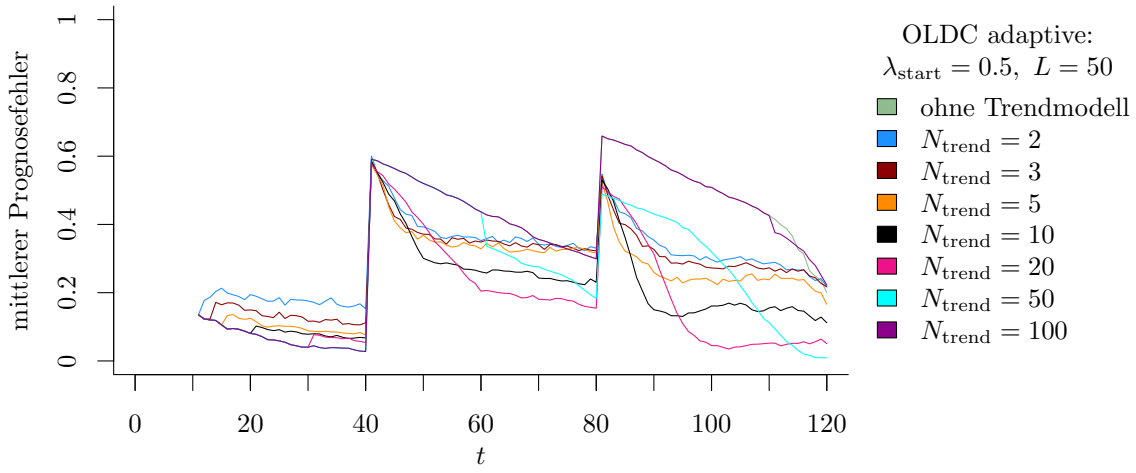
Abbildung 9.14: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **STAGGER**.



(a) **OLDC** mit $\lambda_{start} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

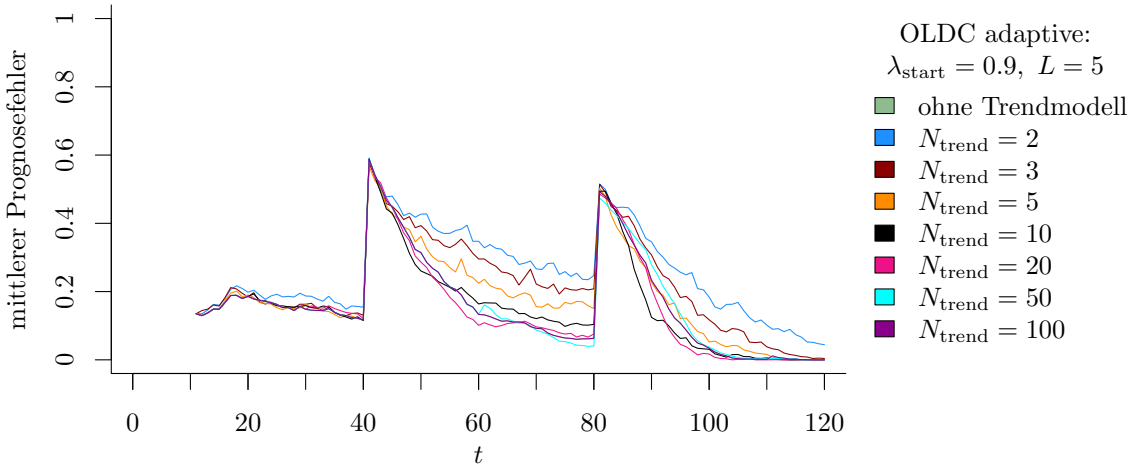


(b) **OLDC** mit $\lambda_{start} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

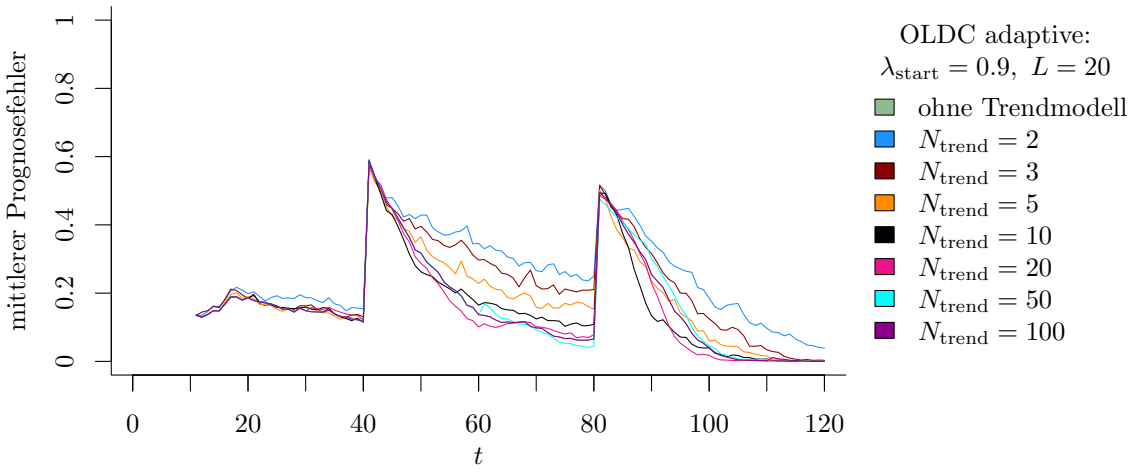


(c) **OLDC** mit $\lambda_{start} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

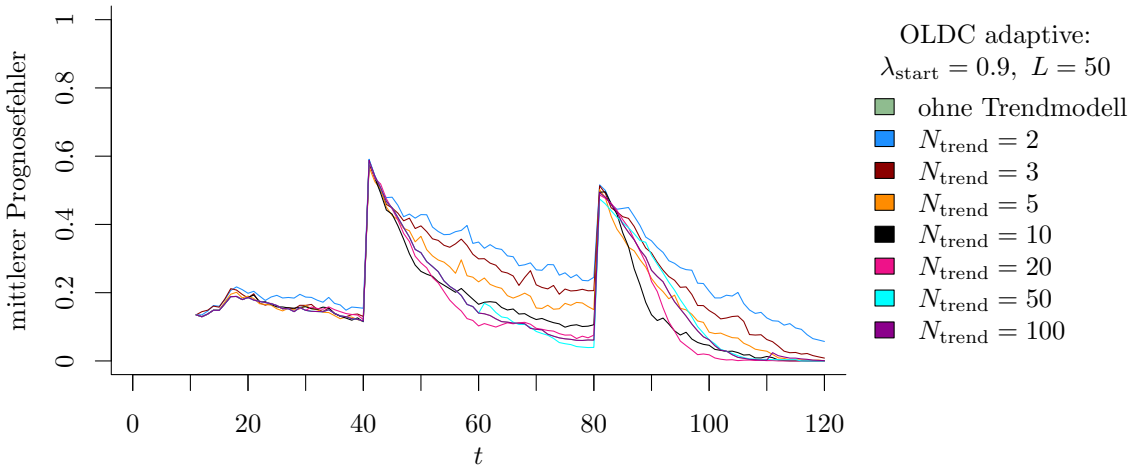
Abbildung 9.15: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **STAGGER**.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.16: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **STAGGER**.

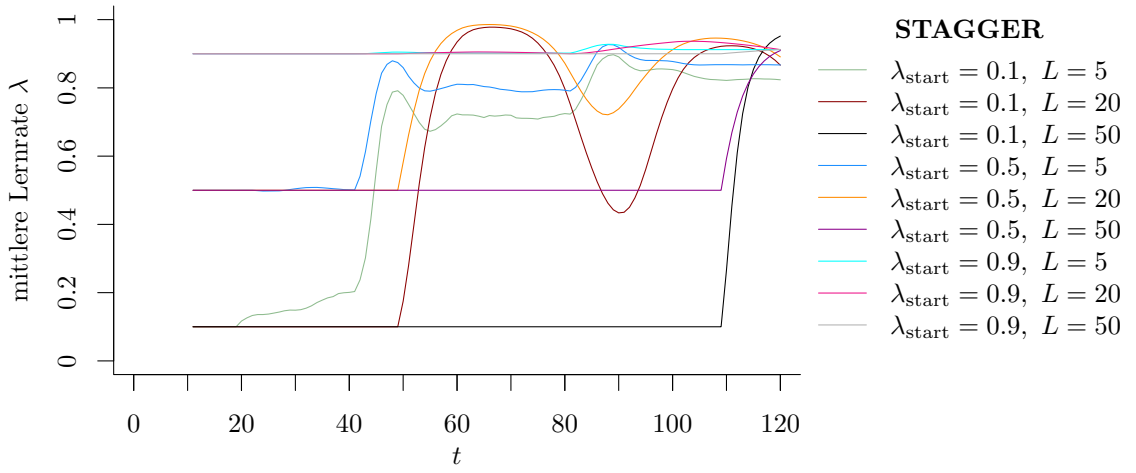


Abbildung 9.17: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **STAGGER**.

Abbildung 9.17 veranschaulicht die Veränderung der adaptiven Lernrate als Reaktion auf die Strukturbrüche im Laufe der Zeit. Bei hohem Startwert $\lambda_{\text{start}} = 0.9$ verändert sich die Lernrate im Laufe der Zeit unabhängig von L kaum. Sie wird lediglich für $L \in \{5, 20\}$ nach dem zweiten Strukturbruch noch etwas größer. Dies schadet der Prognosegüte jedoch nicht wie anhand Tabelle 9.4 ersichtlich wird. Bei den kleineren Startwerten wird deutlich, dass ein Fenster von $L = 50$ zu groß ist. Die beiden Strukturbrüche werden nicht (zeitnah) erkannt und die Lernrate bleibt bis kurz vor Ende des Datenstroms konstant. Als Resultat hängt die Prognosegüte in diesem Fall nur vom festgelegten Startwert λ_{start} ab. Je schmaler das Fenster L ist, desto schneller kann eine Reaktion auf den Strukturbruch erfolgen. Daher sinkt der durchschnittliche mittlere Prognosefehler mit schrumpfendem L.

Ein ähnlicher Trend für N_{trend} ist für die Erweiterungen von *QDA-AF* und *LDA-AF* festzustellen. Der Prognosefehler sinkt bis $N_{\text{trend}} = 20$. Der ursprüngliche Fehler wird erstmals bei $N_{\text{trend}} = 5$ für *QDA-AF* bzw. bei $N_{\text{trend}} = 10$ für *LDA-AF* unterschritten.

Zudem sei zu beachten, dass bei dieser Datensituation die Methode *ILDA* nicht bei allen 100 Simulationsdurchläufen durchgeführt werden konnte. Aufgrund der diskreten Einflussvariablen werden für die Diskriminanzanalyse Dummy-Variablen betrachtet (vgl. Seite 215). *ILDA* ist eine Update-Methode für die Fisher LDA. Bei der Initialisierung der Methode auf Basis der ersten zehn Beobachtungen traten in ein paar der 100 Simulationsdurchläufe klassenbasierte Varianzen von 0 in einer Dummy-Variable auf, da die Dummy-Variable für Beobachtungen der ersten Klasse lediglich Nullen und für jene der anderen Klasse lediglich Einsen aufwies. Dies führt zu einer Spalte und Zeile mit Nullen in der gepoolten Kovarianzmatrix $\boldsymbol{S}$ innerhalb der Klassen. Die Inverse der Quadratwurzel der Kovarianzmatrix (vgl. (3.39) auf Seite 48) ist folglich aufgrund eines Eigenwertes Null nicht definiert. Die Klassifikationsregel, welche auf dem Eigenwertproblem basiert, kann daher nicht aufgestellt werden. Die dargestellten Ergebnisse von *ILDA* basieren daher auf weniger Simulationsdurchläufen (Anzahl in der ersten Spalte in Tabelle 9.4 in eckigen Klammern).

Tabelle 9.4: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **STAGGER** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend}); **grau fettgedruckt**: minimaler Wert pro „Spalte“ (pro Methode); ***schwarz kursiv und fettgedruckt***: minimaler mittlerer Prognosefehler insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	[73] 0.3561 (0.006)	0.2054 (0.013)	0.1703 (0.011)	0.1	0.5144 (0.009)	5	0.2030 (0.009)
						20	0.2621 (0.013)
						50	0.5035 (0.009)
				0.3	0.4703 (0.008)		
				0.5	0.3610 (0.006)	5	0.1675 (0.007)
						20	0.2095 (0.010)
						50	0.3549 (0.006)
				0.7	0.2688 (0.007)		
				0.9	0.1676 (0.007)	5	*0.1613* (0.007)
						20	0.1651 (0.007)
						50	0.1676 (0.007)
2	[73] 0.3037 (0.012)	*0.2299* (0.017)	0.2656 (0.022)	0.1	0.3324 (0.018)	5	0.2903 (0.022)
						20	0.2996 (0.023)
						50	0.3347 (0.018)
				0.3	0.3151 (0.013)		
				0.5	0.3067 (0.012)	5	0.2637 (0.019)
						20	0.2710 (0.019)
						50	0.3052 (0.012)
				0.7	0.2972 (0.013)		
				0.9	0.2644 (0.021)	5	0.2580 (0.021)
						20	0.2587 (0.021)
						50	0.2644 (0.021)
3	[73] 0.2797 (0.011)	*0.2153* (0.015)	0.2311 (0.018)	0.1	0.3011 (0.015)	5	0.2492 (0.016)
						20	0.2624 (0.018)
						50	0.3031 (0.015)
				0.3	0.2882 (0.011)		
				0.5	0.2826 (0.011)	5	0.2241 (0.015)
						20	0.2350 (0.016)
						50	0.2810 (0.011)
				0.7	0.2729 (0.011)		
				0.9	0.2261 (0.017)	5	0.2205 (0.017)
						20	0.2220 (0.017)
						50	0.2276 (0.017)
5	[73] 0.2521 (0.010)	0.2030 (0.014)	0.2052 (0.015)	0.1	0.2708 (0.012)	5	0.2147 (0.013)
						20	0.2346 (0.015)
						50	0.2753 (0.012)
				0.3	0.2597 (0.010)		
				0.5	0.2554 (0.010)	5	0.1885 (0.011)
						20	0.2050 (0.013)
						50	0.2538 (0.010)
				0.7	0.2454 (0.011)		
				0.9	0.1922 (0.013)	5	*0.1863* (0.013)
						20	0.1885 (0.013)
						50	0.1921 (0.013)
10	[73] 0.2046 (0.008)	0.1861 (0.012)	0.1686 (0.012)	0.1	0.2268 (0.011)	5	0.1792 (0.009)
						20	0.2142 (0.014)
						50	0.2339 (0.011)
				0.3	0.2124 (0.008)		
				0.5	0.2061 (0.008)	5	*0.1566* (0.008)
						20	0.1833 (0.012)
						50	0.2054 (0.008)
				0.7	0.1963 (0.009)		
				0.9	0.1625 (0.010)	5	0.1594 (0.010)
						20	0.1613 (0.010)
						50	0.1624 (0.010)

Fortsetzung auf der nächsten Seite

	ILDA	QDA-AF	LDA-AF	OLDC fix		OLDC adaptive	
N_{trend}				λ,λ_{start}		L	
20	[73] **0.1820** (0.007)	**0.1773** (0.011)	**0.1559** (0.010)	0.1	0.2159 (0.009)	5	0.1679 (0.008)
						20	0.2195 (0.014)
						50	0.2240 (0.010)
				0.3	0.1944 (0.007)		
				0.5	0.1839 (0.007)	5	***0.1449*** (0.007)
						20	0.1816 (0.011)
						50	0.1836 (0.007)
				0.7	0.1740 (0.007)		
				0.9	**0.1569** (0.009)	5	0.1543 (0.009)
						20	0.1567 (0.009)
						50	0.1568 (0.009)
50	[73] 0.2567 (0.007)	0.2001 (0.015)	0.1755 (0.011)	0.1	0.3072 (0.009)	5	0.1923 (0.008)
						20	0.2652 (0.013)
						50	0.3124 (0.009)
				0.3	0.2828 (0.007)		
				0.5	0.2592 (0.007)	5	0.1642 (0.007)
						20	0.2174 (0.011)
						50	0.2594 (0.007)
				0.7	0.2249 (0.007)		
				0.9	0.1703 (0.008)	5	*0.1641* (0.008)
						20	0.1684 (0.008)
						50	0.1702 (0.008)
100	[73] 0.3521 (0.006)	0.2184 (0.016)	0.1800 (0.012)	0.1	0.4808 (0.009)	5	0.2061 (0.009)
						20	0.2674 (0.014)
						50	0.4774 (0.009)
				0.3	0.4454 (0.008)		
				0.5	0.3575 (0.006)	5	0.1686 (0.007)
						20	0.2125 (0.011)
						50	0.3550 (0.006)
				0.7	0.2742 (0.007)		
				0.9	0.1688 (0.007)	5	*0.1615* (0.007)
						20	0.1654 (0.007)
						50	0.1683 (0.007)

Fazit: Insgesamt kann der Prognosefehler für alle betrachteten Methoden (bei geeigneter Wahl der Parameterwerte) durch Einführung lokaler linearer Regressionsmodelle zur Modellierung des Trends der Erwartungswerte und deren Prognose verringert werden, obwohl die Annahme eines linearen Trends der Erwartungswertvektoren nicht gerechtfertigt ist. Die Approximation des nicht-linearen Trends durch die lokalen linearen Regressionsmodelle greift somit auch bei diesem sudden drift und wirkt sich positiv auf den Prognosefehler aus. Die optimale Wahl von N_{trend} hängt dabei von den Zeitabständen zwischen den einzelnen Strukturbrüchen ab und ist in dieser Datensituation $N_{\text{trend}} = 20$ für alle betrachteten Methoden der Online Diskriminanzanalyse. Ab $N_{\text{trend}} = 20$ liegen die durchschnittlichen Varianzen des Prognosefehlers auch wieder auf demselben Niveau wie dem der Ausgangsmethoden (vgl. Tabelle 9.4). Liegen die Strukturbrüche weiter auseinander, sollte ein größerer Wert für N_{trend} gewählt werden. Zwar wird bei größerem N_{trend} teilweise eine längere Zeit für die Regeneration des starken Anstiegs des Prognosefehlers nach einem Strukturbruch benötigt, jedoch kann der Prognosefehler im Laufe des Datenstroms weiter sinken, sodass dieser Vorteil im Laufe der Zeit überwiegt. Sind in der Praxis die Zeitabstände zwischen einzelnen sudden drifts nicht bekannt, so könnte der Prognosefehler eine Zeit lang beobachtet werden. Wie bereits erwähnt, wirken sich sudden drifts vorübergehend stark (und unvermeidbar) auf den Prognosefehler aus. Der Parameter N_{trend} könnte dann anhand des Musters zeitlich wiederkehrender Strukturbrüche angepasst werden.

9.6.2 Ergebnisse weiterer Datensituationen in $p = 2$

Datensituation „Kreisen“ ($p = 2$) Die Datensituation „Kreisen“ repräsentiert einen ähnlichen Typ des concept drifts wie die Datensituation „moving hyperplane“, da die Erwartungswertvektoren ebenfalls im Laufe der Zeit auf einem Kreis rotieren (vgl. Seite 218). Allerdings sind in dieser Situation – im Gegensatz zu „moving hyperplane“ – zu jedem Zeitpunkt die Annahmen der Linearen Diskriminanzanalyse erfüllt, da die Beobachtungen jeweils aus einer multivariaten Normalverteilung mit identischer Kovarianzmatrix in beiden Klassen generiert werden. Die Annahme eines zeitlichen linearen Trends der Erwartungswertvektoren ist hier jedoch ebenfalls aufgrund der Rotation der Erwartungswertvektoren nicht erfüllt. Der nicht-lineare Trend soll daher durch die Betrachtung lokaler linearer Regressionsmodelle approximiert werden.

In den Abbildungen 9.19–9.23 ist der Verlauf des mittleren Prognosefehlers für den gesamten Datenstrom für die betrachteten Methoden sowie ihre Erweiterungen mit verschiedenen Werten von N_{trend} veranschaulicht. Zusätzlich ist der Bayesfehler für die Verteilungen zu jedem Zeitpunkt und somit der Verlauf des Bayesfehlers eingezeichnet. Für den quantitativen Vergleich anhand einer einzelnen Maßzahl sind in Tabelle 9.5 auf Seite 263 f. die durchschnittlichen mittleren Prognosefehler sowie die durchschnittlichen Varianzen des Prognosefehlers über die Zeit für alle Methoden und Erweiterungen zusammengefasst.

In den Abbildungen der Methoden ohne Anpassung an einen möglichen concept drift (*ILDA* in Abbildung 9.19 (a) und *OLDC* mit fester Lernrate $\lambda = 0.5$ in Abbildung 9.20 (b)) ist anhand der grünen Kurve (ohne Erweiterung durch Trendmodell) deutlich erkennbar, dass die Erwartungswerte elf mal vollständig auf dem Kreis rotieren. Immer nach 360 Beobachtungen wird wieder die Ausgangsverteilung betrachtet, da mit jedem neuen Zeitpunkt bzw. bei jeder neuen Beobachtung eine Rotation der Erwartungswertvektoren um ein Grad betrachtet wird. Dies wirkt sich deutlich auf den Prognosefehler aus. Immer nach etwa 360 Beobachtungen beträgt der Prognosefehler ungefähr 0.5. Es erfolgt eine zufällige Klassifikation in eine der beiden Klassen, da für die Bildung der Klassifikationsregel zu diesem Zeitpunkt in beiden Klassen aufgrund der vollständigen Rotation der Erwartungswertvektoren auf dem Kreis Beobachtungen aus denselben Verteilungen vorliegen. Zu den dazwischen liegenden Zeitpunkten sinkt der Prognosefehler zunächst unter 0.2 und steigt auf über 0.8, wobei die Varianz der Schwankungen im Laufe der Zeit nach mehreren Rotationen abnimmt. Dies liegt daran, dass das Verhältnis von Anzahl Beobachtungen aus Verteilungen mit unterschiedlichen Erwartungswertvektoren in beiden Klassen zu jener aus Verteilungen mit identischen Erwartungswertvektoren im Laufe der Zeit abnimmt. In der Abbildung 9.25 am Ende des Abschnittes auf Seite 270 ist die Klassifikationsregel bzw. ihre Veränderung über die Zeit für diese Datensituation für ausgewählte Zeitpunkte veranschaulicht.

Die Schwankungen des Prognosefehlers sind bei einer kleineren Lernrate $\lambda = 0.1$ (vgl. Abbildung 9.20 (a)) stärker ausgeprägt, bei stärkerer Gewichtung der aktuellen Beobach-

tungen bei den Updates durch eine Lernrate von $\lambda = 0.9$ sind die Schwankungen kaum noch sichtbar (vgl. Abbildung 9.20 (c)). Jedoch steigt der Prognosefehler trotzdem im Laufe der Zeit an und liegt deutlich über dem Bayesfehler, welcher für diese Datensituation zu jedem Zeitpunkt 0.0786 beträgt (vgl. graue Kurve).

Durch die Einführung lokaler linearer Regressionsmodelle zur Modellierung des Drifts der Erwartungswerte durch lineare Approximation und daraufhin Prognose der Erwartungswerte kann der Verlauf des Prognosefehlers bei *ILDA* und *OLDC* deutlich verringert werden. Für Werte von $N_{\text{trend}} \leq 50$ liegt der mittlere Prognosefehler zu jedem Zeitpunkt nur leicht über dem Bayesfehler, wobei die Fehlerrate für $N_{\text{trend}} = 20$ in allen Fällen am meisten verringert werden kann. Insgesamt resultieren bei Integration von lokalen linearen Regressionsmodellen keine vergleichbar starken Schwankungen über die Zeit mehr, allerdings steigt der Prognosefehler im Mittel mit steigendem Fenster $N_{\text{trend}} \geq 20$ für die Regressionsmodelle.

Für *QDA-AF* und *LDA-AF* wird die Kurve des Prognosefehlers ohne Trendmodell von den anderen Kurven mit $N_{\text{trend}} \leq 100$ überlagert (vgl. Abbildungen 9.19 (b) und (c)). Bei diesen Methoden liegt damit der Prognosefehler generell näher am Bayesfehler – auch ohne Erweiterung. Die Verbesserung durch Einbindung der Regressionsmodelle wird hier jedoch anhand der Ergebnisse in Tabelle 9.5 (vgl. Seite 263 f.) deutlich. Der durchschnittliche mittlere Prognosefehler über die Zeit steigt für $N_{\text{trend}} = 10$ erst einmal leicht gegenüber jenem der Methode ohne Trendmodell an. Für $N_{\text{trend}} \in \{20, 50, 100\}$ kann jedoch eine Verbesserung erzielt werden, bevor der Prognosefehler für größere Fenster wieder ansteigt. Dies liegt daran, dass kein linearer Trend der Erwartungswerte vorliegt und der Trend nur bis zu einer gewissen Fenstergröße linear approximiert werden kann. Es sei jedoch darauf hingewiesen, dass selbst bei großem N_{trend} die Prognosefehler von *QDA-AF* und *LDA-AF* im Vergleich zu jenen der Methoden *ILDA* und *OLDC fix* verhältnismäßig klein sind.

Dies ist ebenfalls bei *OLDC* mit adaptiver Lernrate der Fall. Zudem sind die durchschnittlichen mittleren Prognosefehler über die Zeit bereits ohne Erweiterung der Methode vergleichsweise gering (vgl. Spalte „OLDC adaptive“ in Tabelle 9.5). Nach Erweiterung der Methode steigt der Prognosefehler bei kleinem N_{trend} zunächst für fast alle Kombinationen aus λ_{start} und L an. Mit wachsendem Fenster für das Trendmodell können jedoch auch die vergleichsweise geringen Prognosefehler der ursprünglichen Methode noch etwas verringert werden (durch $N_{\text{trend}} \in \{50, 100\}$). Dies zeigt sich auch am Verlauf des mittleren Prognosefehlers in den Abbildungen 9.21–9.23. Für alle Kombinationen aus λ_{start} und L liegt die orangefarbene ($N_{\text{trend}} = 50$) und schwarze ($N_{\text{trend}} = 100$) Kurve zu den meisten Zeitpunkten unterhalb der grünen (ohne Trendmodell).

Es zeigt sich auch, dass ein großes Fenster $L = 50$ zur Adaption der Lernrate zu tendenziell höheren Prognosefehlern und einem Anstieg des Fehlers im Laufe des Datenstroms führt. Abbildung 9.18 veranschaulicht die Veränderung der Lernrate λ im Laufe des Datenstroms für verschiedene Kombinationen aus λ_{start} und L. Während bei kleinem Fenster $L = 5$

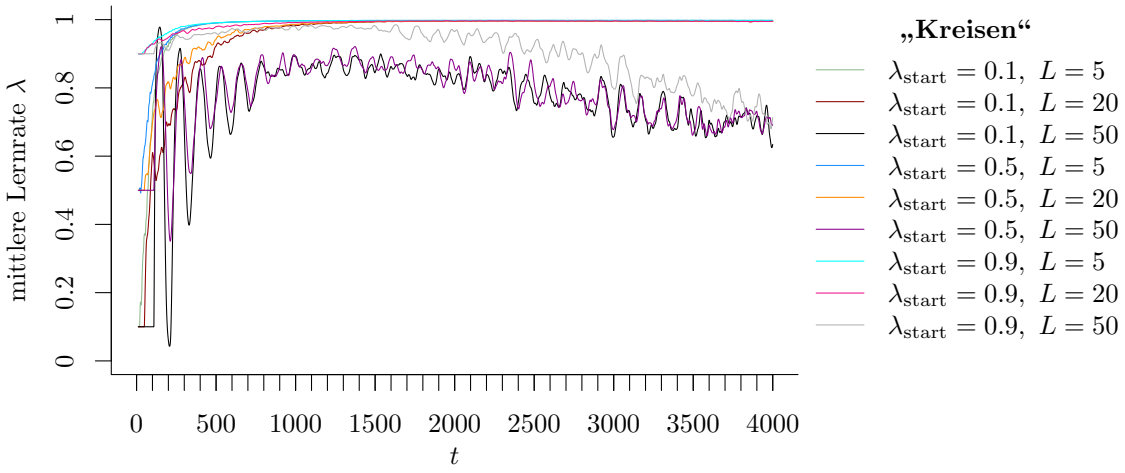


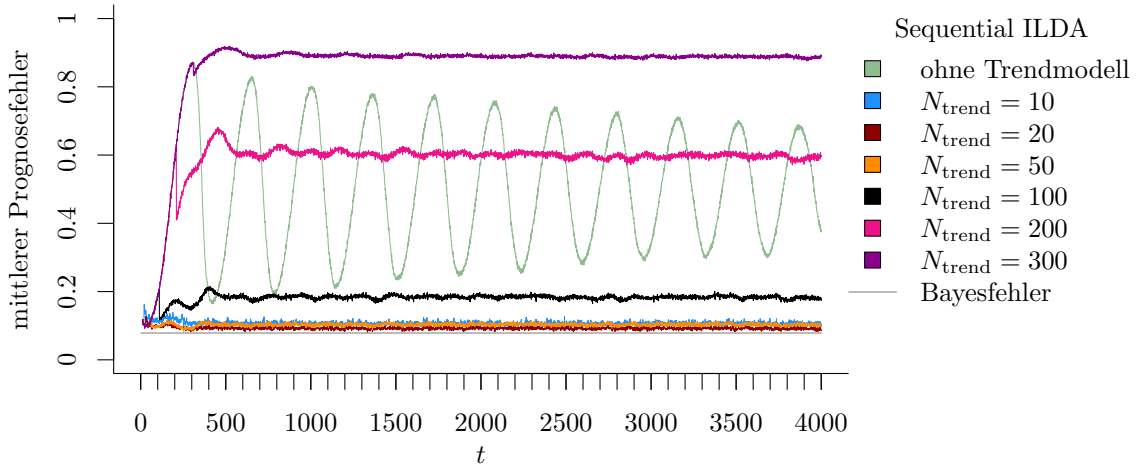
Abbildung 9.18: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **„Kreisen“**.

die Lernrate relativ unabhängig vom Startwert gegen ihr Maximum konvergiert, resultieren bei großem Fenster L zur Adaption große Schwankungen der Lernrate über die Zeit. Eine hohe Lernrate λ scheint auf dieser Datensituation jedoch intuitiv geeignet, da sich die Verteilungen stetig ändern und demnach aktuelle Beobachtungen mit einem stärkeren Gewicht in die Aktualisierung des Modells einfließen sollten. Aufgrund der datenbasierten Anpassung der Lernrate (vgl. Seite 75) ist ein großes Fenster zur Adaption hier ungeeignet. Es wird deutlich, dass die Wahl von L wichtiger ist als der Startwert λ_{start}. Dies zeigt sich auch an den Prognosefehlern. Die Grafiken (a), (b) und (c) der Abbildungen 9.21–9.23 sind qualitativ jeweils ähnlich (unterschiedlicher Startwert λ_{start}).

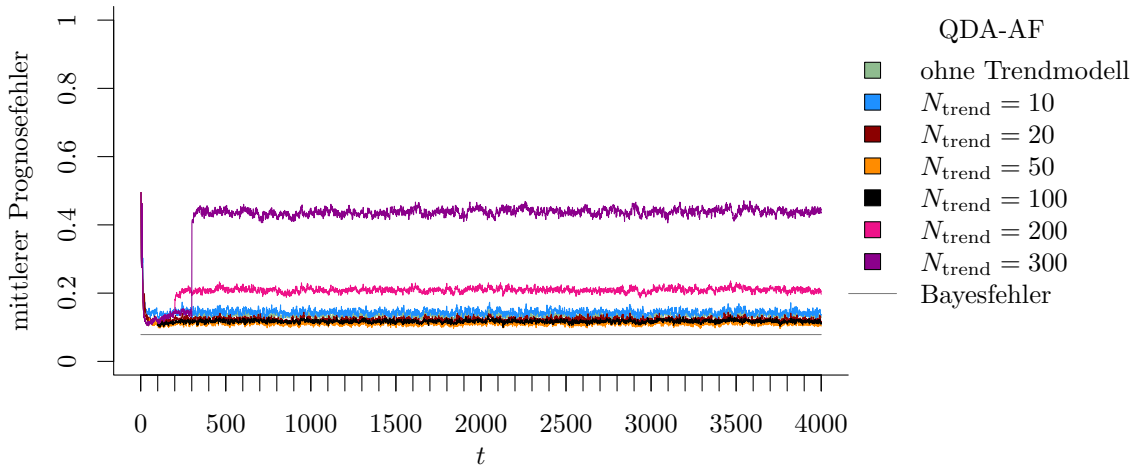
Bei *ILDA* und *OLDC* mit fester Lernrate reduziert die Einführung lokaler linearer Regressionsmodelle den durchschnittlichen mittleren Prognosefehler über die Zeit und dieser sinkt bei den betrachteten Werten bis zu $N_{\text{trend}} = 20$, bevor er mit steigendem N_{trend} wieder größer wird. In den meisten Fällen sinkt der durchschnittliche mittlere Prognosefehler über die Zeit (vgl. „Zeilen“) zudem mit wachsender Lernrate λ bei *OLDC fix*. Eine hohe Lernrate ist in dieser Datensituation allgemein von Vorteil. Im optimalen Bereich des Fensters $N_{\text{trend}} \in \{20, 50\}$ für das Trendmodell spielt die Wahl von λ jedoch keine große Rolle mehr. Die Prognosefehler unterscheiden sich hier nur leicht.

Auch bei *QDA-AF*, *LDA-AF* und *OLDC* mit adaptiver Lernrate können die Prognosefehler durch die Erweiterung verringert werden. Der Unterschied zu den anderen Methoden besteht darin, dass für kleine N_{trend} der durchschnittliche mittlere Prognosefehler zunächst höher liegt als ohne Erweiterung der Methoden. Für steigendes Fenster N_{trend} für die lokalen linearen Regressionsmodelle sinken jedoch die Prognosefehler und die minimalen Werte werden auf dieser Datensituation für $N_{\text{trend}} = 50$ bzw. $N_{\text{trend}} = 100$ erzielt.

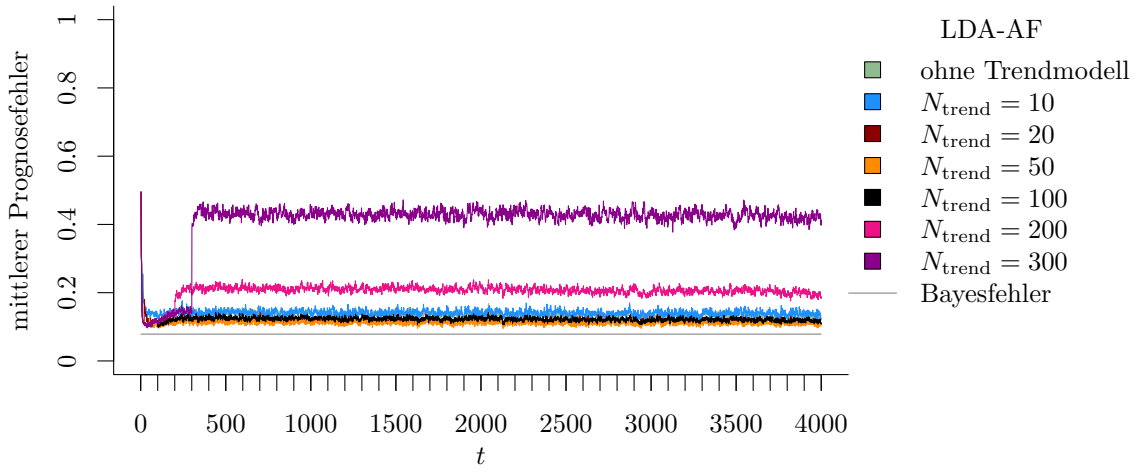
Die durchschnittliche Varianz des Prognosefehlers über die Zeit steigt bei Erweiterung von *OLDC adaptive* zunächst bei kleinen Fenstern N_{trend} minimal an. Mit wachsendem N_{trend} sinkt die Varianz (bei allen Methoden) jedoch wieder recht schnell (vgl. Tabelle 9.5).



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

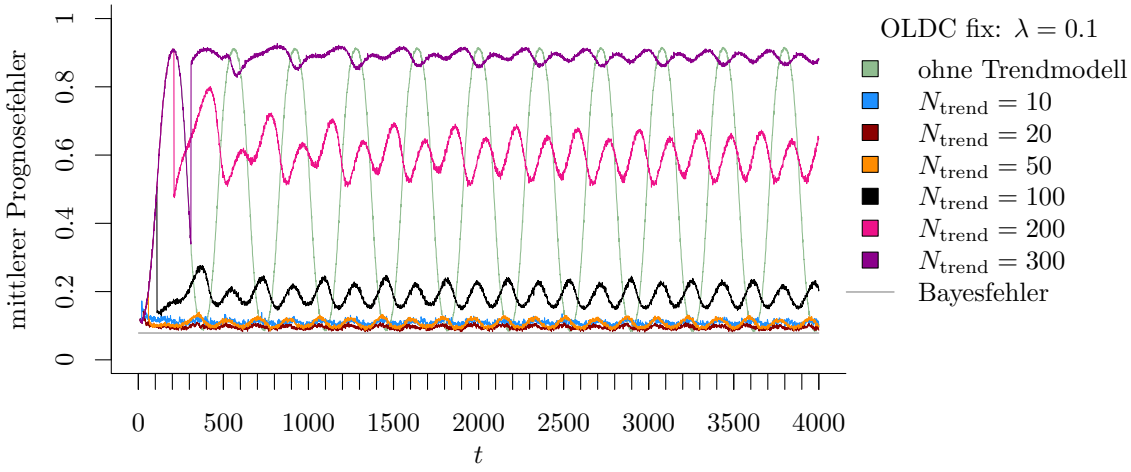


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

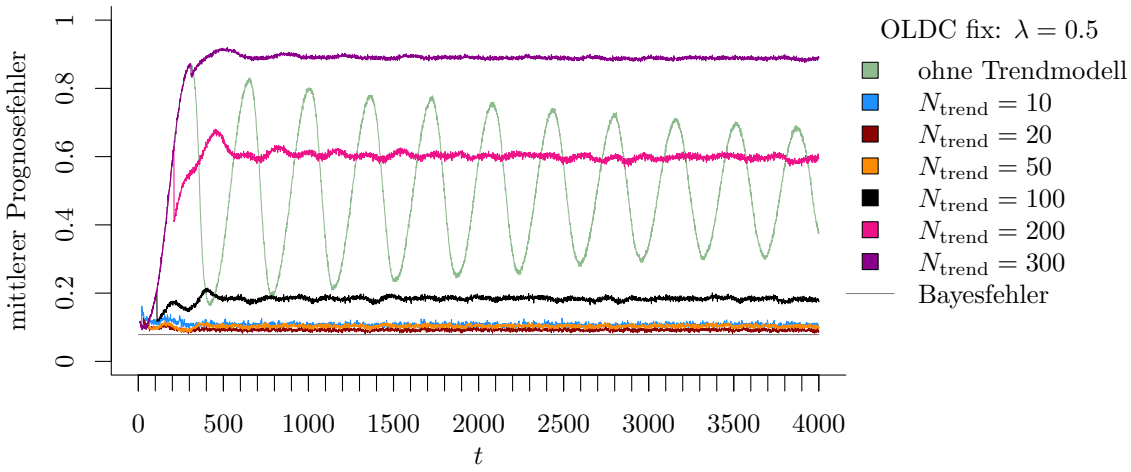


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

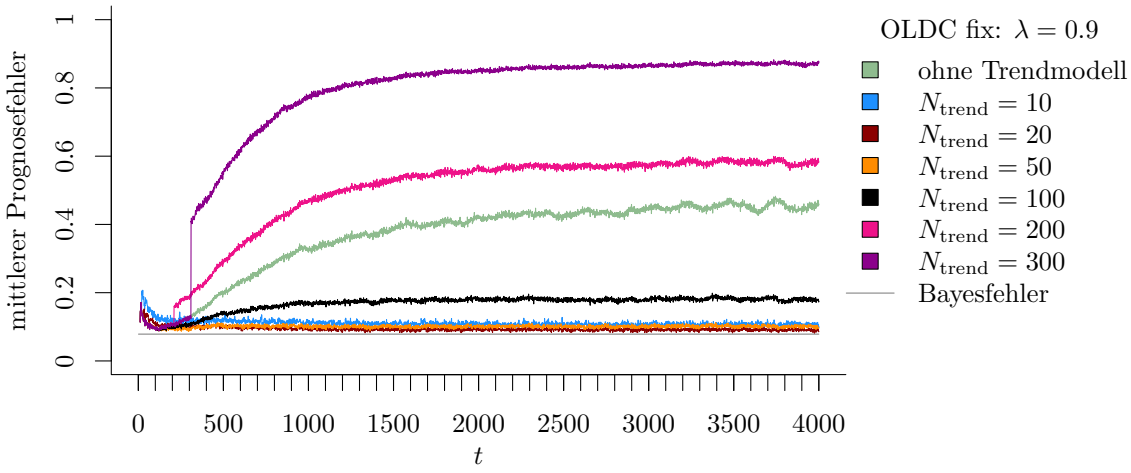
Abbildung 9.19: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreisen“** im zweidimensionalen Raum.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.

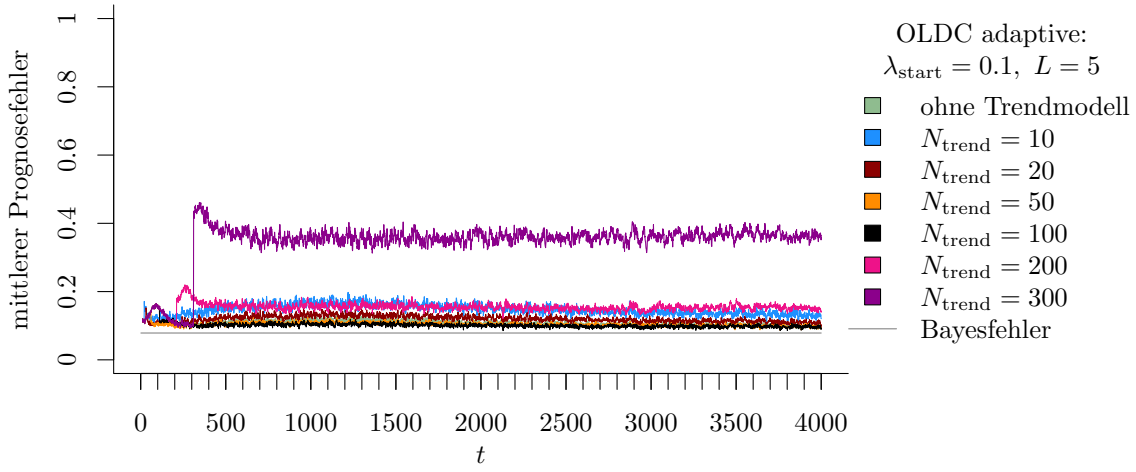


(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.

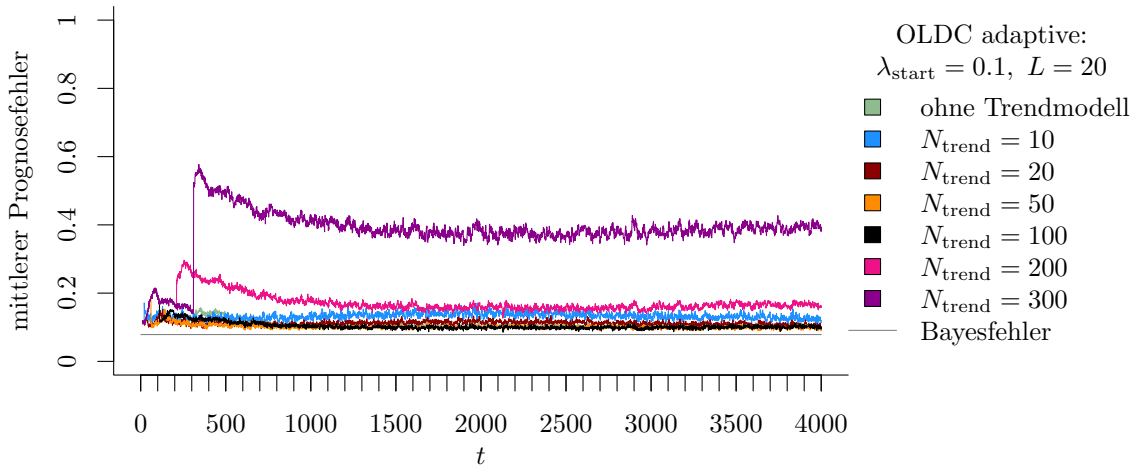


(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

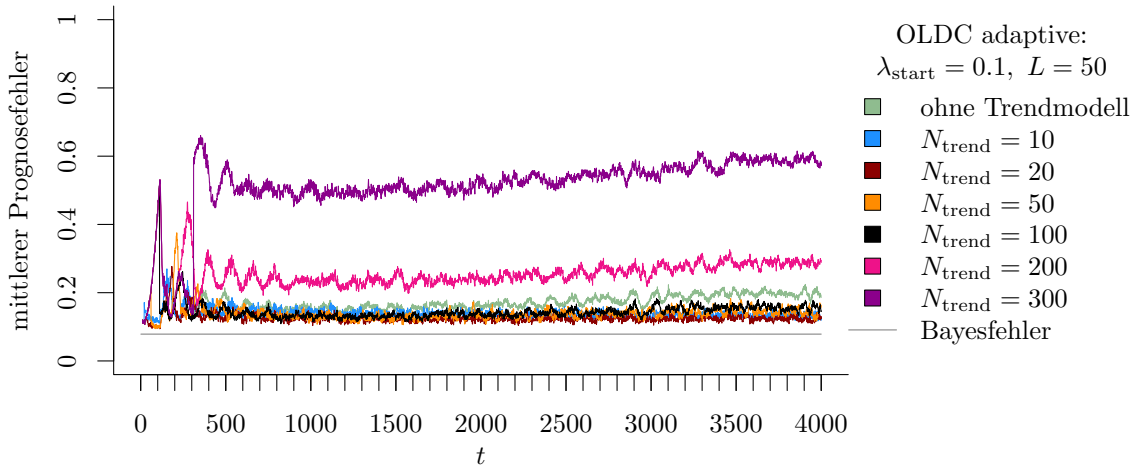
Abbildung 9.20: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreisen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

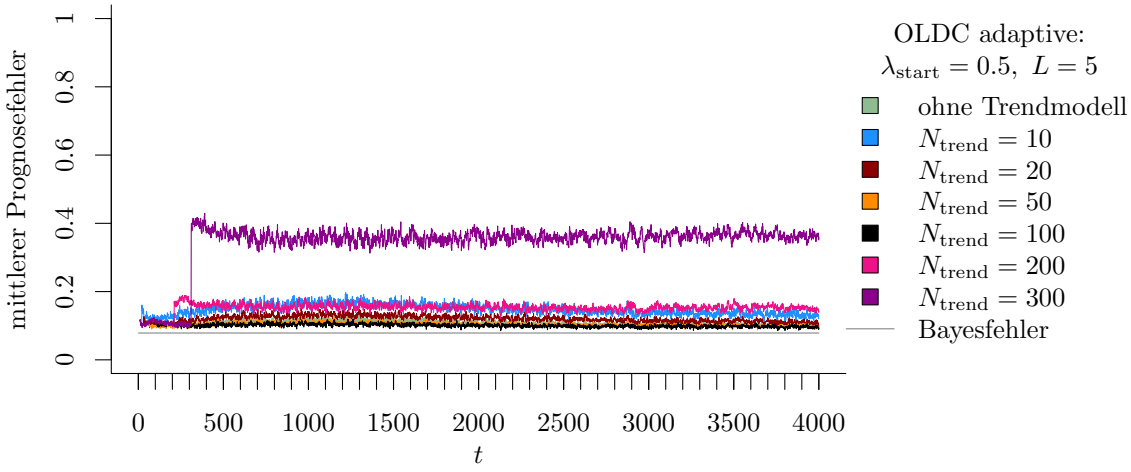


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

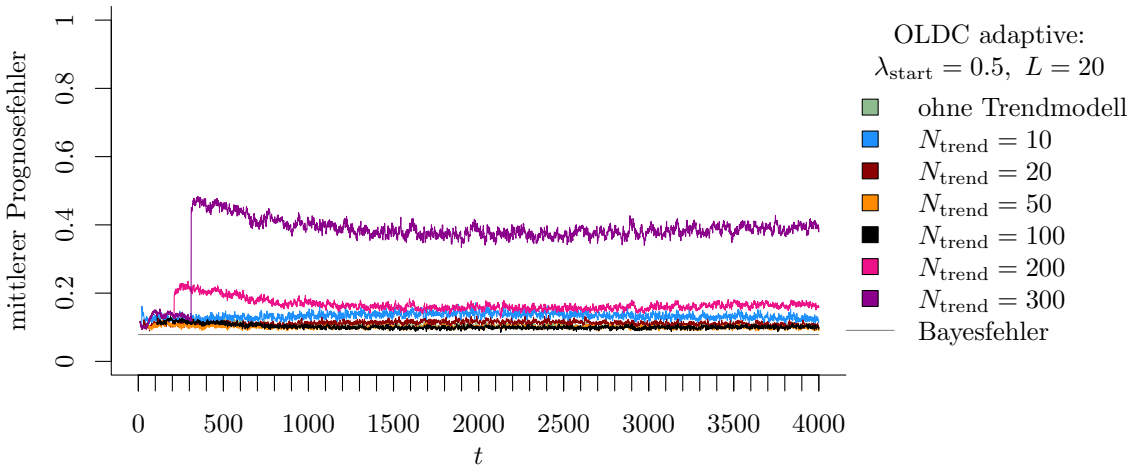


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

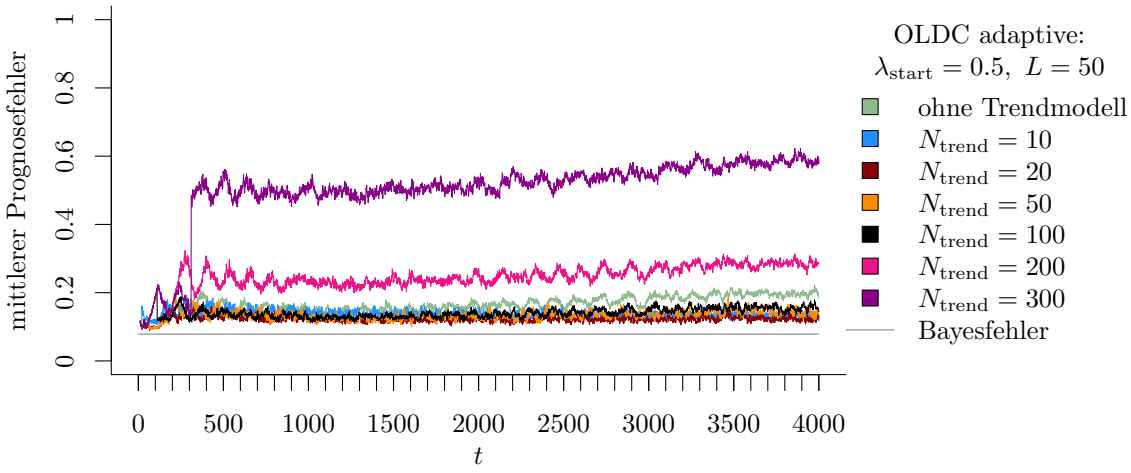
Abbildung 9.21: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreisen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

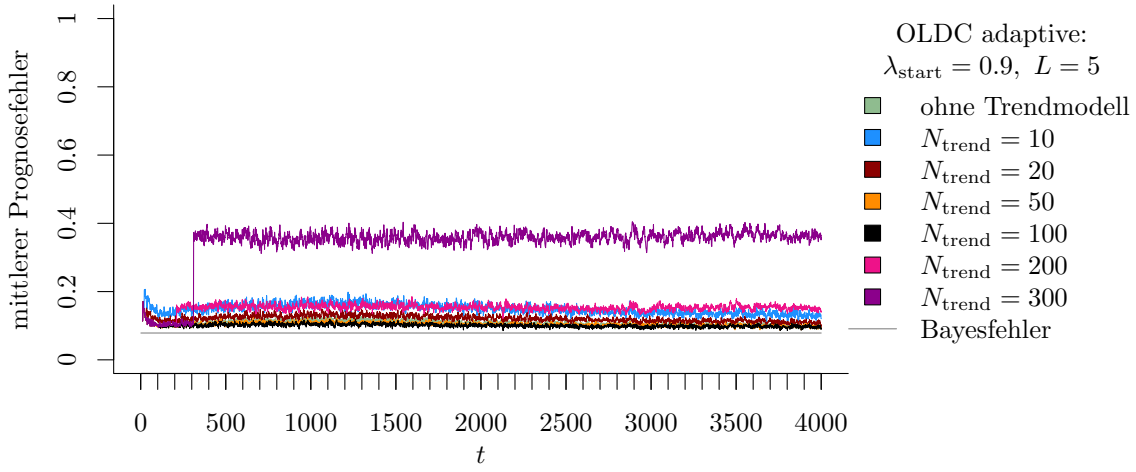


(b) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

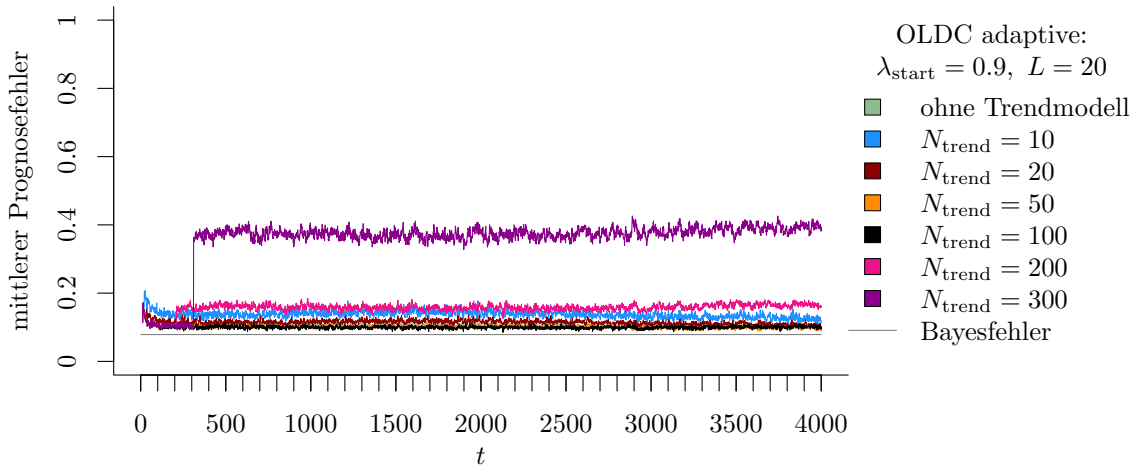


(c) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

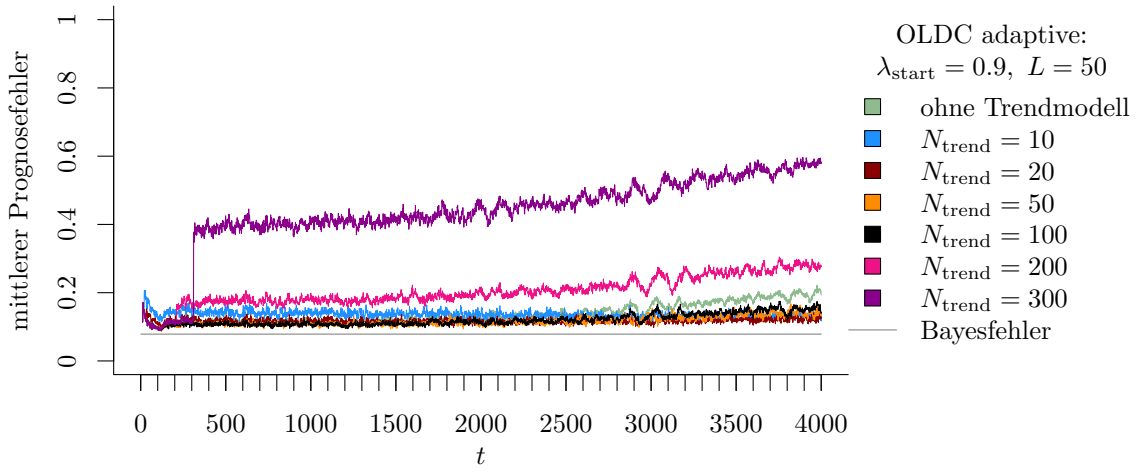
Abbildung 9.22: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreisen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.23: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreisen“** im zweidimensionalen Raum.

Tabelle 9.5: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Kreisen“ ($p = 2$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4976 (0.024)	0.1357 (0.004)	0.1369 (0.007)	0.1	0.4979 (0.007)	5	0.1085 (0.002)
						20	0.1116 (0.002)
						50	0.1759 (0.011)
				0.3	0.4974 (0.004)		
				0.5	0.4976 (0.024)	5	0.1074 (0.002)
						20	0.1072 (0.002)
						50	0.1692 (0.010)
				0.7	0.4695 (0.019)		
				0.9	0.3672 (0.007)	5	0.1076 (0.002)
						20	*0.1044* (0.002)
						50	0.1384 (0.006)
10	*0.1077* (0.002)	0.1443 (0.006)	0.1410 (0.007)	0.1	0.1103 (0.002)	5	0.1470 (0.008)
						20	0.1331 (0.006)
						50	0.1398 (0.007)
				0.3	0.1080 (0.002)		
				0.5	*0.1077* (0.002)	5	0.1472 (0.008)
						20	0.1333 (0.006)
						50	0.1382 (0.007)
				0.7	0.1082 (0.002)		
				0.9	0.1111 (0.003)	5	0.1489 (0.008)
						20	0.1381 (0.007)
						50	0.1367 (0.007)
20	***0.0928*** (0.001)	0.1235 (0.003)	0.1192 (0.003)	0.1	0.0956 (0.001)	5	0.1201 (0.003)
						20	0.1120 (0.003)
						50	0.1259 (0.005)
				0.3	0.0932 (0.001)		
				0.5	***0.0928*** (0.001)	5	0.1201 (0.003)
						20	0.1116 (0.003)
						50	0.1235 (0.005)
				0.7	***0.0928*** (0.001)		
				0.9	0.0942 (0.001)	5	0.1212 (0.003)
						20	0.1143 (0.003)
						50	0.1176 (0.004)
50	0.1032 (0.001)	**0.1111** (0.002)	**0.1113** (0.003)	0.1	0.1078 (0.001)	5	0.1037 (0.002)
						20	0.1016 (0.002)
						50	0.1394 (0.010)
				0.3	0.1044 (0.001)		
				0.5	0.1032 (0.001)	5	0.1033 (0.002)
						20	*0.1003* (0.001)
						50	0.1349 (0.009)
				0.7	0.1024 (0.001)		
				0.9	0.1011 (0.001)	5	0.1038 (0.002)
						20	0.1007 (0.002)
						50	0.1167 (0.005)
100	0.1811 (0.002)	0.1180 (0.002)	0.1224 (0.004)	0.1	0.1914 (0.004)	5	0.1018 (0.002)
						20	0.1052 (0.002)
						50	0.1460 (0.008)
				0.3	0.1841 (0.002)		
				0.5	0.1811 (0.002)	5	0.1008 (0.002)
						20	0.1022 (0.002)
						50	0.1403 (0.007)
				0.7	0.1780 (0.002)		
				0.9	0.1678 (0.002)	5	0.1008 (0.002)
						20	***0.1001*** (0.001)
						50	0.1203 (0.004)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.5832 (0.004)	0.2056 (0.007)	0.2041 (0.012)	0.1	0.5995 (0.008)	5	0.1545 (0.006)
						20	0.1718 (0.006)
						50	0.2552 (0.018)
				0.3	0.5946 (0.004)		
				0.5	0.5832 (0.004)	5	0.1526 (0.006)
						20	0.1646 (0.005)
						50	0.2462 (0.017)
				0.7	0.5627 (0.004)		
				0.9	0.4898 (0.003)	5	*0.1519* (0.006)
						20	0.1565 (0.005)
						50	0.2042 (0.011)
300	0.8565 (0.001)	0.4156 (0.014)	0.4065 (0.022)	0.1	0.8642 (0.002)	5	0.3446 (0.020)
						20	0.3806 (0.017)
						50	0.5115 (0.023)
				0.3	0.8664 (0.001)		
				0.5	0.8565 (0.001)	5	0.3423 (0.020)
						20	0.3709 (0.016)
						50	0.5014 (0.023)
				0.7	0.8308 (0.001)		
				0.9	0.7568 (0.002)	5	*0.3409* (0.020)
						20	0.3560 (0.017)
						50	0.4353 (0.021)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0786 (Standardabweichung 0)

Zusammengefasst lässt sich erkennen:

- Der durchschnittliche Bayesfehler über den gesamten Datenstrom für diese Datensituation beträgt 0.0786 mit einer Standardabweichung von 0.
- Der minimale Wert von 0.0928 wird für *ILDA* bzw. *OLDC* mit $\lambda \in \{0.5, 0.7\}$ und Erweiterung durch lokale lineare Regressionsmodelle auf Fenstern der Breite $N_{\text{trend}} = 20$ erreicht.
- Bei *ILDA* und *OLDC fix* erfolgt direkt eine Verbesserung hinsichtlich des Prognosefehlers durch Einführen lokaler linearer Regressionsmodelle bis $N_{\text{trend}} = 20$. Für größere Werte von N_{trend} steigt der Prognosefehler wieder aufgrund des nicht-linearen Trends der Erwartungswerte. Erst ab $N_{\text{trend}} \geq 200$ werden jedoch die durchschnittlichen Prognosefehler der ursprünglichen Methoden überschritten.
- *QDA-AF*, *LDA-AF* und *OLDC adaptive* (außer $L = 50$): Einführen von Trendmodellen auf kleinen Fenstern der Größe $N_{\text{trend}} = 10$ (bei *OLDC adaptive* teilweise $N_{\text{trend}} \leq 20$) führt erst einmal zu einem größeren durchschnittlichen mittleren Prognosefehler entgegen der ursprünglichen Methoden. Für $N_{\text{trend}} \in \{20, 50\}$ (bei *OLDC adaptive* $N_{\text{trend}} \in \{50, 100\}$) kann eine Verringerung erzielt werden, bevor auch bei diesen Methoden der Prognosefehler für größere Werte wieder steigt.
- In den meisten Fällen ist bei *OLDC* mit fester Lernrate ein höheres λ besser, der Prognosefehler sinkt mit steigendem λ.
- Bei *OLDC* mit adaptiver Lernrate hat das Fenster L zur Adaption der Lernrate einen größeren Einfluss als der Startwert λ_{start}. Bei Erweiterung der Methode und geeigneter Wahl von N_{trend} (hier: $N_{\text{trend}} \in \{20, 50\}$) spielen die Parameter λ_{start} und L jedoch keine so große Rolle mehr.

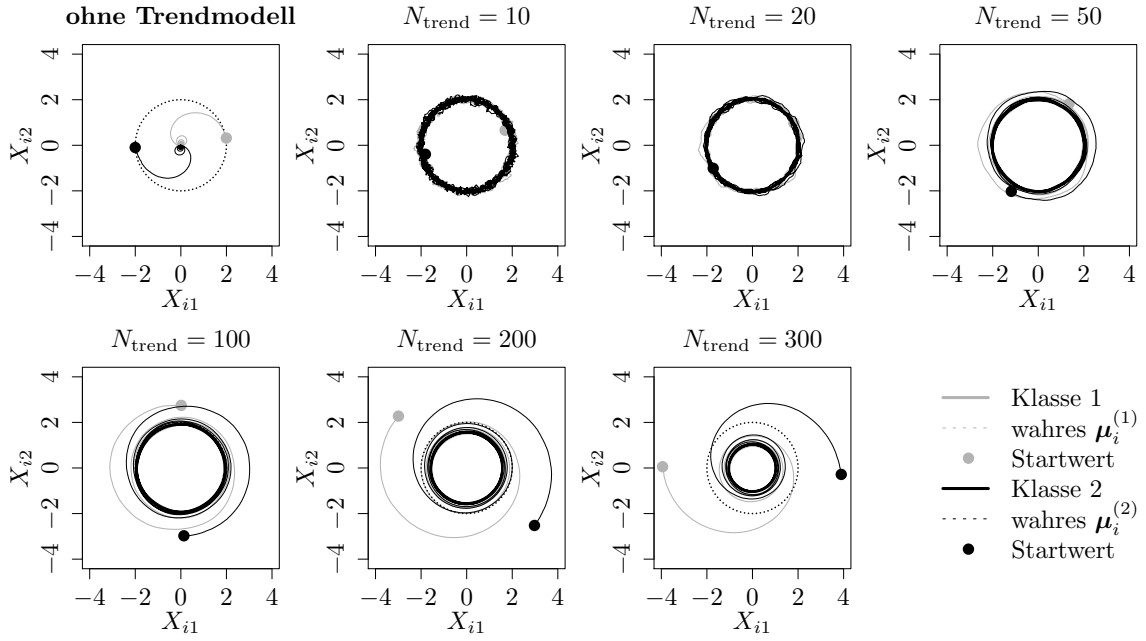


Abbildung 9.24: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren auf der Datensituation **„Kreisen“** für den gesamten Datenstrom. Ergebnisse aus *ILDA* und Erweiterung mit verschiedenen N_{trend}.

In Abbildung 9.24 ist der Verlauf der geschätzten bzw. prognostizierten Erwartungswertvektoren der Update-Methode *ILDA* sowie der Erweiterung durch lokale lineare Regressionsmodelle auf verschiedenen Fenstern N_{trend} dargestellt. Die dicken Punkte markieren erneut jeweils den ersten prognostizierten Erwartungswertvektor im Datenstrom in beiden Klassen. Die gepunkteten Linien zeigen den Verlauf der wahren Erwartungswertvektoren über die Zeit (vgl. Abbildung 9.3 (a) auf Seite 219). In der linken oberen Grafik ist zu erkennen, dass die Erwartungswertvektoren mit der einfachen Update-Methode aufgrund des Drifts nicht gut geschätzt werden können. Durch Erweiterung der Methode und Modellierung des Trends durch lokale lineare Regressionsmodelle sowie anschließender Prognose können die Schätzer deutlich verbessert werden. Für $N_{\text{trend}} = 10$ sind diese noch recht unsicher und weisen eine verhältnismäßig große Varianz auf. Für $N_{\text{trend}} \geq 50$ sind die Fenster für diese Datensituation zu groß. Der nicht-lineare Trend kann nicht mehr gut approximiert werden bzw. die Schätzer für die Erwartungswertvektoren unterscheiden sich wieder stärker von den wahren Werten. Insgesamt scheint von den betrachteten Fenstergrößen $N_{\text{trend}} = 20$ ideal zu sein. Die Analyse der Prognosefehler ließ dasselbe Fazit zu.

Werden die durchschnittlichen euklidischen Abstände zwischen wahren und mittleren prognostizierten Erwartungswertvektoren bzw. mittleren aktualisierten Mittelwertvektoren (ursprüngliche Methoden) über die Zeit als repräsentative Werte für die Prognosegüte herangezogen, ist bei allen Methoden die Erweiterung durch lokale lineare Regressionsmodelle auf kleinen Fenstern am besten (vgl. Tabelle 9.6). Bei allen Update-Methoden für die Diskriminanzanalyse führt die jeweilige Erweiterung zu einer Verringerung der durch-

schnittlichen euklidischen Abstände über die Zeit, mit steigendem N_{trend} steigen diese jedoch wieder an, sodass ab $N_{\text{trend}} = 100$ bzw. $N_{\text{trend}} = 200$ die durchschnittlichen euklidischen Abstände der ursprünglichen Methode wieder überschritten werden.

Wie bei den Prognosefehlern ist auch hier zu sehen, dass bei *OLDC* mit fester Lernrate zunächst ein großes λ am besten für die Prognosegüte ist. Erst bei großen Fenstern $N_{\text{trend}} \geq 200$ steigt der durchschnittliche euklidische Abstand mit steigendem λ leicht an. Bei *OLDC* mit adaptiver Lernrate hängt der euklidische Abstand stark von den Parametern λ_{start} und L ab. Nach Erweiterung der Methoden ist für kleine N_{trend} jedoch keine offensichtliche Abhängigkeit mehr zu erkennen. Die euklidischen Abstände unterscheiden sich für die verschiedenen Parameterkombinationen kaum, die Wahl wird somit weniger wichtig.

Insgesamt lässt sich schlussfolgern, dass für diese Datensituation die Erweiterung der Methoden durch lokale lineare Regressionsmodelle auf recht kleinen Fenstern durch $N_{\text{trend}} \in \{10, 20\}$ am besten ist. In diesem Fall weichen die geschätzten bzw. prognostizierten Erwartungswertvektoren der beiden Klassen im Durchschnitt über den gesamten Datenstrom am geringsten von den wahren Erwartungswertvektoren der Verteilungen ab. Es wird hier ein anderes Fazit geschlossen als anhand der Abbildung 9.24, da die Ergebnisse auf einen einzelnen Wert beschränkt werden. In der Grafik zu $N_{\text{trend}} = 10$ ist zu sehen, dass die wahren Erwartungswertvektoren über den gesamten Datenstrom sehr gut approximiert werden, was zu dem geringen durchschnittlichen euklidischen Abstand führt. Jedoch sind die Schätzungen hier noch recht „unruhig“ (vgl. auch hohe Varianz in Tabelle 9.6). Bei $N_{\text{trend}} = 50$ weichen die Schätzer zu Beginn des Datenstroms stärker von den wahren Erwartungswertvektoren ab, was zu einem höheren durchschnittlichen euklidischen Abstand über die Zeit führt. Im Laufe der Zeit werden die wahren Erwartungswertvektoren jedoch auch gut approximiert und die Schätzungen weisen eine geringere Varianz auf.

Fazit: Der nicht-lineare Trend der Erwartungswertvektoren kann in dieser Datensituation gut durch lokale lineare Regressionsmodelle auf Basis der aktualisierten Mittelwertvektoren der Update-Methoden der Diskriminanzanalyse auf nicht allzu großen Fenstern aus N_{trend} Aktualisierungen approximiert werden. Durch die Erweiterung der Methoden kann die Prognosegüte deutlich verbessert werden, was sich sowohl anhand der Prognosefehler als auch der euklidischen Abstände zwischen wahren und geschätzten bzw. prognostizierten Erwartungswerten als Repräsentant für „Erwartungstreue“ erkennen lässt. Die Erweiterung der Methoden kann somit auch für nicht-lineare Trends der Erwartungswertvektoren und demnach nicht erfüllte Voraussetzungen sinnvoll sein. Dieses Resultat ist insbesondere in Hinblick auf praktische Anwendungen nützlich, da in der Praxis in den seltensten Fällen die Art des concept drifts bzw. die genaue Form und der Verlauf über die Zeit bekannt ist. In dieser Datensituation ist je nach Methode für die Online Diskriminanzanalyse eine Fenstergröße von $N_{\text{trend}} \in \{20, 50\}$ für die Anpassung der lokalen linearen Regressionsmodelle am besten geeignet, um geringe Prognosefehler zu erzielen. In praktischen Anwendungen könnte zunächst eine gewisse Zeit der Parameter N_{trend} auf dem Datenstrom getuned werden, bevor er für die weiteren Modellaktualisierungen festgesetzt wird.

Tabelle 9.6: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Kreisen“ ($p = 2$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

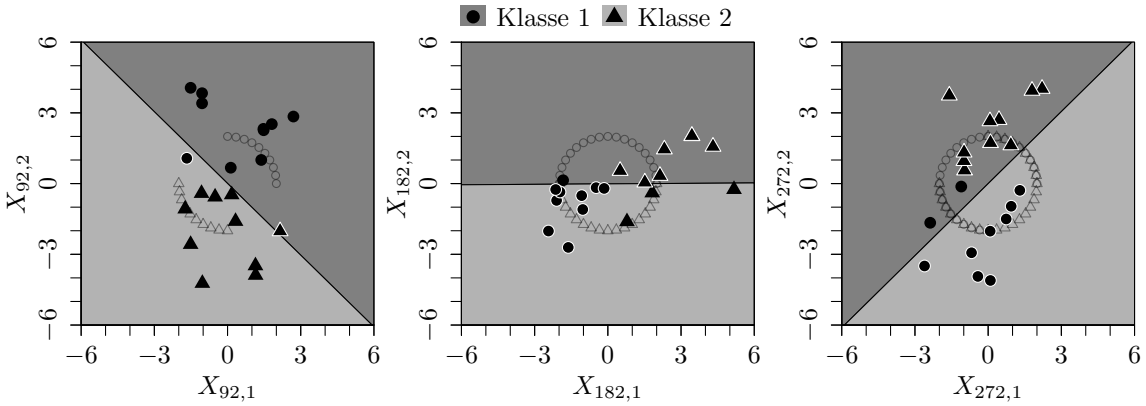
	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	1.9910 (0.007)	0.6203 (0.216)	0.6082 (0.240)	0.1	2.2309 (0.145)	5	0.3036 (0.167)
	1.9933 (0.007)	0.6227 (0.217)	0.6042 (0.241)		2.2261 (0.129)		0.3037 (0.166)
						20	0.5368 (0.160)
							0.5393 (0.163)
						50	1.0806 (0.507)
							1.0695 (0.500)
				0.3	2.0333 (0.016)		
					2.0349 (0.014)		
				0.5	1.9910 (0.007)	5	0.2893 (0.164)
					1.9933 (0.007)		0.2885 (0.163)
						20	0.4934 (0.149)
							0.4918 (0.150)
						50	1.0526 (0.488)
							1.0377 (0.487)
				0.7	1.9513 (0.010)		
					1.9538 (0.010)		
				0.9	1.7798 (0.025)	5	*0.2698* (0.167)
					1.7833 (0.025)		*0.2681* (0.166)
						20	0.3945 (0.149)
							0.3921 (0.149)
						50	0.7678 (0.348)
							0.7674 (0.352)
10	**0.1824** (0.353)	***0.1428*** (0.623)	**0.1430** (0.632)	0.1	0.1832 (0.349)	5	0.1622 (0.901)
	0.1778 (0.354)	**0.1373** (0.615)	***0.1363*** (0.624)		0.1791 (0.351)		0.1567 (0.864)
						20	0.1475 (0.706)
							0.1429 (0.689)
						50	0.1552 (0.794)
							0.1506 (0.793)
				0.3	0.1829 (0.351)		
					0.1787 (0.352)		
				0.5	0.1824 (0.353)	5	0.1617 (0.905)
					0.1778 (0.354)		0.1556 (0.865)
						20	0.1469 (0.703)
							0.1426 (0.700)
						50	0.1516 (0.735)
							0.1495 (0.776)
				0.7	0.1813 (0.358)		
					0.1761 (0.360)		
				0.9	**0.1758** (0.385)	5	0.1623 (0.929)
					0.1690 (0.392)		0.1556 (0.896)
						20	0.1499 (0.766)
							0.1451 (0.774)
						50	0.1506 (0.772)
							0.1438 (0.765)

Fortsetzung auf der nächsten Seite

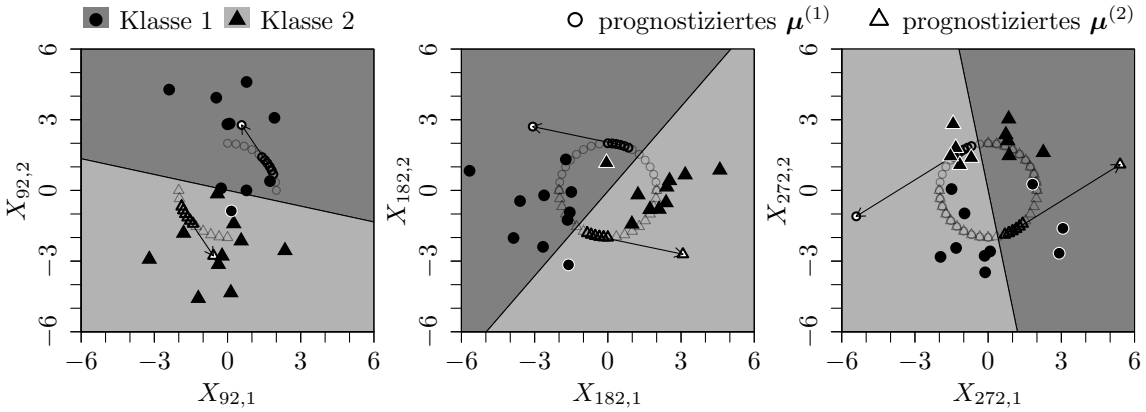
	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
20	0.3321 (0.147)	0.1481 (0.268)	0.1464 (0.272)	0.1	0.3350 (0.145)	5	0.1412 (0.357)
	0.3302 (0.147)	0.1440 (0.270)	0.1417 (0.273)		0.3336 (0.145)		0.1378 (0.356)
						20	0.1544 (0.354)
							0.1498 (0.349)
						50	0.1837 (0.554)
							0.1756 (0.557)
				0.3	0.3340 (0.146)		
					0.3323 (0.146)		
				0.5	0.3321 (0.147)	5	0.1394 (0.357)
					0.3302 (0.147)		0.1353 (0.356)
						20	0.1487 (0.338)
							0.1456 (0.338)
						50	0.1801 (0.514)
							0.1737 (0.519)
				0.7	0.3277 (0.151)		
					0.3254 (0.150)		
				0.9	0.3062 (0.163)	5	***0.1366*** (0.361)
					0.3028 (0.163)		***0.1317*** (0.361)
						20	0.1389 (0.343)
							0.1342 (0.344)
						50	0.1573 (0.408)
							0.1520 (0.415)
50	0.8499 (0.104)	0.3645 (0.149)	0.3567 (0.148)	0.1	0.8551 (0.102)	5	0.2573 (0.164)
	0.8488 (0.104)	0.3651 (0.147)	0.3549 (0.146)		0.8550 (0.103)		0.2572 (0.161)
						20	0.3670 (0.231)
							0.3670 (0.234)
						50	0.4792 (1.754)
							0.4598 (1.747)
				0.3	0.8533 (0.103)		
					0.8527 (0.103)		
				0.5	0.8499 (0.104)	5	0.2538 (0.162)
					0.8488 (0.104)		0.2531 (0.160)
						20	0.3528 (0.200)
							0.3513 (0.200)
						50	0.4736 (1.502)
							0.4643 (1.539)
				0.7	0.8410 (0.108)		
					0.8393 (0.108)		
				0.9	0.7918 (0.122)	5	*0.2460* (0.161)
					0.7893 (0.120)		*0.2443* (0.159)
						20	0.3099 (0.182)
							0.3072 (0.180)
						50	0.3883 (0.758)
							0.3860 (0.781)
100	1.6784 (0.072)	0.9010 (0.210)	0.8826 (0.213)	0.1	1.6786 (0.071)	5	0.6857 (0.138)
	1.6790 (0.074)	0.9028 (0.210)	0.8797 (0.214)		1.6811 (0.074)		0.6860 (0.135)
						20	0.8828 (0.220)
							0.8847 (0.222)
						50	1.2276 (1.973)
							1.2206 (2.025)
				0.3	1.6787 (0.070)		
					1.6803 (0.073)		
				0.5	1.6784 (0.072)	5	0.6805 (0.137)
					1.6790 (0.074)		0.6798 (0.134)
						20	0.8543 (0.201)
							0.8534 (0.197)
						50	1.2258 (1.825)
							1.2302 (1.911)
				0.7	1.6742 (0.077)		
					1.6739 (0.079)		
				0.9	1.6226 (0.096)	5	*0.6693* (0.136)
					1.6220 (0.097)		*0.6677* (0.132)
						20	0.7794 (0.171)
							0.7777 (0.167)
						50	1.0277 (1.001)
							1.0378 (1.012)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	2.8801 (0.051)	2.1782 (0.263)	2.1400 (0.258)	0.1	2.8424 (0.055)	5	1.8723 (0.109)
	2.8795 (0.052)	2.1792 (0.268)	2.1348 (0.262)		2.8443 (0.055)		1.8698 (0.107)
						20	2.1472 (0.184)
							2.1463 (0.184)
						50	2.7627 (1.112)
							2.7471 (1.092)
				0.3	2.8564 (0.050)		
					2.8571 (0.051)		
				0.5	2.8801 (0.051)	5	1.8671 (0.108)
					2.8795 (0.052)		1.8647 (0.106)
						20	2.1091 (0.169)
							2.1070 (0.164)
						50	2.7409 (1.081)
							2.7305 (1.077)
				0.7	2.9205 (0.057)		
					2.9191 (0.058)		
				0.9	2.9892 (0.077)	5	*1.8570* (0.107)
					2.9885 (0.078)		*1.8545* (0.105)
						20	2.0145 (0.142)
							2.0117 (0.139)
						50	2.4250 (0.678)
							2.4297 (0.686)
300	3.1275 (0.038)	3.0487 (0.177)	3.0034 (0.167)	0.1	3.0447 (0.048)	5	2.8021 (0.090)
	3.1247 (0.039)	3.0490 (0.181)	2.9994 (0.169)		3.0422 (0.046)		2.7991 (0.089)
						20	3.0248 (0.135)
							3.0212 (0.137)
						50	3.4614 (0.505)
							3.4495 (0.494)
				0.3	3.0749 (0.038)		
					3.0719 (0.039)		
				0.5	3.1275 (0.038)	5	2.8017 (0.089)
					3.1247 (0.039)		2.7994 (0.089)
						20	3.0039 (0.128)
							3.0005 (0.128)
						50	3.4579 (0.489)
							3.4529 (0.486)
				0.7	3.2251 (0.044)		
					3.2225 (0.046)		
				0.9	3.4689 (0.066)	5	*2.7991* (0.089)
					3.4676 (0.067)		*2.7970* (0.089)
						20	2.9412 (0.114)
							2.9385 (0.114)
						50	3.2564 (0.342)
							3.2623 (0.346)



(a) Ursprüngliche Methode *ILDA* bzw. *OLDC* mit $\lambda = 0.5$.



(b) Erweiterung durch lokale lineare Regressionsmodelle mit $N_{\text{trend}} = 50$.

Abbildung 9.25: Veränderung der Klassifikationsgrenze der LDA im Laufe der Zeit durch Aktualisierung des Modells mit neuen Beobachtungen aus Verteilungen mit verschobenem Erwartungswertvektor und Auswirkung auf den Prognosefehler für die Datensituation **„Kreisen“**. Graue ∘ und △ veranschaulichen Erwartungswertvektoren der Verteilungen von Klasse 1 bzw. Klasse 2, aus denen Beobachtungen im Datenstrom realisiert werden, die in (a) für die Anpassung der LDA herangezogen werden.
In (b): Schwarze ∘ und △ skizzieren Erwartungswertvektoren, die von den aktualisierten Mittelwerten der letzten $N_{\text{trend}} = 50$ Updates der LDA für das lokale lineare Regressionsmodell repräsentiert werden.

Datensituation „Kreuzen“ ($p = 2$) Bei der Datensituation der kreuzenden Klassen (vgl. Seite 220 f.) steigt der Prognosefehler für die Methoden ohne Anpassung an einen concept drift (grüne Kurven in Abbildungen 9.27 (a) und 9.28 (b)) im Laufe der Zeit immer weiter an, sodass fast der Wert 1 erreicht wird und alle Beobachtungen des Testdatensatzes falsch prognostiziert werden. Zum Zeitpunkt $t = 2000$ liegt der Prognosefehler bei etwa 0.5, was sich dadurch erklären lässt, dass sich zu diesem Zeitpunkt die Verteilungen in beiden Klassen vollständig überlagern (vgl. Abbildung 9.3 (b) auf Seite 219). Daher kann dieser Wert nicht unterschritten werden, was auch anhand des Verlaufs des Bayesfehlers deutlich wird. Zum Ende des Datenstroms fällt der Prognosefehler wieder auf 0.5 ab. Zu diesem Zeitpunkt liegen in beiden Klassen Beobachtungen aus Verteilungen mit Erwartungswertvektoren auf den zwei (in der Mitte) kreuzenden Geraden vor, sodass die lineare Trenngerade durch die LDA plötzlich von senkrecht zur x-Achse auf vertikal wechselt (vgl. Veranschaulichung der Klassifikationsregel in Abbildung 9.33 am Ende des Abschnittes auf Seite 285).

Eine höhere feste Lernrate λ bei *OLDC* und demnach stärkere Gewichtung der aktuellen Beobachtungen bei der Aktualisierung der LDA führt dazu, dass der Prognosefehler nach $t = 2000$ nicht mehr so stark ansteigt und zu einem früheren Zeitpunkt wieder kleiner wird (vgl. Abbildung 9.28 (c)). Dies ist auch bei *OLDC* mit adaptiver Lernrate der Fall. Unabhängig vom Startwert λ_{start} und Fenster L kann sich die Lernrate im Laufe des Datenstroms so gut an den concept drift adaptieren, dass der Prognosefehler den Bayesfehler zu jedem Zeitpunkt annähernd approximiert (vgl. Abbildungen 9.29–9.31).

Für alle Update-Methoden ist zu erkennen, dass die Einführung lokaler linearer Regressionsmodelle – unabhängig vom Wert für N_{trend} – den Prognosefehler des ursprünglichen Modells bis zum Zeitpunkt $t = 2000$ nicht verringern kann, da dieser bereits sehr nahe am Bayesfehler liegt. Für kleinere Werte von N_{trend} liegt die Kurve des Prognosefehlers sogar etwas über jener der ursprünglichen nicht erweiterten Methode (grüne Kurven). Ab dem Zeitpunkt $t = 2000$, sobald sich die Verteilungen in beiden Klassen überlappen bzw. sich die Erwartungswertvektoren kreuzen, wird der positive Effekt der Erweiterung der Methoden *ILDA* und *OLDC fix* jedoch deutlich (vgl. Abbildungen 9.27 (a) und 9.28). Statt weiter anzusteigen, fällt der Prognosefehler relativ schnell wieder ab, sodass die Kurven dem Verlauf des Bayesfehlers für den Datenstrom ähneln. Auch bei den anderen Methoden kann der Prognosefehler zu vielen Zeitpunkten bei genügend großem Fenster N_{trend} für die lokalen linearen Regressionsmodelle noch verkleinert werden.

Lediglich bei großem Fenster L zur Adaption der Lernrate bei *OLDC adaptive* kommt es zu kleinen „Ausschlägen“ des Prognosefehlers. Dies lässt sich durch den Verlauf der Lernrate über die Zeit für die verschiedenen Parameterkombinationen erklären, welcher in Abbildung 9.26 dargestellt ist. Unabhängig vom Startwert λ_{start} steigt die Lernrate auf Werte > 0.9 bei Adaption auf kleinen Fenstern $L < 50$. Bei großem Fenster $L = 50$ nimmt die Lernrate jedoch im Gegenteil unabhängig vom Startwert λ_{start} über die Zeit wieder ab. Da aufgrund des andauernden Drifts in dieser Datensituation eine hohe Lernrate geeignet ist, wird deutlich, dass ein großes Fenster L zur Adaption ungeeignet ist. Der starke

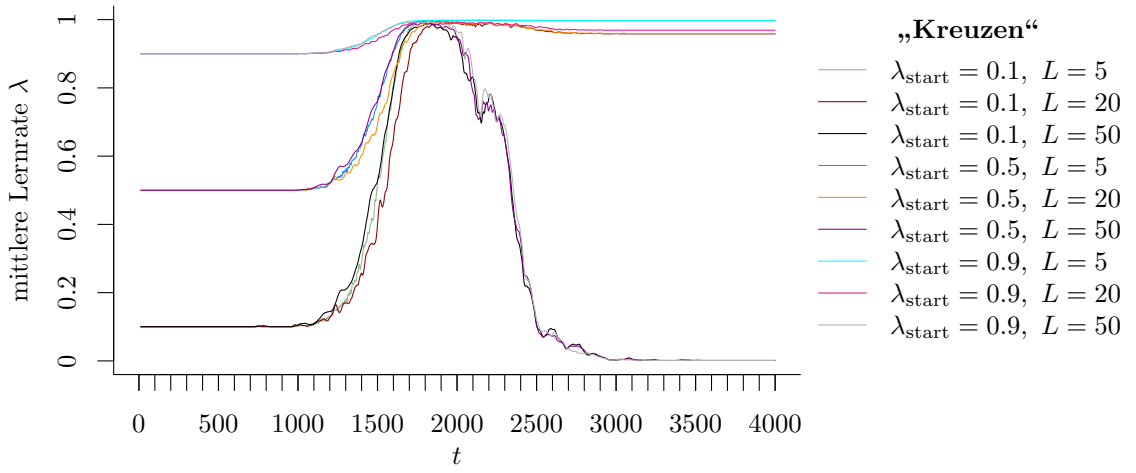


Abbildung 9.26: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **„Kreuzen“**.

Abfall der Lernrate resultiert in vergleichsweise höheren Prognosefehlern um den Zeitpunkt $t = 2500$ (vgl. Abbildungen 9.29–9.31), da bei kleinen Lernraten aktuelle Beobachtungen bei den Modellaktualisierungen weniger stark gewichtet werden und demnach keine gute Anpassung an einen Drift erfolgt, sondern ein veraltetes Modell betrachtet wird.

Die starke Reduzierung des Prognosefehlers ab Mitte des Datenstroms wirkt sich auch auf die einzelne quantitative Maßzahl, genauer den durchschnittlichen mittleren Prognosefehler über die Zeit für alle Methoden aus (vgl. Tabelle 9.7). Bei *ILDA* und *OLDC fix* wird die Prognosegüte durch Einführung der Regressionsmodelle direkt besser. Zudem sinkt der durchschnittliche mittlere Prognosefehler für alle Methoden für steigendes N_{trend} immer weiter, was auf die erfüllte Annahme des linearen Trends der Erwartungswertvektoren zurückzuführen ist.

Für *QDA-AF* und *LDA-AF* ist der durchschnittliche mittlere Prognosefehler über die Zeit für kleine Fenster $N_{\text{trend}} \in \{10, 20, 50\}$ zunächst größer als jener der Methode ohne Trendmodell. Ein größerer Wert für N_{trend} kann jedoch auch hier eine Verbesserung hinsichtlich der Prognosegüte entgegen der ursprünglichen Methoden erzielen. Der Prognosefehler nimmt zu Beginn für *QDA-AF* und *LDA-AF* (vgl. Abbildung 9.27 (b) und (c)) einen Wert von 0.5 an, da diese Methoden datenunabhängig initialisiert werden.

Für *OLDC adaptive* ist ein ähnliches Muster zu erkennen. Für die meisten Fenstergrößen L und Startwerte λ_{start} wird der Prognosefehler erstmals bei Erweiterung des Modells durch lokale lineare Regressionsmodelle auf Fenstern der Größe $N_{\text{trend}} = 100$ verringert. Für große Fenster $L = 50$ werden größere Werte für N_{trend} für kleinere Prognosefehler benötigt. Der Parameter λ_{start} hat dabei kaum einen Einfluss. Die Prognosefehler unterscheiden sich bei festem N_{trend} und identischem L für verschiedene Startwerte λ_{start} nur sehr gering. Ein größeres Fenster L führt teilweise zu leicht erhöhten Prognosefehlern wie bereits oben erläutert. Generell ist zu sagen, dass die durchschnittlichen mittleren Prognosefehler über die Zeit auch ohne Erweiterung des Modells bereits sehr nahe am Bayesfehler liegen. Durch

die Integration lokaler linearer Regressionsmodelle auf genügend großen Fenstern N_{trend} kann der Prognosefehler jedoch noch leicht verringert werden.

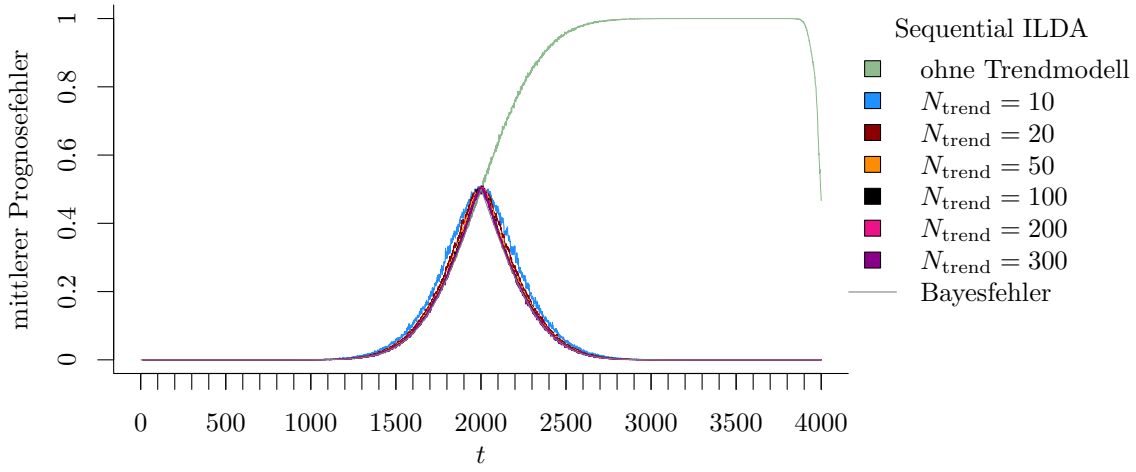
Eine weitere Erkenntnis ist, dass falls die Erweiterung durch Trendmodelle betrachtet wird, die Wahl von λ bei *OLDC* unerheblich ist. Während in der ursprünglichen Methode der durchschnittliche mittlere Prognosefehler über die Zeit mit größerem λ stark sinkt (vgl. erste „Zeile“ in Spalte „OLDC fix“ in Tabelle 9.7), was auch anhand der grünen Kurven in Abbildung 9.28 deutlich wird, hat λ nach Integration der Trendmodelle keinen Einfluss mehr auf den Prognosefehler (vgl. Spalte „OLDC fix“ in Tabelle 9.7). Daher ist sogar der Prognosefehler bei $\lambda = 0.1$ und Trendmodellen auf Fenstern der Größe $N_{\text{trend}} = 10$ kleiner als im ursprünglichen nicht erweiterten Modell mit fester Lernrate $\lambda = 0.9$.

Generell wird deutlich, dass die Auswahl der Update-Methode für die Diskriminanzanalyse mit steigendem N_{trend} für das inkludierte Trendmodell zur Modellierung und Prognose des Trends der Erwartungswertvektoren weniger relevant ist. Während die Prognosefehler sich bei den ursprünglichen Methoden deutlich voneinander unterscheiden (vgl. grüne Kurven in Abbildungen 9.27–9.31 bzw. erste „Zeile“ in Tabelle 9.7), variieren die Prognosefehler bei entsprechend hohem N_{trend} für die verschiedenen Update-Methoden kaum noch. Durch die entwickelte Erweiterung kann der Prognosefehler aller Methoden im Datenstrom zu vielen Zeitpunkten annähernd auf den Wert des Bayesfehlers minimiert werden.

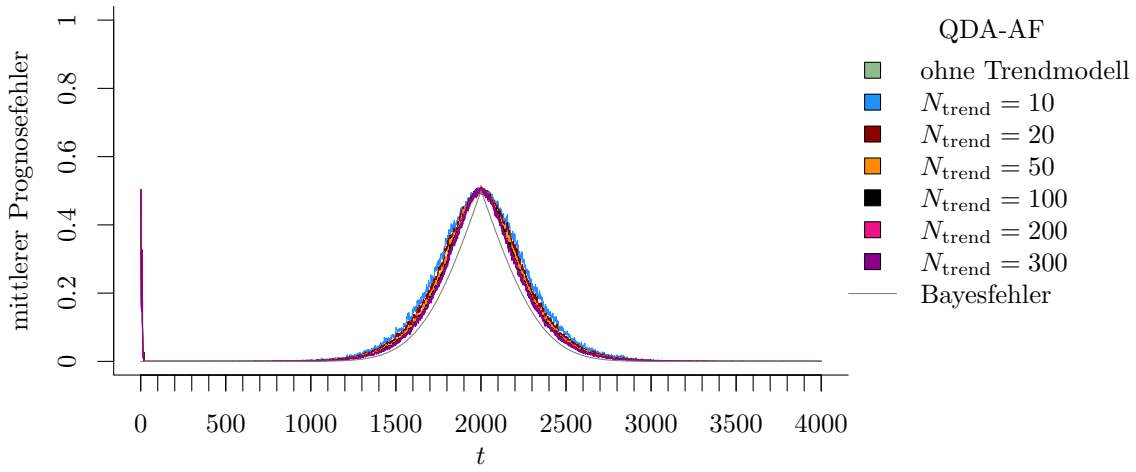
Die durchschnittliche Varianz des Prognosefehlers über die Zeit ist bereits in allen ursprünglichen Methoden sehr gering (< 0.002) und wird durch die Erweiterung der Methoden nicht signifikant vergrößert (vgl. Tabelle 9.7).

Zusammengefasst lässt sich für diese Datensituation schließen:

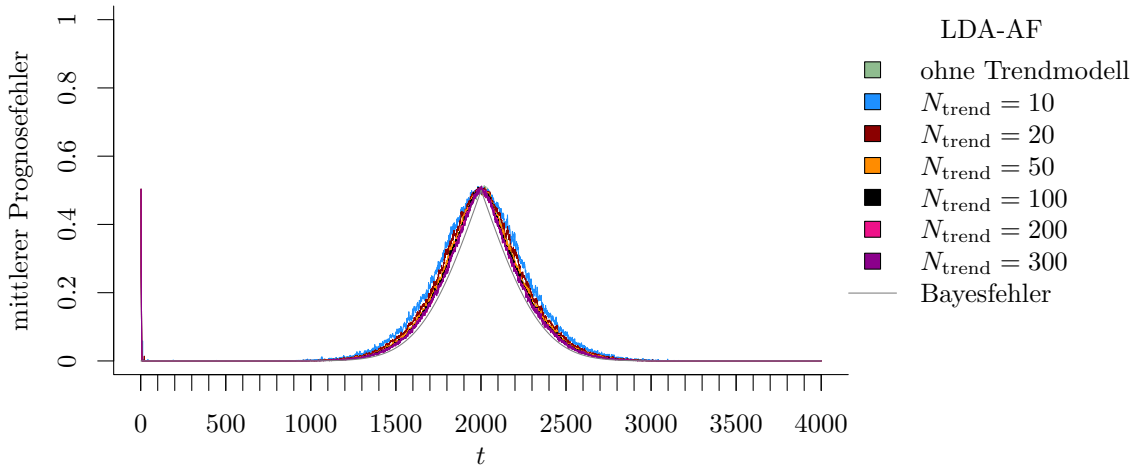
- Der durchschnittliche Bayesfehler über den gesamten Datenstrom für diese Datensituation beträgt 0.0564 mit einer Standardabweichung von 0.116.
- Die Prognosegüte wird durch Integration lokaler linearer Regressionsmodelle für *IL-DA* und *OLDC fix* direkt verbessert.
- Für *QDA-AF* und *LDA-AF* sinkt der durchschnittliche mittlere Prognosefehler über die Zeit bei Einführen der Trendmodelle im Gegensatz zu jenem der ursprünglichen Methoden ab der Betrachtung von $N_{\text{trend}} = 100$, für *OLDC adaptive* für die betrachteten Startwerte λ_{start} und Fenster $L < 50$ ab $N_{\text{trend}} = 100$, bei $L = 50$ erst für größere Werte für N_{trend}.
- *QDA-AF*, *LDA-AF* und *OLDC adaptive* weisen auch in der ursprünglichen Variante geringe Prognosefehler nahe am Bayesfehler auf.
- Mit steigendem Wert für N_{trend} sinkt der Prognosefehler in allen Methoden.
- Die Wahl der Lernrate λ bei *OLDC* spielt nach Erweiterung der Methode keine Rolle mehr, der Prognosefehler ist annähernd unabhängig von der Lernrate.
- Bei genügend großem Fenster N_{trend} bei Einbindung lokaler linearer Regressionsmodelle ist die Auswahl der Ausgangsmethode für die Diskriminanzanalyse weniger relevant. Die Prognosefehler liegen alle sehr nahe beieinander.



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

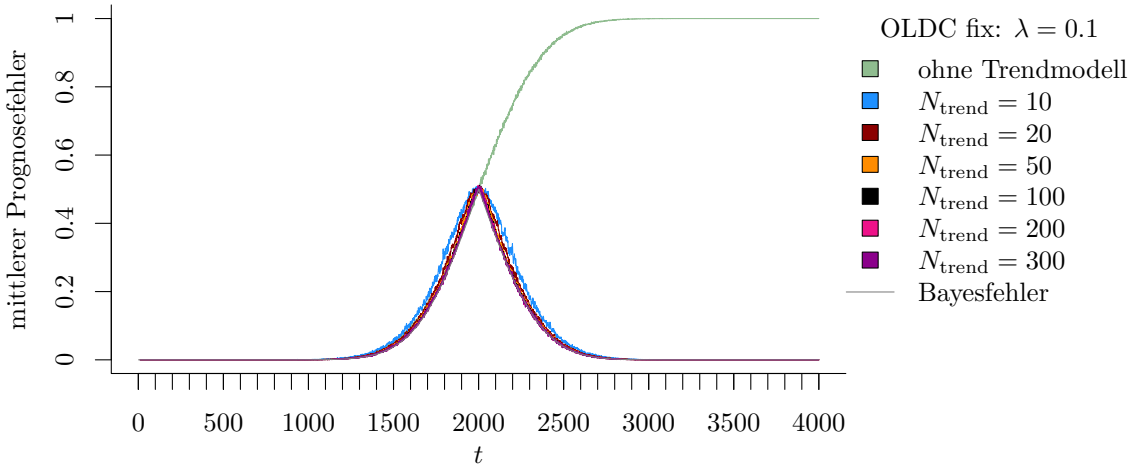


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

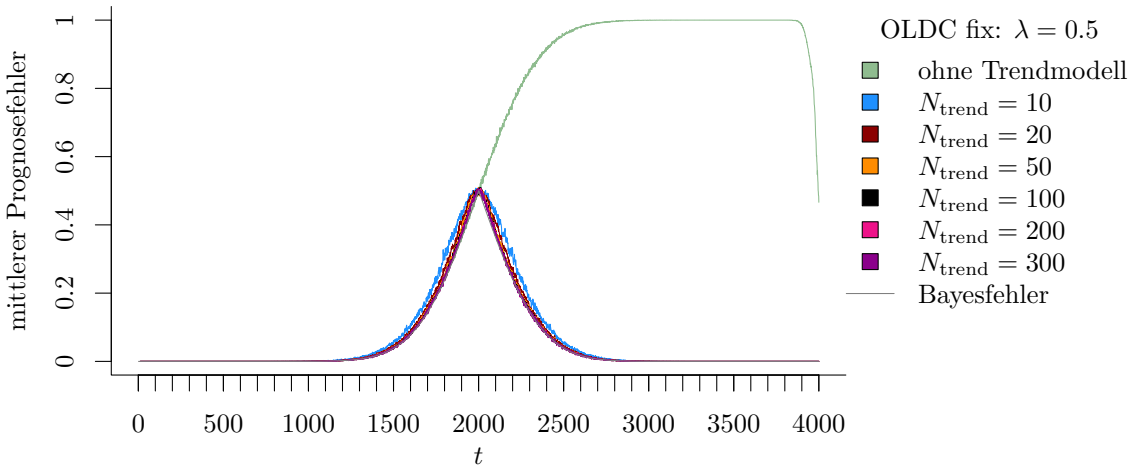


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

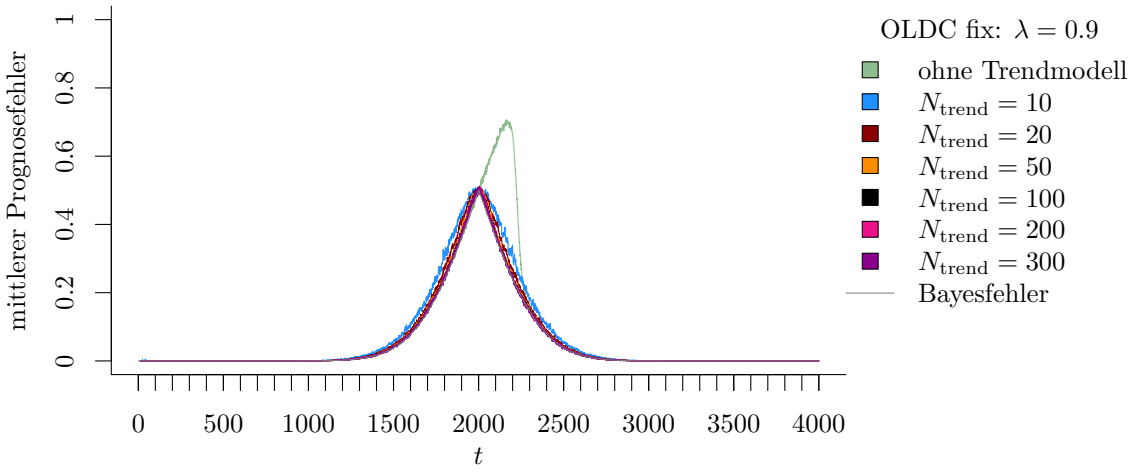
Abbildung 9.27: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreuzen“** im zweidimensionalen Raum.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.

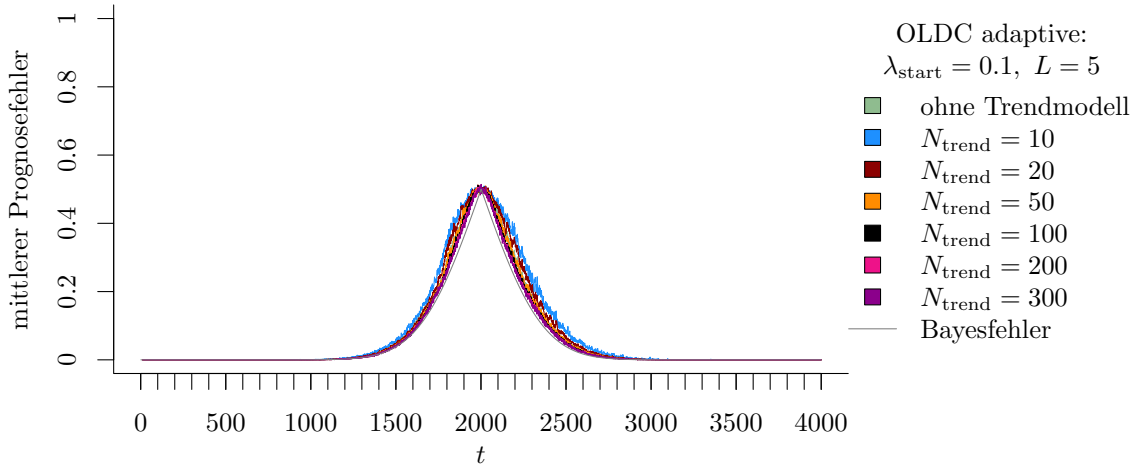


(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.

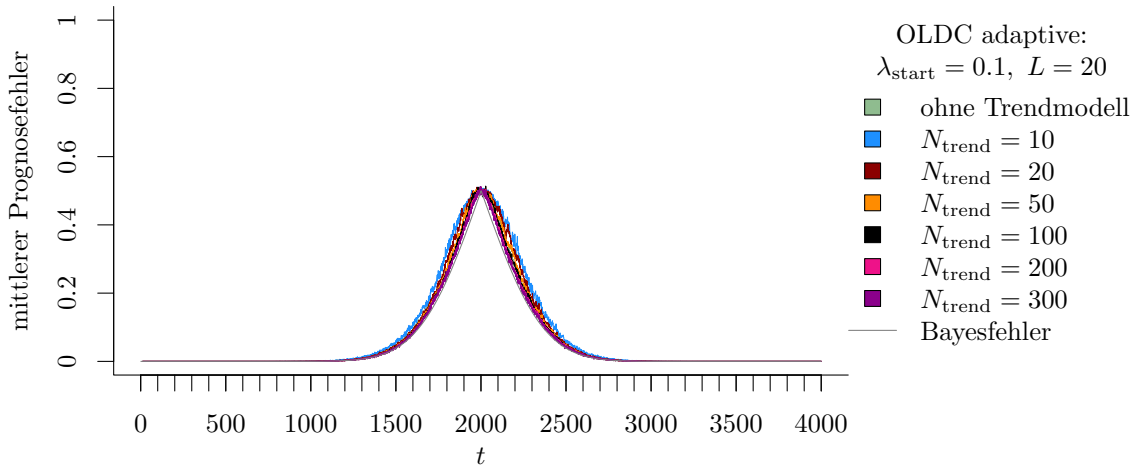


(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

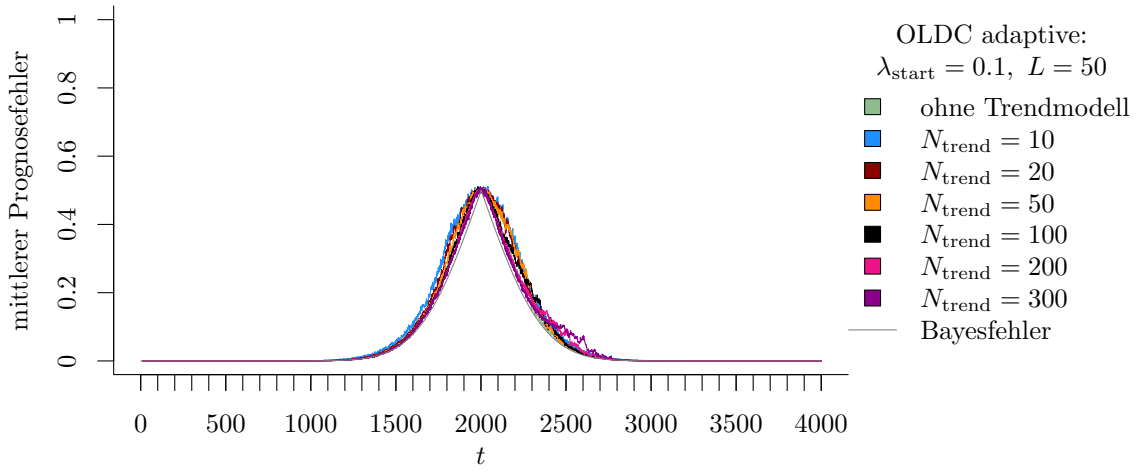
Abbildung 9.28: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreuzen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

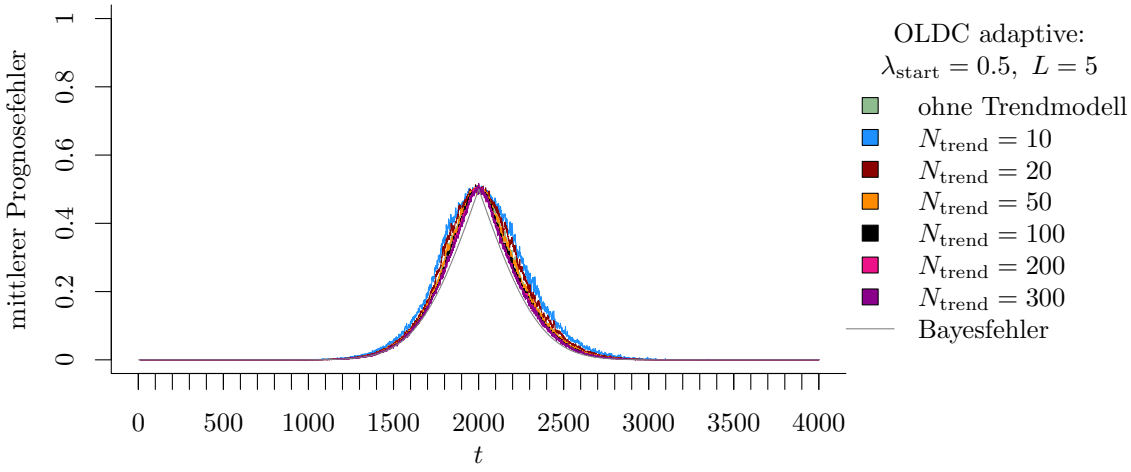


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

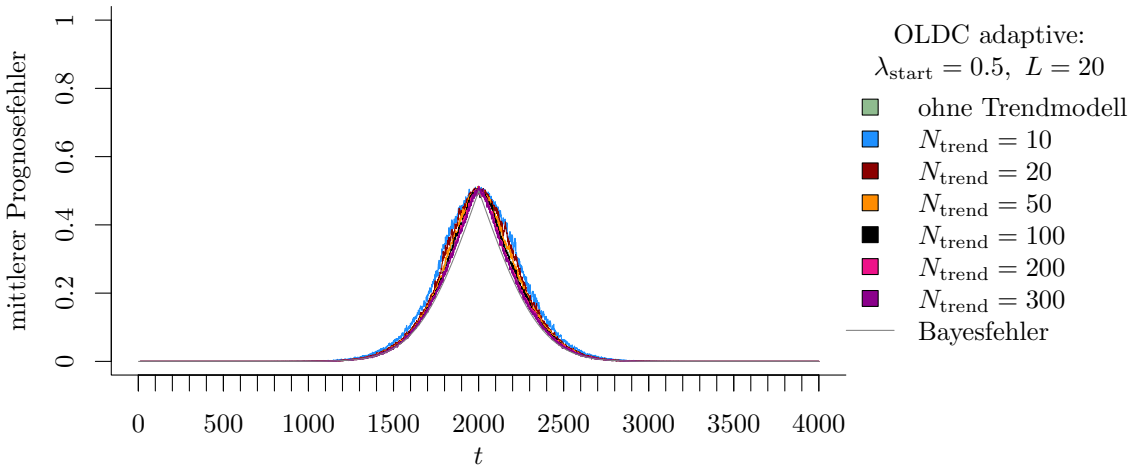


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

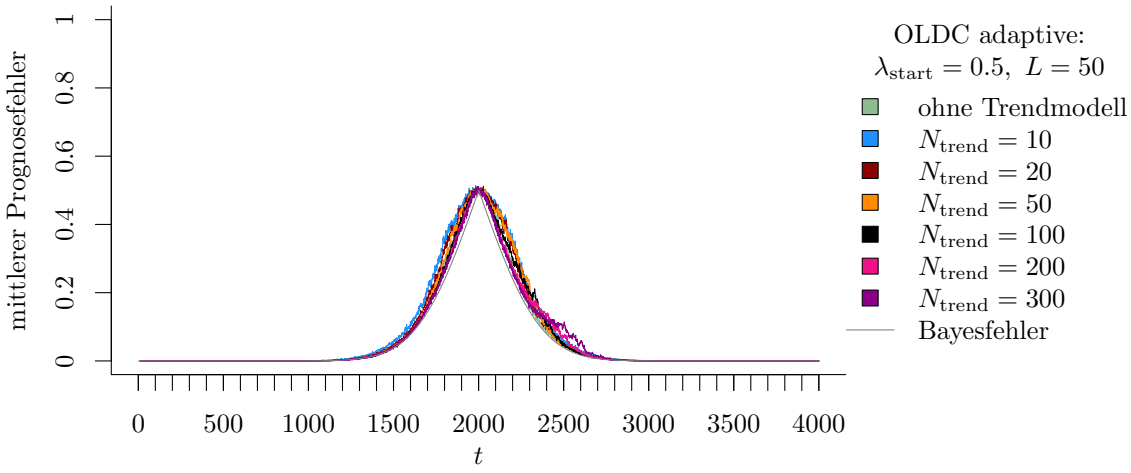
Abbildung 9.29: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreuzen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

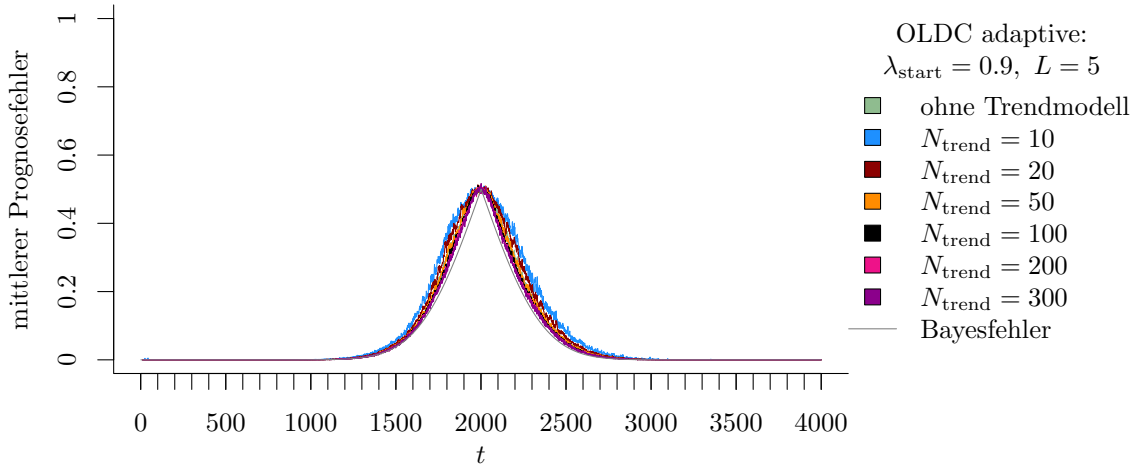


(b) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

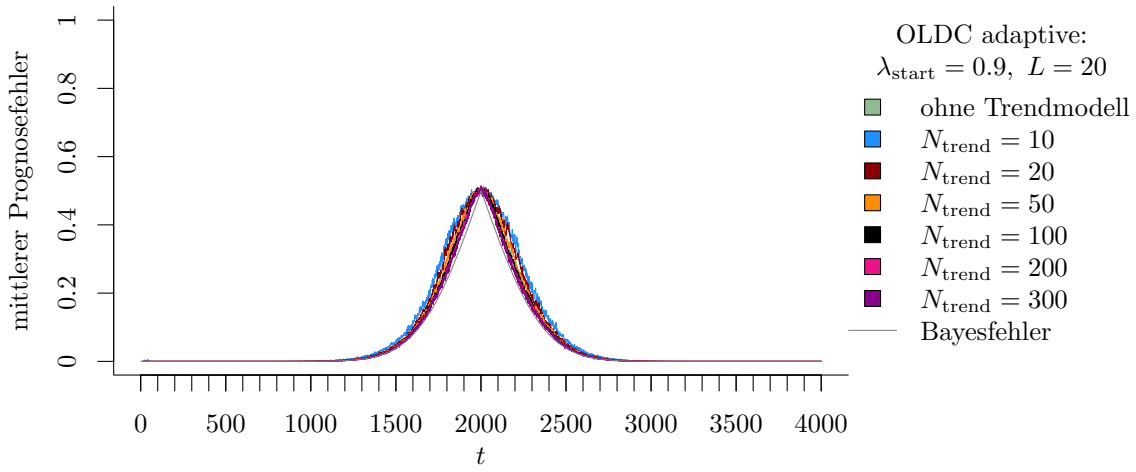


(c) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

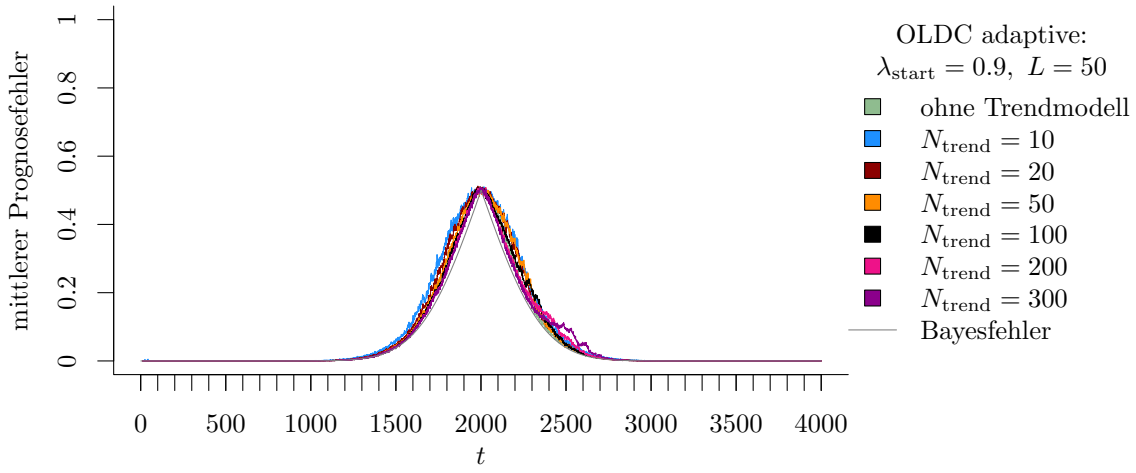
Abbildung 9.30: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreuzen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.31: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Kreuzen“** im zweidimensionalen Raum.

Tabelle 9.7: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Kreuzen“** ($p = 2$) getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4964 (0.001)	0.0708 (0.001)	0.0660 (0.001)	0.1	0.5017 (0.001)	5	0.0669 (0.001)
						20	0.0634 (0.001)
						50	0.0641 (0.001)
				0.3	0.5010 (0.000)		
				0.5	0.4964 (0.001)	5	0.0660 (0.001)
						20	*0.0627* (0.001)
						50	0.0638 (0.001)
				0.7	0.2135 (0.001)		
				0.9	0.0736 (0.001)	5	0.0665 (0.001)
						20	0.0635 (0.001)
						50	0.0641 (0.001)
10	0.0694 (0.001)	0.0808 (0.002)	0.0782 (0.002)	0.1	0.0694 (0.001)	5	0.0774 (0.002)
						20	0.0731 (0.002)
						50	0.0734 (0.002)
				0.3	0.0694 (0.001)		
				0.5	0.0694 (0.001)	5	0.0775 (0.002)
						20	0.0735 (0.002)
						50	0.0738 (0.002)
				0.7	*0.0693* (0.001)		
				0.9	0.0694 (0.001)	5	0.0785 (0.002)
						20	0.0749 (0.002)
						50	0.0749 (0.002)
20	0.0629 (0.001)	0.0754 (0.001)	0.0716 (0.001)	0.1	0.0628 (0.001)	5	0.0711 (0.001)
						20	0.0679 (0.001)
						50	0.0690 (0.001)
				0.3	0.0630 (0.001)		
				0.5	0.0629 (0.001)	5	0.0711 (0.001)
						20	0.0683 (0.001)
						50	0.0694 (0.001)
				0.7	*0.0627* (0.001)		
				0.9	0.0629 (0.001)	5	0.0720 (0.001)
						20	0.0695 (0.001)
						50	0.0701 (0.001)
50	0.0598 (0.000)	0.0715 (0.001)	0.0673 (0.001)	0.1	*0.0596* (0.000)	5	0.0661 (0.001)
						20	0.0646 (0.001)
						50	0.0679 (0.001)
				0.3	0.0598 (0.000)		
				0.5	0.0598 (0.000)	5	0.0661 (0.001)
						20	0.0648 (0.001)
						50	0.0683 (0.001)
				0.7	*0.0596* (0.000)		
				0.9	0.0597 (0.000)	5	0.0668 (0.001)
						20	0.0657 (0.001)
						50	0.0685 (0.001)
100	0.0587 (0.000)	0.0694 (0.001)	0.0653 (0.001)	0.1	*0.0584* (0.000)	5	0.0637 (0.001)
						20	0.0627 (0.001)
						50	0.0654 (0.001)
				0.3	0.0587 (0.000)		
				0.5	0.0587 (0.000)	5	0.0637 (0.001)
						20	0.0628 (0.001)
						50	0.0658 (0.001)
				0.7	0.0585 (0.000)		
				0.9	0.0586 (0.000)	5	0.0642 (0.001)
						20	0.0634 (0.001)
						50	0.0660 (0.001)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.0582 (0.000)	0.0680 (0.001)	0.0639 (0.001)	0.1	*0.0579* (0.000)	5	0.0622 (0.001)
						20	0.0609 (0.001)
						50	0.0642 (0.001)
				0.3	0.0582 (0.000)		
				0.5	0.0582 (0.000)	5	0.0622 (0.001)
						20	0.0610 (0.001)
						50	0.0644 (0.001)
				0.7	*0.0579* (0.000)		
				0.9	0.0580 (0.000)	5	0.0626 (0.001)
						20	0.0615 (0.001)
						50	0.0646 (0.001)
300	**0.0580** (0.000)	**0.0676** (0.001)	**0.0634** (0.001)	0.1	***0.0576*** (0.000)	5	0.0618 (0.001)
						20	**0.0602** (0.001)
						50	0.0640 (0.001)
				0.3	0.0580 (0.000)		
				0.5	0.0580 (0.000)	5	0.0617 (0.001)
						20	0.0603 (0.001)
						50	0.0644 (0.001)
				0.7	0.0577 (0.000)		
				0.9	0.0577 (0.000)	5	0.0622 (0.001)
						20	0.0608 (0.001)
						50	0.0645 (0.001)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0564 (Standardabweichung 0.116)

Abbildung 9.32 stellt wieder den zeitlichen Verlauf der geschätzten bzw. prognostizierten Erwartungswertvektoren durch *ILDA* und die Erweiterungen mit verschiedenen Werten für N_{trend} grafisch dar. Aufgrund des linearen Trends der Erwartungswertvektoren werden die Richtungen in allen Fällen im Durchschnitt (über alle Simulationsdurchläufe) perfekt geschätzt. An der oberen linken Grafik ist jedoch die zeitliche Verzögerung der Schätzer bei der ursprünglichen Methode zu erkennen. Zum Ende des Datenstroms werden die Erwartungswertvektoren durch $(0, 0)$ geschätzt und nicht durch $(10, 10)$ bzw. $(-10, 10)$ (vgl. Seite 220 f.). Durch die zusätzliche Modellierung des Trends durch lokale lineare Regressionsmodelle und Prognose der Erwartungswertvektoren durch diese Regressionsmodelle kann diese zeitliche Verzögerung korrigiert werden. Für kleine Fenster N_{trend} ist die Schätzung aufgrund der geringen Anzahl an Beobachtungen für jedes Regressionsmodell noch etwas unsicher. Ab $N_{\text{trend}} = 20$ ist hier bereits im Mittel über 100 Wiederholungen ein fast linearer Verlauf zu erkennen, das heißt die Erwartungswertvektoren für die LDA werden im Laufe des Datenstroms sehr gut geschätzt und prognostiziert.

Tabelle 9.8 umfasst die durchschnittlichen euklidischen Abstände zwischen den wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit für alle betrachteten Methoden und ihre Erweiterungen. Beim Vergleich aller ursprünglichen Methoden ist die Methode *LDA-AF* hervorzuheben (vgl. erste „Zeile“). Die Werte 0.1910 (Klasse 1) und 0.1887 (Klasse 2) sind deutlich kleiner als jene der anderen Methoden. Dies bedeutet, dass die wahren Erwartungswertvektoren zu allen Zeitpunkten durch die Methode gut geschätzt werden. Der Vorteil der Erweiterung der Methoden durch Einführung lokaler linearer Regressionsmodelle zur Modellierung des Trends zeigt sich deutlich in den euklidischen Abständen. Für alle Methoden kann der durchschnittliche euklidische Abstand für beide Klassen verringert werden und wird

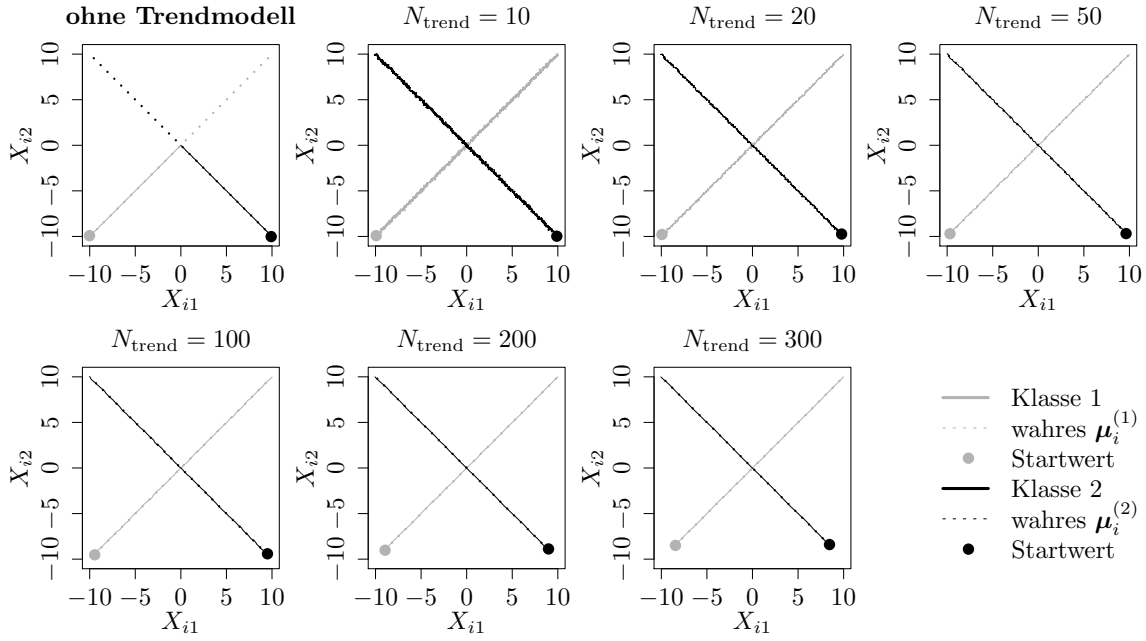


Abbildung 9.32: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren auf der Datensituation **„Kreuzen“** für den gesamten Datenstrom. Ergebnisse aus *ILDA* und Erweiterung mit verschiedenen N_{trend}.

darüber hinaus aufgrund des linearen Trends mit wachsendem N_{trend} immer kleiner. Lediglich *OLDC* mit adaptiver Lernrate bildet hier eine Ausnahme. Bis $N_{\text{trend}} = 200$ sinkt der durchschnittliche euklidische Abstand mit steigendem N_{trend}. Für $N_{\text{trend}} > 200$ steigt er bei Betrachtung eines großen Fensters $L = 50$ zur Adaption der Lernrate wieder.

Generell variieren die Ergebnisse der einzelnen Methoden nach Einführung der Trendmodelle weniger stark. Mit steigendem N_{trend} werden die Unterschiede zudem geringer. Für die Methoden mit ursprünglich großen durchschnittlichen euklidischen Abständen über die Zeit erfolgt demnach absolut gesehen eine stärkere Verringerung der Abstände.

In Abbildung 9.33 auf Seite 285 am Ende des Abschnittes ist die Veränderung der Klassifikationsregel und des Prognosefehlers für verschiedene Zeitpunkte bei *ILDA* und zusätzlicher Einbindung lokaler linearer Regressionsmodelle auf den jeweils letzten $n_{\text{trend}}^{(c)}$ Mittelwertvektoren aus Fenstern der Breite $N_{\text{trend}} = 50$ veranschaulicht.

Fazit: In dieser Datensituation, in der die Voraussetzung eines linearen Trends der Erwartungswertvektoren erfüllt ist, kann die Prognosegüte durch die Erweiterung stark verbessert werden. Sowohl die Prognosefehler als auch die euklidischen Abstände zwischen wahren und geschätzten bzw. prognostizierten Erwartungswertvektoren werden deutlich verringert. Aufgrund der erfüllten Voraussetzung bezüglich des Trends ist eine möglichst große Fenstergröße N_{trend} für die Anpassung der lokalen linearen Regressionsmodelle zur Modellierung des Trends der Erwartungswerte am besten, da mehr Beobachtungen in die Modellierung einfließen können und die Schätzung stabiler wird.

Tabelle 9.8: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Kreuzen“ ($p = 2$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	7.0965 (0.013)	0.5012 (1.823)	*0.1910* (0.129)	0.1	12.7386 (0.105)	5	2.1932 (0.331)
	7.0908 (0.013)	0.4609 (1.718)	*0.1887* (0.135)		12.7313 (0.165)		2.1870 (0.365)
						20	2.5889 (0.421)
							2.5849 (0.458)
						50	4.6904 (0.409)
							4.7081 (0.435)
				0.3	9.9271 (0.015)		
					9.9188 (0.019)		
				0.5	7.0965 (0.013)	5	1.1922 (0.122)
					7.0908 (0.013)		1.1822 (0.124)
						20	1.5192 (0.254)
							1.5072 (0.262)
						50	3.7118 (0.194)
							3.7525 (0.221)
				0.7	4.2617 (0.016)		
					4.2625 (0.015)		
				0.9	1.4307 (0.022)	5	0.2627 (0.089)
					1.4321 (0.022)		0.2596 (0.088)
						20	0.4937 (0.150)
							0.4893 (0.159)
						50	2.8139 (0.140)
							2.8456 (0.146)
10	0.1010 (0.345)	0.1328 (0.630)	0.1336 (0.644)	0.1	*0.1005* (0.341)	5	0.1273 (0.584)
	0.1021 (0.345)	0.1320 (0.618)	0.1332 (0.634)		*0.1016* (0.342)		0.1271 (0.597)
						20	0.1101 (0.423)
							0.1100 (0.429)
						50	0.1082 (0.418)
							0.1080 (0.416)
				0.3	0.1007 (0.342)		
					0.1018 (0.343)		
				0.5	0.1010 (0.345)	5	0.1278 (0.590)
					0.1021 (0.345)		0.1274 (0.607)
						20	0.1108 (0.431)
							0.1119 (0.445)
						50	0.1093 (0.431)
							0.1096 (0.428)
				0.7	0.1017 (0.351)		
					0.1029 (0.351)		
				0.9	0.1049 (0.382)	5	0.1316 (0.633)
					0.1060 (0.378)		0.1315 (0.652)
						20	0.1165 (0.484)
							0.1168 (0.494)
						50	0.1132 (0.475)
							0.1137 (0.470)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
20	0.0646 (0.126)	0.0932 (0.272)	0.0937 (0.274)	0.1	*0.0640* (0.124)	5	0.0882 (0.241)
	0.0666 (0.126)	0.0944 (0.270)	0.0947 (0.272)		*0.0659* (0.124)		0.0894 (0.244)
						20	0.0778 (0.212)
							0.0789 (0.211)
						50	0.0742 (0.202)
							0.0759 (0.204)
				0.3	0.0642 (0.124)		
					0.0661 (0.124)		
				0.5	0.0646 (0.126)	5	0.0889 (0.244)
					0.0666 (0.126)		0.0895 (0.246)
						20	0.0785 (0.212)
							0.0801 (0.218)
						50	0.0752 (0.204)
							0.0777 (0.213)
				0.7	0.0655 (0.129)		
					0.0675 (0.129)		
				0.9	0.0687 (0.144)	5	0.0920 (0.263)
					0.0710 (0.144)		0.0937 (0.265)
						20	0.0827 (0.233)
							0.0853 (0.231)
						50	0.0790 (0.220)
							0.0803 (0.221)
50	0.0398 (0.048)	0.0638 (0.128)	0.0637 (0.126)	0.1	*0.0390* (0.047)	5	0.0579 (0.105)
	0.0439 (0.049)	0.0672 (0.126)	0.0653 (0.125)		*0.0429* (0.048)		0.0615 (0.104)
						20	0.0571 (0.092)
							0.0585 (0.092)
						50	0.0601 (0.314)
							0.0647 (0.317)
				0.3	0.0393 (0.047)		
					0.0432 (0.048)		
				0.5	0.0398 (0.048)	5	0.0588 (0.105)
					0.0439 (0.049)		0.0621 (0.105)
						20	0.0562 (0.094)
							0.0577 (0.094)
						50	0.0608 (0.293)
							0.0644 (0.324)
				0.7	0.0410 (0.051)		
					0.0451 (0.052)		
				0.9	0.0451 (0.061)	5	0.0625 (0.117)
					0.0490 (0.062)		0.0663 (0.117)
						20	0.0612 (0.106)
							0.0643 (0.105)
						50	0.0635 (0.282)
							0.0667 (0.282)
100	0.0284 (0.026)	0.0467 (0.068)	0.0470 (0.067)	0.1	*0.0277* (0.026)	5	0.0421 (0.056)
	0.0339 (0.026)	0.0534 (0.067)	0.0501 (0.066)		*0.0324* (0.027)		0.0470 (0.057)
						20	0.0436 (0.056)
							0.0476 (0.056)
						50	0.0465 (0.176)
							0.0529 (0.199)
				0.3	0.0279 (0.025)		
					0.0330 (0.026)		
				0.5	0.0284 (0.026)	5	0.0434 (0.056)
					0.0339 (0.026)		0.0484 (0.055)
						20	0.0437 (0.056)
							0.0478 (0.055)
						50	0.0499 (0.185)
							0.0527 (0.209)
				0.7	0.0296 (0.028)		
					0.0355 (0.028)		
				0.9	0.0336 (0.036)	5	0.0463 (0.063)
					0.0395 (0.036)		0.0527 (0.063)
						20	0.0478 (0.064)
							0.0529 (0.062)
						50	0.0502 (0.177)
							0.0561 (0.180)

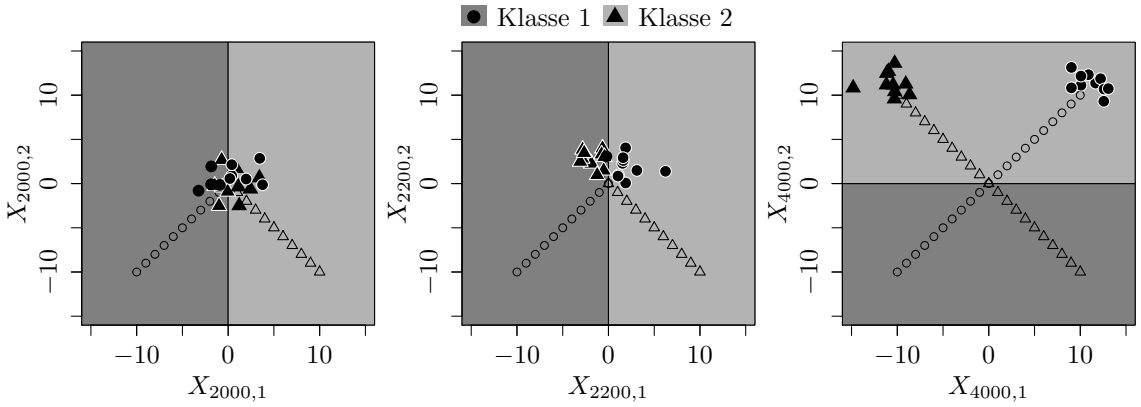
Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.0194 (0.015)	0.0329 (0.034)	0.0328 (0.034)	0.1	0.0193 (0.017)	5	0.0296 (0.032)
	0.0243 (0.015)	0.0381 (0.034)	0.0367 (0.034)		*0.0228* (0.018)		0.0356 (0.033)
						20	0.0357 (0.035)
							0.0367 (0.037)
						50	0.0414 (0.187)
							0.0446 (0.266)
				0.3	*0.0189* (0.015)		
					0.0234 (0.015)		
				0.5	0.0194 (0.015)	5	0.0301 (0.029)
					0.0243 (0.015)		0.0364 (0.029)
						20	0.0346 (0.032)
							0.0376 (0.033)
						50	0.0460 (0.195)
							0.0491 (0.210)
				0.7	0.0208 (0.016)		
					0.0261 (0.017)		
				0.9	0.0250 (0.022)	5	0.0331 (0.033)
					0.0301 (0.022)		0.0402 (0.033)
						20	0.0383 (0.036)
							0.0418 (0.037)
						50	0.0436 (0.211)
							0.0493 (0.239)
300	**0.0153** (0.011)	**0.0259** (0.022)	**0.0243** (0.022)	0.1	0.0165 (0.015)	5	0.0243 (0.025)
	0.0203 (0.011)	**0.0312** (0.023)	**0.0303** (0.023)		***0.0191*** (0.017)		**0.0288** (0.027)
						20	0.0326 (0.029)
							0.0314 (0.030)
						50	0.0454 (0.248)
							0.0469 (0.457)
				0.3	***0.0151*** (0.012)		
					0.0195 (0.012)		
				0.5	0.0153 (0.011)	5	**0.0237** (0.020)
					0.0203 (0.011)		0.0289 (0.021)
						20	0.0299 (0.024)
							0.0322 (0.025)
						50	0.0516 (0.293)
							0.0503 (0.273)
				0.7	0.0166 (0.012)		
					0.0222 (0.013)		
				0.9	0.0200 (0.016)	5	0.0259 (0.023)
					0.0268 (0.017)		0.0329 (0.023)
						20	0.0324 (0.026)
							0.0356 (0.026)
						50	0.0457 (0.307)
							0.0502 (0.369)

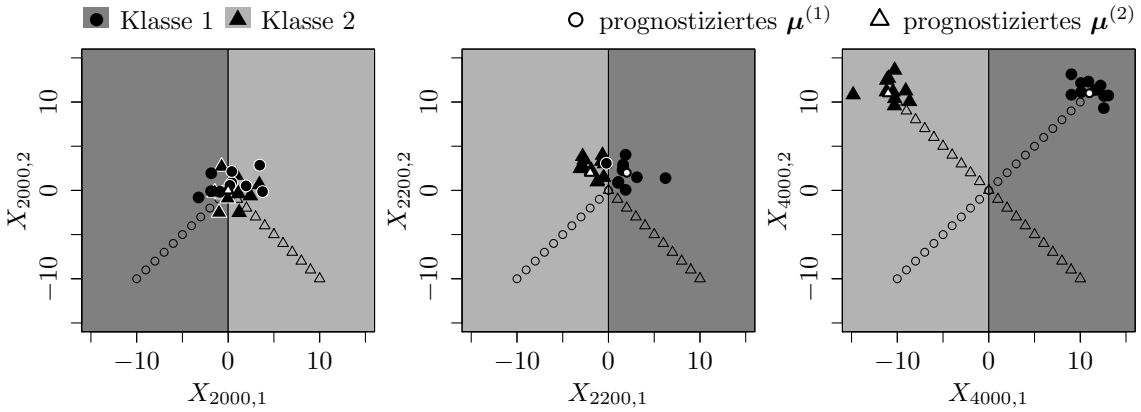
Ein weiterer Vorteil der Erweiterung besteht darin, dass sich die erweiterten Methoden in Bezug auf die Prognosegüte repräsentiert durch die euklidischen Abstände (vgl. Tabelle 9.8) bei genügend großem N_{trend} kaum noch unterscheiden. Dies ist ebenfalls bei Fokus auf den durchschnittlichen mittleren Prognosefehler der Fall (vgl. Tabelle 9.7).

Zudem ist die Wahl der Lernrate λ bei *OLDC* nach Erweiterung der Methode von geringer Bedeutung, während sich in der ursprünglichen Methode die Prognosefehler in Abhängigkeit von λ stark unterscheiden (vgl. Tabelle 9.7). Bei *OLDC* mit adaptiver Lernrate sollte das Fenster L zur Adaption nicht allzu groß gewählt werden. Für $L < 50$ schwanken die Prognosefehler unabhängig vom Startwert λ_{start} bei großem, festem N_{trend} nur leicht.

Die konkrete Auswahl der speziellen Methode für Online Diskriminanzanalyse ist insgesamt nach Erweiterung somit weniger relevant.



(a) Ursprüngliche Methode *ILDA* bzw. *OLDC* mit $\lambda = 0.5$.



(b) Erweiterung durch lokale lineare Regressionsmodelle mit $N_{\text{trend}} = 50$.

Abbildung 9.33: Veränderung der Klassifikationsgrenze der LDA im Laufe der Zeit durch Aktualisierung des Modells mit neuen Beobachtungen aus Verteilungen mit verschobenem Erwartungswertvektor und Auswirkung auf den Prognosefehler für die Datensituation **„Kreuzen“**.
Schwarze ∘ und △ veranschaulichen Erwartungswertvektoren der Verteilungen von Klasse 1 bzw. Klasse 2, aus denen Beobachtungen im Datenstrom realisiert werden, die in (a) für die Anpassung der LDA herangezogen werden.

Datensituation „Vorbeilaufen“ ($p = 2$) In dieser Datensituation laufen die Verteilungen beider Klassen im Laufe der Zeit auf zwei parallelen Geraden aufeinander zu und aneinander vorbei (vgl. Seite 221 f.). Dieses Verhalten sorgt dafür, dass der Bayesfehler zu den meisten Zeitpunkten fast Null ist, da die Verteilungen beider Klassen aufgrund des Abstandes zueinander perfekt trennbar sind. Nähern sich die Verteilungen im Laufe des Datenstroms an, überlagern sie sich irgendwann teilweise, was einen leichten Anstieg des Bayesfehlers mit einem Maximum von etwa 0.0668 um den Zeitpunkt $t = 2000$ nach sich zieht, bevor ein erneuter Abfall erfolgt (vgl. Abbildungen 9.35–9.39).

Dieser Verlauf wird durch den Prognosefehler bei *QDA-AF*, *LDA-AF* und *OLDC* mit großer Lernrate $\lambda = 0.9$ bzw. *OLDC adaptive* angenähert (vgl. Abbildungen 9.35 (b) und (c), 9.36 (c) sowie 9.37–9.39). Für diese Methoden ist der durchschnittliche mittlere Prognosefehler über die Zeit auch entsprechend gering (vgl. Tabelle 9.9).

Die Update-Methoden für LDA ohne bzw. mit identischer Gewichtung der Beobachtungen (*ILDA* und *OLDC* mit $\lambda = 0.5$) verhalten sich im Laufe der Zeit folgendermaßen auf dieser Datensituation: Bis zum Zeitpunkt $t = 2000$, d. h. dem Zeitpunkt, an dem sich die Verteilungen der Klassen kreuzen, wird der Bayesfehler zu jedem Zeitpunkt durch den Prognosefehler nahezu approximiert. Dies ergibt sich zufällig durch die konkrete Datensituation. Intuitiv wäre zunächst damit zu rechnen, dass der Prognosefehler etwas höher läge als der Bayesfehler, da die Mittelwertvektoren aufgrund des linearen Trends die Erwartungswertvektoren des Zeitpunktes $\frac{t+1}{2}$ anstelle jene des Zeitpunktes $t + 1$ repräsentieren und die lineare Trenngerade somit nicht die Verteilungen des kommenden Zeitpunktes $t + 1$ perfekt trennt. Ausschlaggebend ist hier jedoch die gepoolte Kovarianzmatrix. Aufgrund des linearen Trends der Erwartungswertvektoren und der stetigen gleich gewichteten Aktualisierungen der Größen der LDA durch die neuen Beobachtungen „streckt“ sich die gepoolte Kovarianzmatrix im Laufe der Zeit in Richtung des Trends. Dies zieht eine leichte Rotation der linearen Trennebene bei der Klassifizierung durch die LDA zum Zeitpunkt t nach sich, sodass die Verteilungen der beiden Klassen des Zeitpunktes $t + 1$ jeweils fast perfekt voneinander getrennt werden. Ab dem Zeitpunkt $t = 2000$ steigt der Prognosefehler jedoch zunächst weiter an, während der Bayesfehler wieder sinkt. Dies liegt daran, dass zu den Zeitpunkten $t > 2000$ die aktuellen konstanten Verteilungen beider Klassen sich zwar nicht mehr überschneiden (was zu einem Bayesfehler nahe Null führt), die aktualisierten Größen der LDA und folglich die Klassifikationsgrenze jedoch weiterhin auf allen bisherigen und somit auch den vergangenen Verteilungen basieren. Dadurch rotiert die Klassifikationsgrenze und der Prognosefehler steigt (vgl. Veranschaulichung in Abbildung 9.41 auf Seite 301).

Durch stärkere Gewichtung der aktuellen Beobachtungen durch $\lambda = 0.9$ bei *OLDC* bzw. adaptive Lernrate oder stärkere Gewichtung aktueller Likelihood Terme bei *QDA-AF* und *LDA-AF* kann dieser Effekt bereits reduziert werden (vgl. Abbildungen 9.35 (b) und (c), 9.36 (c) sowie 9.37–9.39). Durch die Integration lokaler linearer Regressionsmodelle soll der Prognosefehler noch weiter reduziert werden. Für *OLDC* mit $\lambda = 0.9$ ist anhand der

Kurven zu sehen, dass dies zu vielen Zeitpunkten für alle N_{trend} möglich ist. Für die Methoden *ILDA* und *OLDC* mit kleineren Lernraten (vgl. Abbildungen 9.35 (a) und 9.36 (a) und (b)) kann der Anstieg des Prognosefehlers ab Zeitpunkt $t = 2000$ ebenfalls verringert werden. Jedoch sollte gleichzeitig auch erwähnt werden, dass der geringe Prognosefehler zu früheren Zeitpunkten $t \leq 2000$ etwas vergrößert wird. Dies liegt auch hier an der oben erklärten Streckung der gepoolten Kovarianzmatrix im Laufe der Zeit. Dadurch rotiert die Trenngerade der LDA leicht und die Verteilungen der beiden Klassen werden trotz besser geschätzter Erwartungswertvektoren des Zeitpunktes $t + 1$ als bei der ursprünglichen Methode nicht perfekt getrennt. Dies liegt wieder zufällig an der speziellen Datensituation. Die zuerst genannte Verbesserung überwiegt jedoch diese Verschlechterung, sodass im Durchschnitt über die Zeit der Prognosefehler durch Integration der Regressionsmodelle verringert wird (vgl. Spalten „ILDA“ und „OLDC fix“ in Tabelle 9.9). Mit steigender Fensterbreite N_{trend} sinkt der durchschnittliche mittlere Prognosefehler über die Zeit, da auch hier die Annahme eines linearen Trends der Erwartungswertvektoren erfüllt ist.

Für *QDA-AF* und *LDA-AF* wird die (grüne) Kurve von den Kurven der Erweiterungen überlagert (vgl. Abbildung 9.35 (b) und (c)). Anhand der Mittelwerte in Tabelle 9.9 wird ersichtlich, dass bei diesen Methoden zunächst eine leichte Verschlechterung bezüglich der Prognosegüte für $N_{\text{trend}} \in \{10, 20\}$ (*QDA-AF*) bzw. $N_{\text{trend}} \in \{10, 20, 50\}$ (*LDA-AF*) erfolgt, für größere Fenster für die Regressionsmodelle der Prognosefehler der Diskriminanzanalyse jedoch dann auch sinkt.

Weitere Auffälligkeiten sind, dass eine hohe Lernrate von $\lambda = 0.9$ bei *OLDC* hier immer am besten ist, auch bei der erweiterten Methode (vgl. Spalte „OLDC fix“). Zudem ist eine kleinere Fehlerrate $\lambda = 0.1$ immer noch besser als mittlere Werte für λ. Es ist also ein „kurvenförmiger“ Verlauf des Prognosefehlers für die verschiedenen Werte von λ erkennbar. Dies ist darauf zurückzuführen, dass bei geringen Lernraten die oben beschriebene Verschlechterung des Verlaufs des Prognosefehlers für Zeitpunkte $t \leq 2000$ nicht so stark ausgeprägt ist, was sich auf den Mittelwert über die Zeit auswirkt.

Bei *OLDC adaptive* wird mit 0.0082 der geringste durchschnittliche mittlere Prognosefehler über die Zeit bei einer Wahl von $\lambda_{\text{start}} = 0.9$, $L = 20$ und $N_{\text{trend}} = 300$ erzielt. Anhand der Abbildungen 9.37–9.39 ist zu sehen, dass auch bei der erweiterten Methode ein Startwert von $\lambda_{\text{start}} = 0.9$ am besten ist. Der Prognosefehler approximiert so den Bayesfehler zu jedem Zeitpunkt. Wird ein kleinerer Startwert gewählt, so ergibt sich auch hier das Problem des erhöhten Prognosefehlers zu Zeitpunkten $t \leq 2000$. Bezüglich des Parameters L bzw. der Kombination mit λ_{start} ist kein durchgehender auffälliger Effekt für alle N_{trend} zu erkennen. In Abbildung 9.34 ist zwar zu sehen, dass die adaptive Lernrate für alle Startwerte λ_{start} und Fenstergrößen L kurz vor dem Zeitpunkt $t = 2000$ ansteigt und abhängig von L danach kein ($L = 5$), ein geringer ($L = 20$) oder stärkerer erneuter Abfall ($L = 50$) erfolgt. Dies hat jedoch keinen starken Einfluss auf den durchschnittlichen mittleren Prognosefehler über die Zeit, da λ_{start} alleine aufgrund des Einflusses für Zeitpunkte $t \leq 2000$ eine wichtigere Rolle spielt. Für alle N_{trend} ist somit ein hoher Startwert der

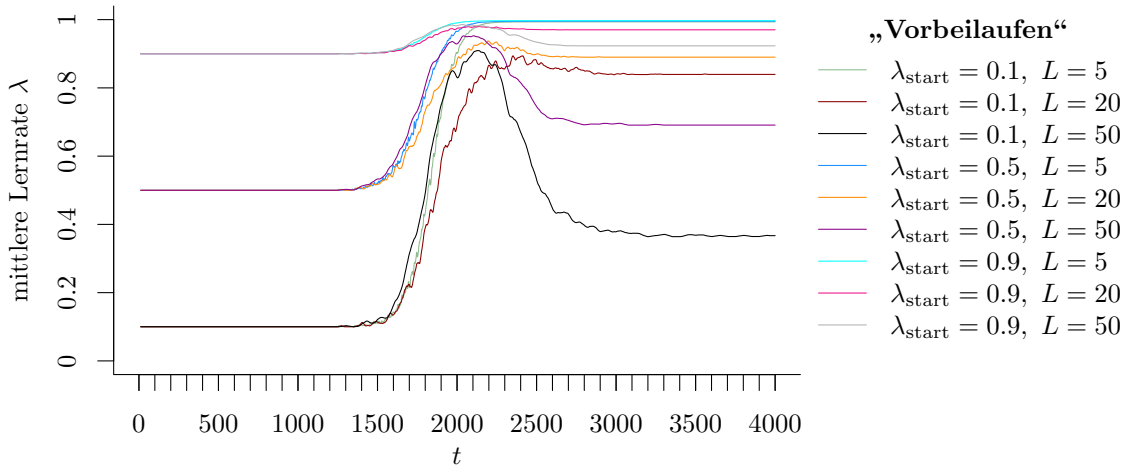
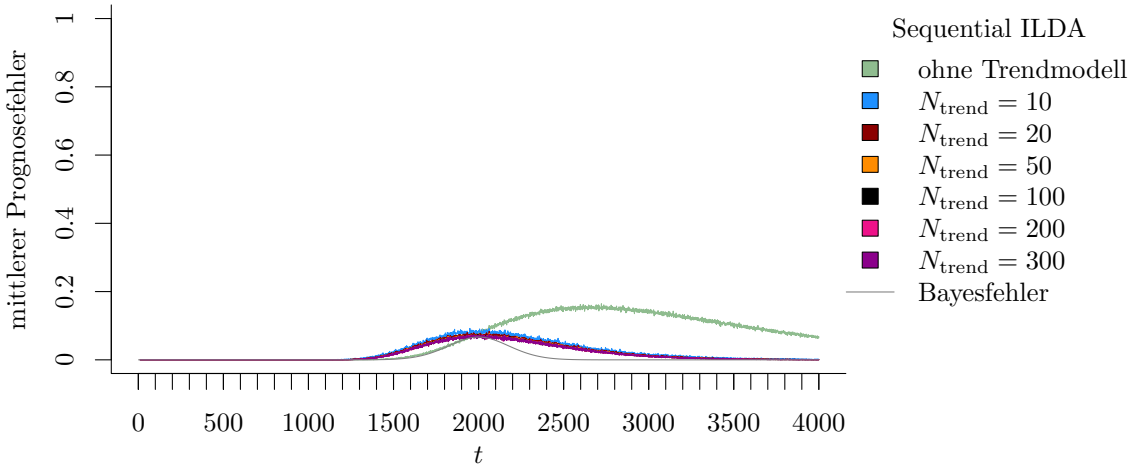


Abbildung 9.34: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **„Vorbeilaufen“**.

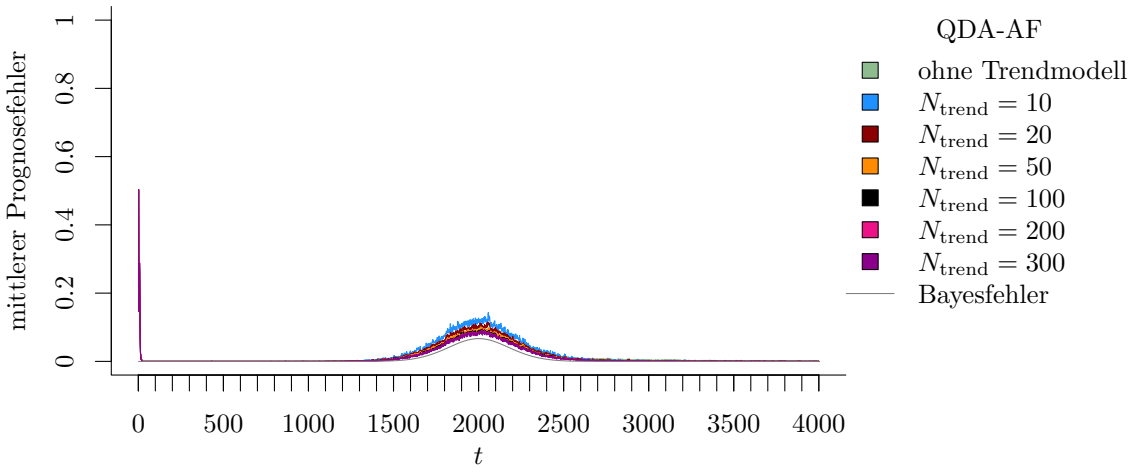
Lernrate am besten. Bei diesem hohen Startwert erweist sich die Kombination mit $L = 20$ als optimal, während bei den kleineren Startwerten die Wahl $L = 20$ häufig die höchsten Prognosefehler nach sich zieht. Generell kann der durchschnittliche Prognosefehler über die Zeit mit steigendem N_{trend} für alle Kombinationen minimal verringert werden. Die bereits geringen Prognosefehler der ursprünglichen Methode können durch Integration zusätzlicher Regressionsmodelle auf Basis von $n_{\text{trend}}^{(c)}$ Beobachtungen zur Modellierung des Trends und Prognose der Erwartungswertvektoren für vereinzelte Kombinationen aus λ_{start} und L zudem noch leicht unterschritten werden (vgl. Tabelle 9.9).

Zusammengefasst lässt sich erkennen:

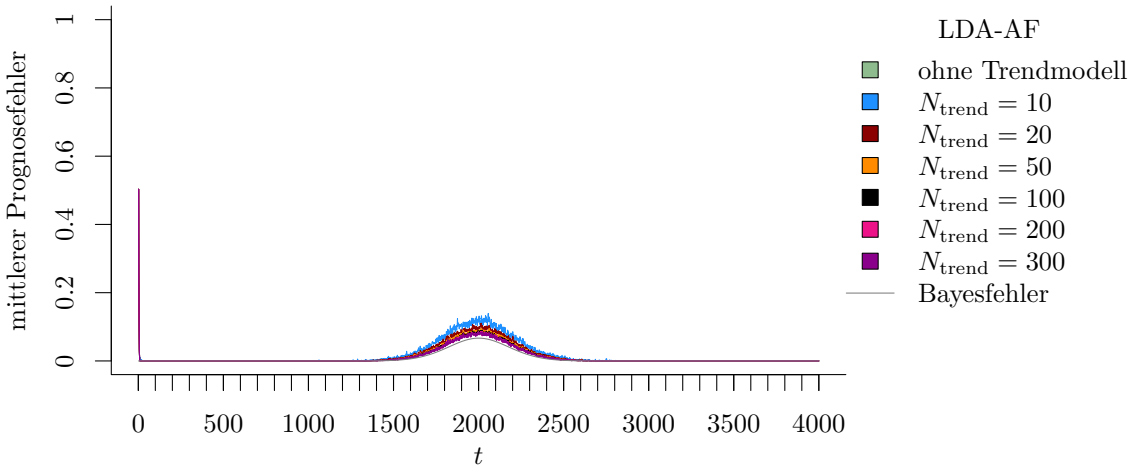
- Der durchschnittliche Bayesfehler über den gesamten Datenstrom für diese Datensituation beträgt 0.0076 mit einer Standardabweichung von 0.017.
- Für *ILDA* und *OLDC fix* wird direkt durch Integration von Regressionsmodellen eine Reduzierung des Prognosefehlers erzielt. Zudem sinkt dieser für steigendes N_{trend} immer weiter aufgrund der erfüllten Voraussetzung eines linearen Trends.
- *QDA-AF* und *LDA-AF*: Der durchschnittliche mittlere Prognosefehler über die Zeit steigt zunächst leicht bei zusätzlicher Betrachtung von Trendmodellen im Vergleich zu den ursprünglichen Methoden. Durch lokale Regressionsmodelle auf breiteren Fenstern N_{trend} sinkt jedoch der durchschnittliche mittlere Prognosefehler über die Zeit.
- Bei *OLDC* mit fester Lernrate ist in allen Fällen eine hohe Lernrate λ am besten.
- *OLDC adaptive*: Für steigendes N_{trend} kann der durchschnittliche Prognosefehler über die Zeit minimal verringert werden. Die niedrigen Werte der ursprünglichen Methode werden für vereinzelte Kombinationen mit $\lambda_{\text{start}} = 0.9$ leicht unterschritten.
- Der minimale durchschnittliche mittlere Prognosefehler über die Zeit von 0.0082 wird durch *OLDC* mit $\lambda_{\text{start}} = 0.9$, $L = 20$ und der Integration lokaler linearer Regressionsmodelle auf Fenstern von $N_{\text{trend}} = 300$ erzielt, direkt gefolgt von 0.0083 durch *OLDC* mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ (sowie zusätzlich bei $N_{\text{trend}} = 200$).



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

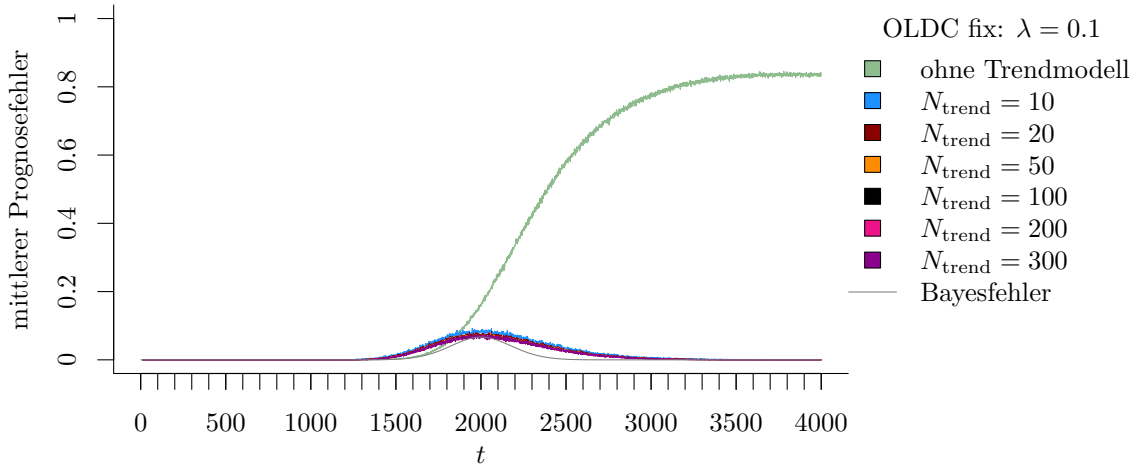


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

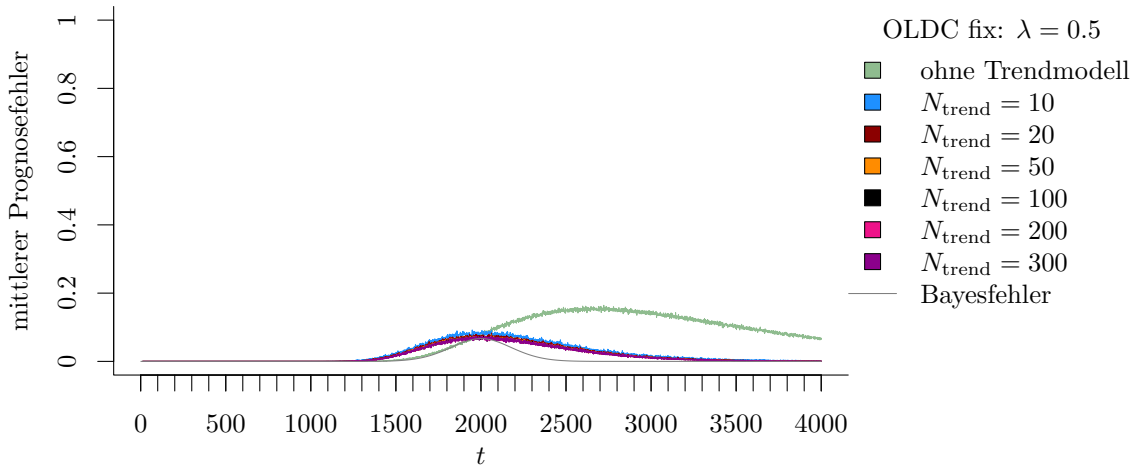


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

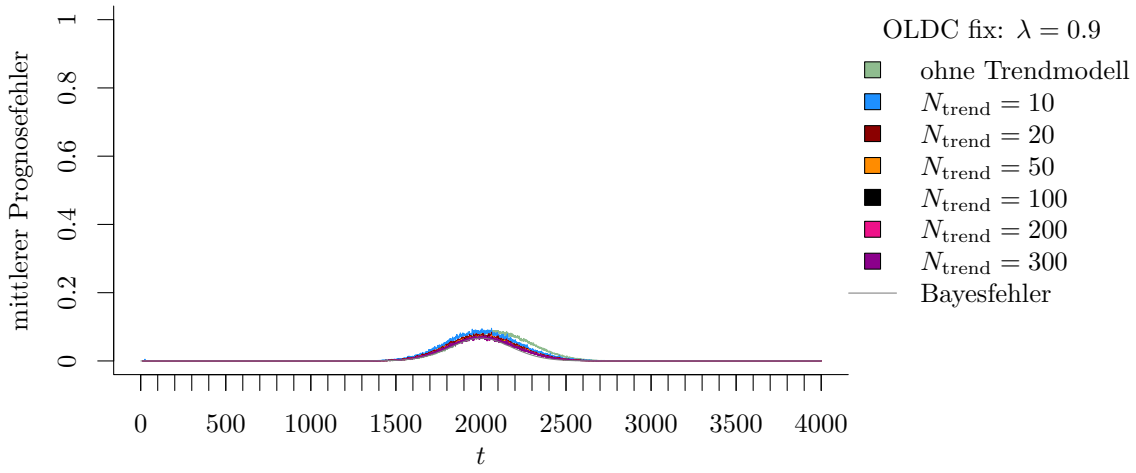
Abbildung 9.35: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“** im zweidimensionalen Raum.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.

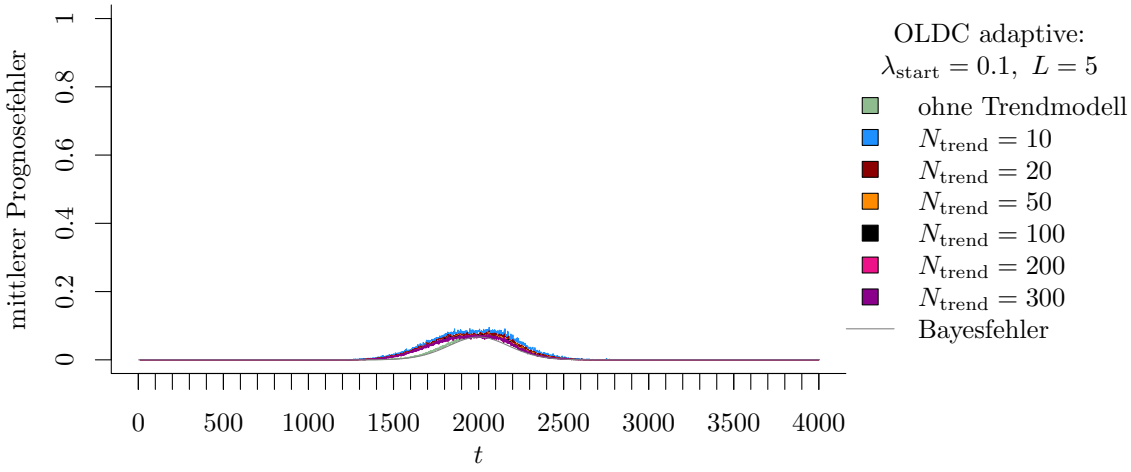


(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.

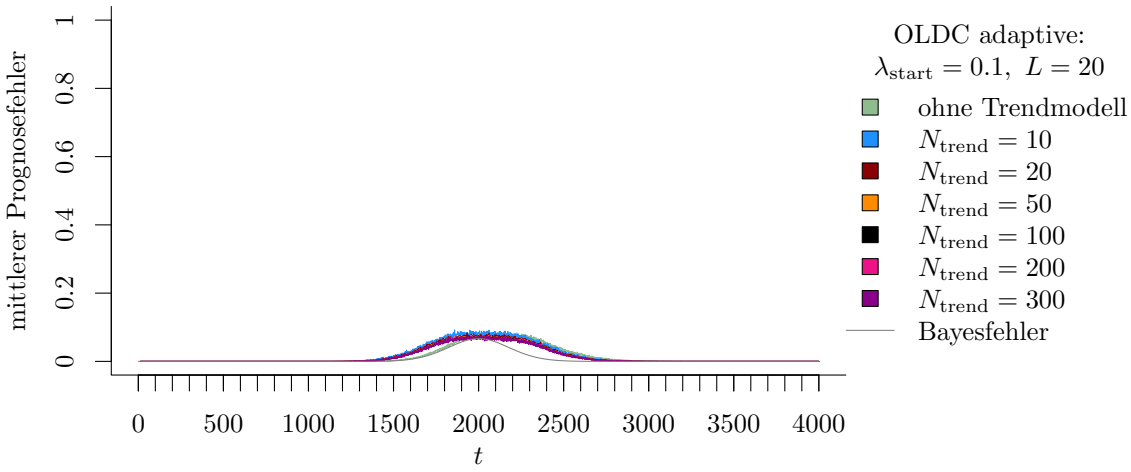


(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

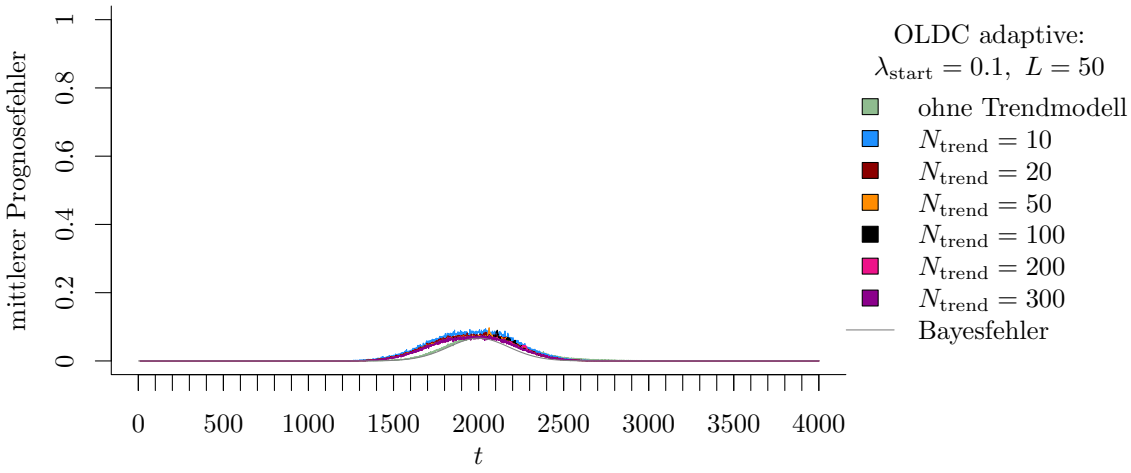
Abbildung 9.36: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

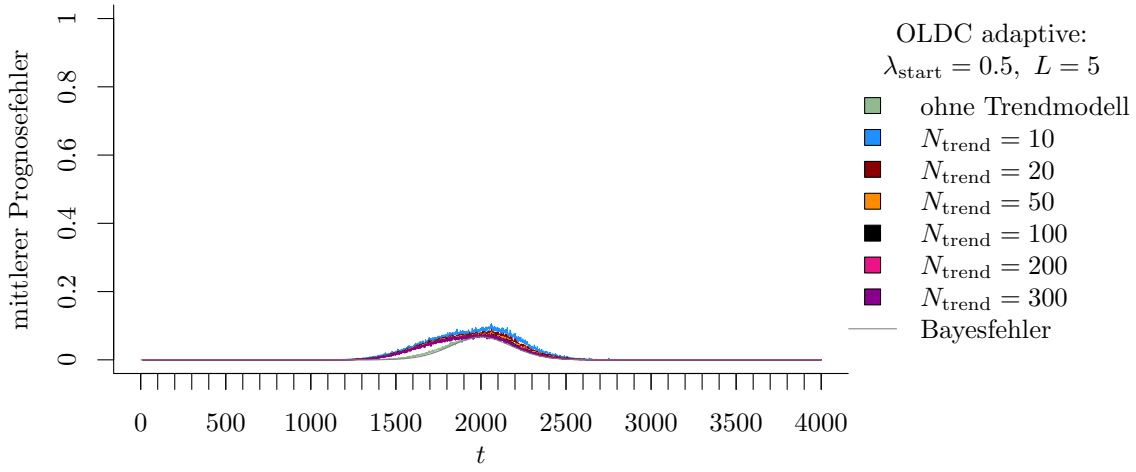


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

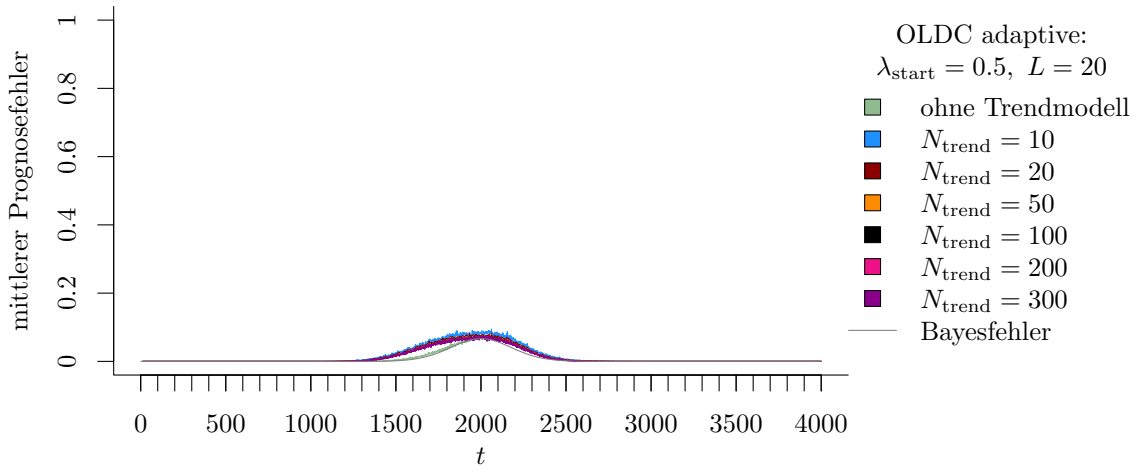


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

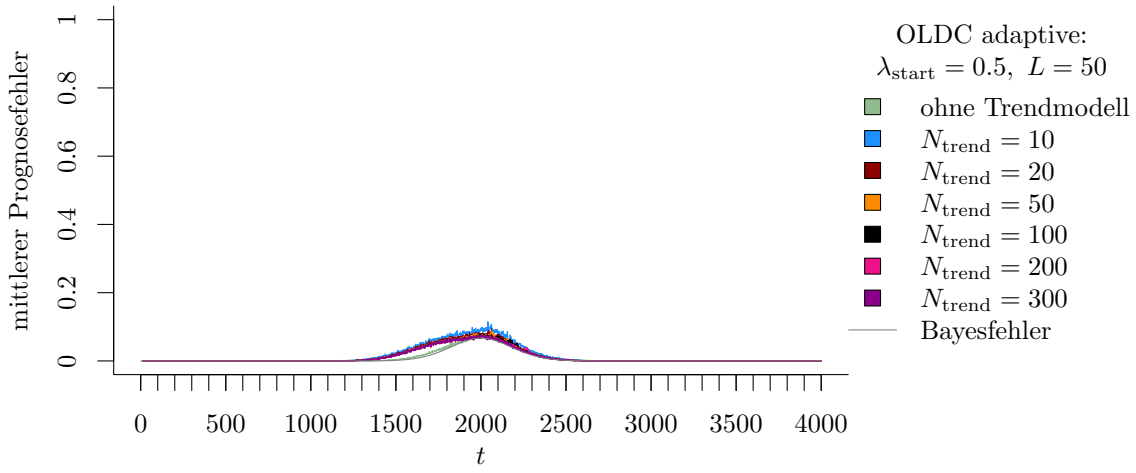
Abbildung 9.37: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

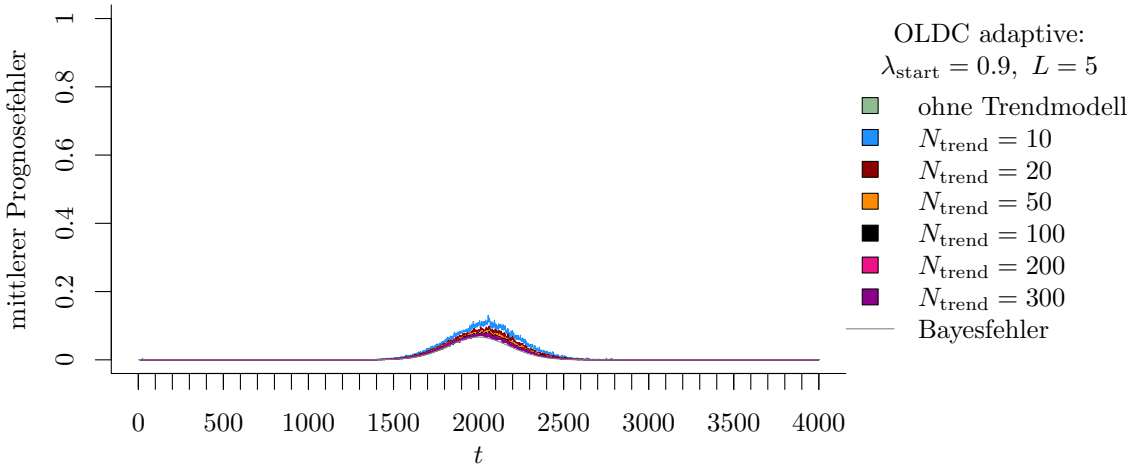


(b) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

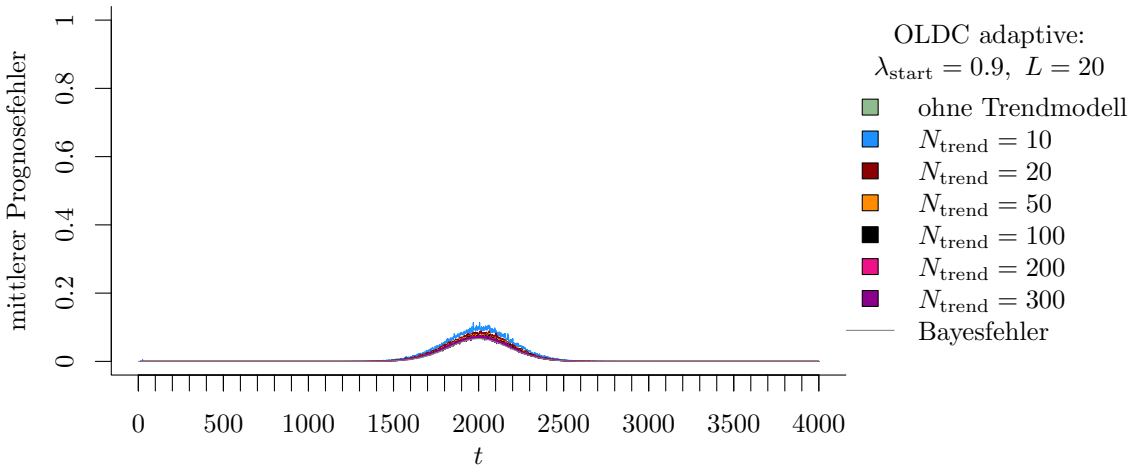


(c) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

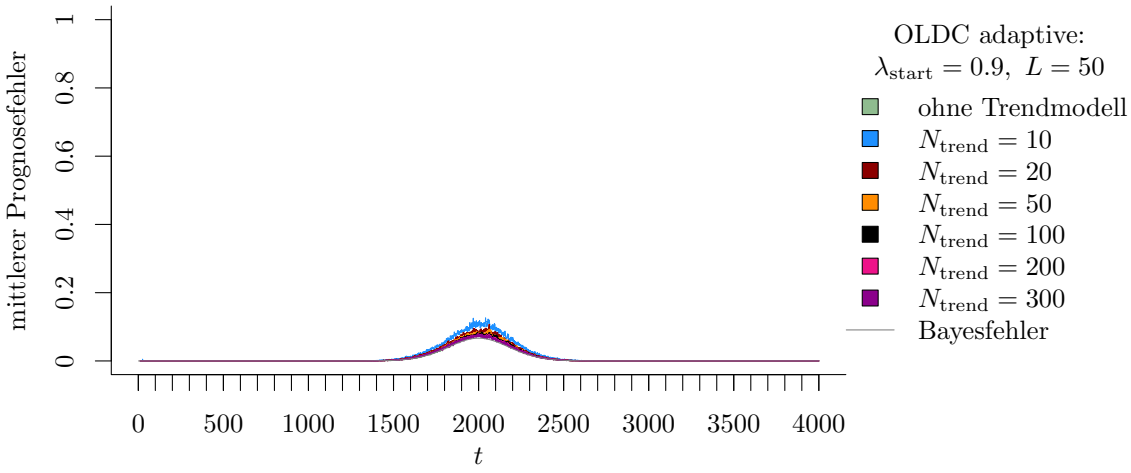
Abbildung 9.38: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.39: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“** im zweidimensionalen Raum.

Tabelle 9.9: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ ($p = 2$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.0640 (0.001)	0.0138 (0.000)	0.0106 (0.000)	0.1	0.3467 (0.008)	5	0.0091 (0.000)
						20	0.0142 (0.000)
						50	0.0103 (0.000)
				0.3	0.1206 (0.001)		
				0.5	0.0640 (0.001)	5	0.0090 (0.000)
						20	0.0099 (0.000)
						50	0.0091 (0.000)
				0.7	0.0316 (0.000)		
				0.9	0.0114 (0.000)	5	0.0090 (0.000)
						20	*0.0083* (0.000)
						50	0.0086 (0.000)
10	0.0219 (0.000)	0.0178 (0.001)	0.0159 (0.001)	0.1	0.0174 (0.000)	5	0.0137 (0.000)
						20	0.0165 (0.000)
						50	0.0138 (0.000)
				0.3	0.0224 (0.000)		
				0.5	0.0219 (0.000)	5	0.0151 (0.000)
						20	0.0144 (0.000)
						50	0.0143 (0.000)
				0.7	0.0174 (0.000)		
				0.9	*0.0115* (0.000)	5	0.0135 (0.000)
						20	0.0118 (0.000)
						50	0.0126 (0.000)
20	0.0189 (0.000)	0.0149 (0.000)	0.0127 (0.000)	0.1	0.0149 (0.000)	5	0.0117 (0.000)
						20	0.0144 (0.000)
						50	0.0119 (0.000)
				0.3	0.0193 (0.000)		
				0.5	0.0189 (0.000)	5	0.0127 (0.000)
						20	0.0124 (0.000)
						50	0.0121 (0.000)
				0.7	0.0149 (0.000)		
				0.9	*0.0097* (0.000)	5	0.0108 (0.000)
						20	*0.0097* (0.000)
						50	0.0102 (0.000)
50	0.0176 (0.000)	0.0132 (0.000)	0.0111 (0.000)	0.1	0.0139 (0.000)	5	0.0108 (0.000)
						20	0.0136 (0.000)
						50	0.0112 (0.000)
				0.3	0.0180 (0.000)		
				0.5	0.0176 (0.000)	5	0.0116 (0.000)
						20	0.0116 (0.000)
						50	0.0114 (0.000)
				0.7	0.0139 (0.000)		
				0.9	0.0089 (0.000)	5	0.0095 (0.000)
						20	*0.0088* (0.000)
						50	0.0093 (0.000)
100	0.0172 (0.000)	0.0123 (0.000)	0.0104 (0.000)	0.1	0.0136 (0.000)	5	0.0105 (0.000)
						20	0.0133 (0.000)
						50	0.0111 (0.000)
				0.3	0.0176 (0.000)		
				0.5	0.0172 (0.000)	5	0.0111 (0.000)
						20	0.0114 (0.000)
						50	0.0111 (0.000)
				0.7	0.0135 (0.000)		
				0.9	0.0087 (0.000)	5	0.0090 (0.000)
						20	*0.0086* (0.000)
						50	0.0089 (0.000)

Fortsetzung auf der nächsten Seite

	ILDA	QDA-AF	LDA-AF	OLDC fix		OLDC adaptive	
N_{trend}				λ,λ_{start}		L	
200	**0.0170** (0.000)	0.0118 (0.000)	0.0100 (0.000)	0.1	0.0134 (0.000)	5	0.0102 (0.000)
						20	0.0132 (0.000)
						50	0.0109 (0.000)
				0.3	0.0174 (0.000)		
				0.5	0.0170 (0.000)	5	0.0109 (0.000)
						20	0.0113 (0.000)
						50	0.0109 (0.000)
				0.7	0.0134 (0.000)		
				0.9	**0.0085** (0.000)	5	0.0087 (0.000)
						20	*0.0083* (0.000)
						50	0.0085 (0.000)
300	**0.0170** (0.000)	**0.0117** (0.000)	**0.0099** (0.000)	0.1	0.0134 (0.000)	5	0.0102 (0.000)
						20	0.0131 (0.000)
						50	0.0107 (0.000)
				0.3	0.0173 (0.000)		
				0.5	0.0170 (0.000)	5	0.0108 (0.000)
						20	0.0112 (0.000)
						50	0.0107 (0.000)
				0.7	0.0134 (0.000)		
				0.9	**0.0085** (0.000)	5	0.0086 (0.000)
						20	***0.0082*** (0.000)
						50	0.0084 (0.000)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0076 (Standardabweichung 0.017)

Wie bei der Datensituation „Kreuzen“ hängen aufgrund des Trends bei *ILDA* die Schätzer für die Erwartungswertvektoren zeitlich hinterher, wodurch keine erwartungstreuen Schätzer zur Bildung der aktuellen Klassifikationsregel zu jedem Zeitpunkt herangezogen werden (vgl. obere linke Grafik in Abbildung 9.40). Durch die Modellierung des Trends durch lokale lineare Regressionsmodelle und Prognose der Erwartungswertvektoren durch diese Modelle wird dieses Problem behoben. In Abbildung 9.40 wird auch für diese Datensituation deutlich, dass für kleine N_{trend} die Schätzer zunächst noch unsicher sind. Bei genügend großem Fenster N_{trend} für die einzelnen Regressionsmodelle kann vermutlich eine erwartungstreue Schätzung der Erwartungswertvektoren der Klassen gewährleistet werden und es fließen repräsentative Schätzer für die Erwartungswertvektoren zu jedem Zeitpunkt in die Klassifikationsregel der LDA ein.

Da die Verteilungen der Klasse 1 zu jedem Zeitpunkt in den Datensituationen „Kreuzen“ und „Vorbeilaufen“ identisch sind (vgl. Seite 220 ff.), sind die durchschnittlichen euklidischen Abstände zwischen den wahren und den mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren über die Zeit in Klasse 1 bei den Methoden *ILDA*,*OLDC fix* und *QDA-AF* ebenfalls identisch (vgl. Tabellen 9.8 und 9.10). Die durchschnittlichen euklidischen Abstände über die Zeit für Klasse 2 unterscheiden sich für die ursprünglichen Methoden aufgrund abweichender Verteilungen von jenen der Datensituation „Kreuzen“. Bei Erweiterung von *ILDA* und *OLDC* mit fester Lernrate sind die Resultate jedoch auch für Klasse 2 identisch zu jenen der vorherigen Datensituation „Kreuzen“. Bei *QDA-AF* sind sie lediglich ähnlich, da anders als bei den anderen beiden Methoden der aktualisierte Schätzer $\tilde{\boldsymbol{\Sigma}}_t^{(2)}$ für die klassenspezifische Kovarianzmatrix durch den Faktor $\lambda^{(2)}_{(n_t^{(2)})}$ bei der Aktualisierung des Schätzers für den Erwartungswertvektor einfließt, welcher in den

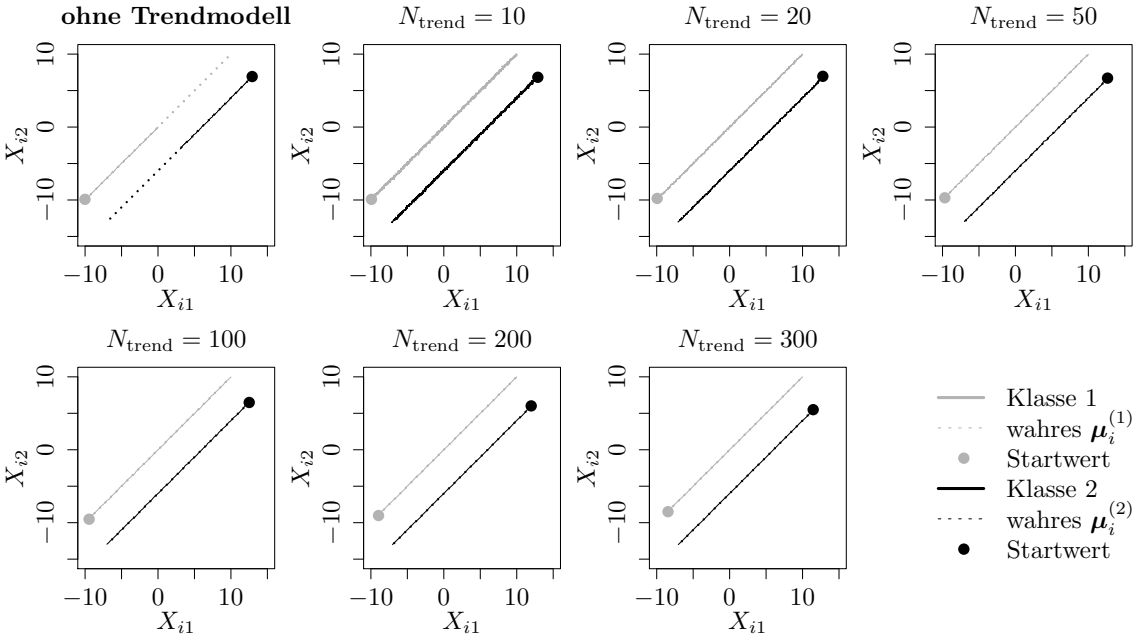


Abbildung 9.40: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren auf der Datensituation **„Vorbeilaufen“** für den gesamten Datenstrom. Ergebnisse aus *ILDA* und Erweiterung mit verschiedenen N_{trend}.

Datensituationen natürlich unterschiedlich ist (vgl. Seiten 79–81). Dies kann sich unterschiedlich auf die Regressionsmodelle und folglich Prognosen des Erwartungswertvektors bei den Datensituationen „Kreuzen“ und „Vorbeilaufen“ auswirken, was die geringen Abweichungen bezüglich der euklidischen Abstände von Klasse 2 der beiden Datensituationen erklärt. Dass die Ergebnisse bei *ILDA* und *OLDC fix* nach Erweiterung hingegen identisch sind liegt daran, dass der Trend in Klasse 2 zwar unterschiedlich verläuft, die „Stärke“ bzw. „Geschwindigkeit“ des Drifts jedoch in beiden Datensituationen identisch ist. Das heißt zu jedem Zeitpunkt bewegt sich der Erwartungswertvektor um 0.005 in beiden Dimensionen. Innerhalb der Klasse 2 ist der „Verlauf“ des linearen Trends für beide Datensituationen daher ähnlich. Die Erwartungswerte werden durch die lokalen linearen Regressionsmodelle gleich gut prognostiziert, sodass schließlich beim euklidischen Abstand in beiden Situationen dieselben Ergebnisse resultieren.

Die durchschnittlichen euklidischen Abstände bei den anderen beiden Methoden (*LDA-AF* und *OLDC adaptive*) unterscheiden sich leicht in beiden Datensituationen trotz identischer Verteilung in Klasse 1, da die jeweiligen betrachteten Gewichtungen in den Methoden über die Zeit variabel sind. Bei *OLDC* mit adaptiver Lernrate hängt die Adaption der Lernrate λ vom Zusammenspiel der Verteilungen aller Klassen bzw. der Trennung zwischen den Klassen im Laufe des Datenstroms ab. Bei *LDA-AF* werden der Startwert für den Faktor $\lambda^{(c)}_{(n_0^{(c)})}$ und die Schrittweite $\alpha_0^{(c)}$ für beide Datensituationen zwar identisch (gleicher Startwert für den Zufallszahlengenerator) aus dem jeweiligen Intervall gezogen (vgl. Abschnitt 4.4). Die Unterschiede resultieren jedoch auch hier durch das Zusammenspiel der Verteilungen beider

Klassen, weil der Faktor $\lambda^{(c)}_{(n_t^{(c)})}$ auf Basis der gepoolten Kovarianzmatrix $\mathbf{\Sigma}_t^{(P)}$ aktualisiert wird (im Gegensatz zu den Aktualisierungen bei *QDA-AF*, vgl. Seite 86 ff.). Nach Erweiterung von *LDA-AF* sind die Ergebnisse jedoch ebenfalls vergleichbar mit denen der Datensituation „Kreuzen“, da die Erwartungswerte ähnlich gut prognostiziert werden und daher beim euklidischen Abstand sehr ähnliche Ergebnisse resultieren.

Die Einführung lokaler linearer Regressionsmodelle zur Modellierung des Trends und Prognose der Erwartungswertvektoren kann bei allen Methoden die durchschnittlichen euklidischen Abstände über die Zeit zwischen wahren und mittleren prognostizierten Erwartungswertvektoren in beiden Klassen deutlich verringern (vgl. „Zeilen“ in Tabelle 9.10). Zudem sinken die Werte mit wachsender Fensterbreite N_{trend} immer weiter.

Aufgrund der hohen durchschnittlichen euklidischen Abstände über die Zeit bei *ILDA* und *OLDC* mit kleinen Lernraten λ führt bei diesen die Erweiterung der Methoden zu einer absolut betrachtet stärkeren Verringerung der durchschnittlichen euklidischen Abstände über die Zeit als bei den übrigen Methoden. Nach Erweiterung der Methoden unterscheiden sich die Werte zwischen den Update-Methoden für die Diskriminanzanalyse kaum noch und die Unterschiede werden mit steigendem N_{trend} geringer, da bei allen Methoden die aktuellen Erwartungswertvektoren im Datenstrom aufgrund einer höheren Anzahl an Beobachtungen für die einzelnen lokalen linearen Regressionsmodelle immer besser geschätzt und prognostiziert werden können.

Fazit: Wie bei der Datensituation „Kreuzen“ ist die Annahme eines linearen Trends der Erwartungswertvektoren, auf Basis derer die Erweiterung der Methoden entwickelt wurde, erfüllt. Daher kann die Prognosegüte fast aller Methoden für Online Diskriminanzanalyse durch die Erweiterung deutlich verbessert werden, sowohl in Hinblick auf den Prognosefehler als auch die euklidischen Abstände zwischen wahren und geschätzten bzw. prognostizierten Erwartungswertvektoren über die Zeit als Maßzahl für die „Erwartungstreue“ der Schätzer. Dabei ist auch hier ein breites Fenster N_{trend} für die einzelnen lokalen linearen Regressionsmodelle am besten. Bei großem N_{trend} wird die konkrete Auswahl der speziellen Methode für Online Diskriminanzanalyse weniger relevant.

Tabelle 9.10: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ ($p = 2$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

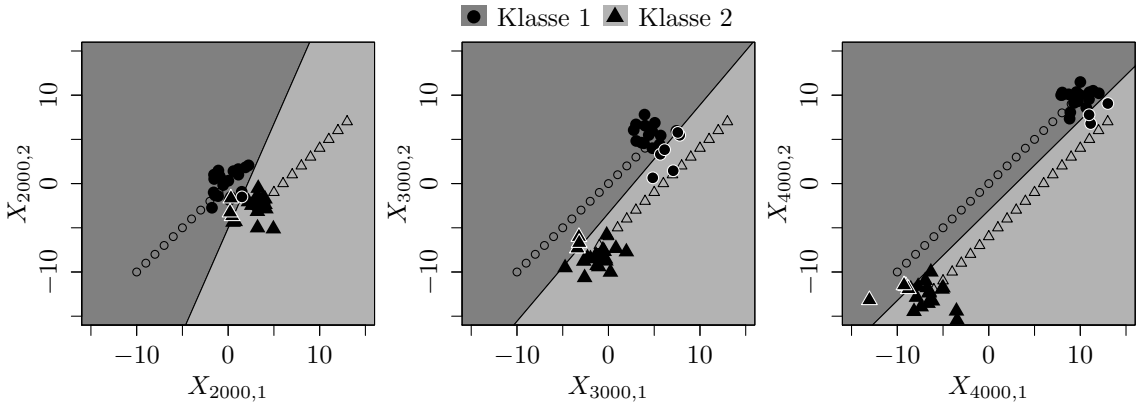
	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	7.0965 (0.013)	0.5012 (1.823)	*0.1942* (0.131)	0.1	12.7386 (0.105)	5	3.2760 (0.554)
	7.0880 (0.013)	0.4768 (1.736)	*0.1920* (0.140)		12.7207 (0.135)		3.2684 (0.580)
						20	4.8454 (0.805)
							4.8485 (0.824)
						50	5.5137 (0.868)
							5.5255 (0.880)
				0.3	9.9271 (0.015)		
					9.9093 (0.017)		
				0.5	7.0965 (0.013)	5	1.6926 (0.189)
					7.0880 (0.013)		1.6824 (0.183)
						20	2.6862 (0.291)
							2.6837 (0.293)
						50	3.3839 (0.321)
							3.3786 (0.316)
				0.7	4.2617 (0.016)		
					4.2614 (0.015)		
				0.9	1.4307 (0.022)	5	0.3589 (0.067)
					1.4317 (0.021)		0.3544 (0.066)
						20	0.6173 (0.046)
							0.6123 (0.046)
						50	0.9508 (0.223)
							0.9482 (0.217)
10	0.1010 (0.345)	0.1328 (0.630)	0.1336 (0.643)	0.1	*0.1005* (0.341)	5	0.1123 (0.431)
	0.1021 (0.345)	0.1320 (0.616)	0.1333 (0.634)		*0.1016* (0.342)		0.1126 (0.428)
						20	0.1015 (0.345)
							0.1024 (0.346)
						50	0.1011 (0.346)
							0.1024 (0.347)
				0.3	0.1007 (0.342)		
					0.1018 (0.343)		
				0.5	0.1010 (0.345)	5	0.1159 (0.466)
					0.1021 (0.345)		0.1159 (0.464)
						20	0.1022 (0.351)
							0.1032 (0.351)
						50	0.1022 (0.354)
							0.1032 (0.356)
				0.7	0.1017 (0.351)		
					0.1029 (0.351)		
				0.9	0.1049 (0.382)	5	0.1238 (0.543)
					0.1060 (0.378)		0.1233 (0.539)
						20	0.1079 (0.402)
							0.1088 (0.399)
						50	0.1074 (0.404)
							0.1088 (0.404)

Fortsetzung auf der nächsten Seite

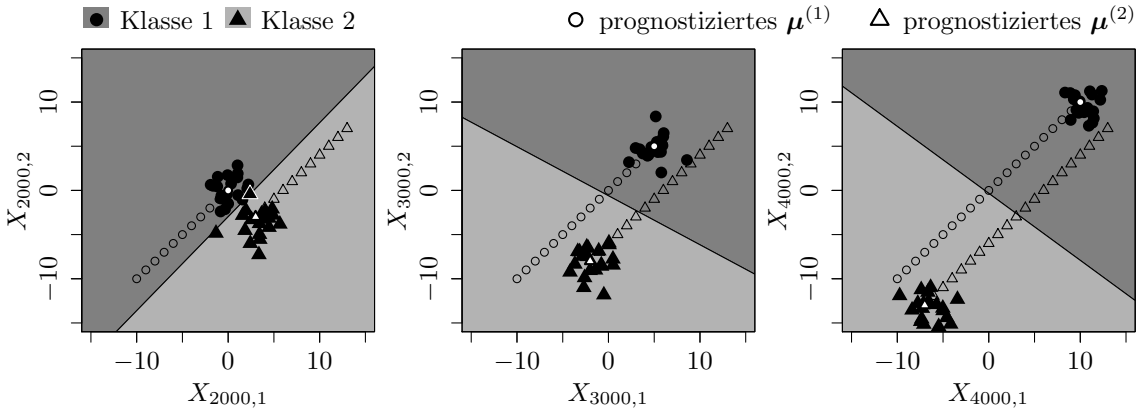
	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
20	0.0646 (0.126)	0.0932 (0.272)	0.0937 (0.274)	0.1	*0.0640* (0.124)	5	0.0761 (0.178)
	0.0666 (0.126)	0.0945 (0.270)	0.0948 (0.273)		*0.0659* (0.124)		0.0783 (0.179)
						20	0.0659 (0.129)
							0.0672 (0.129)
						50	0.0650 (0.129)
							0.0671 (0.129)
				0.3	0.0642 (0.124)		
					0.0661 (0.124)		
				0.5	0.0646 (0.126)	5	0.0789 (0.195)
					0.0666 (0.126)		0.0803 (0.195)
						20	0.0668 (0.132)
							0.0684 (0.131)
						50	0.0664 (0.134)
							0.0682 (0.135)
				0.7	0.0655 (0.129)		
					0.0675 (0.129)		
				0.9	0.0687 (0.144)	5	0.0856 (0.227)
					0.0710 (0.144)		0.0874 (0.228)
						20	0.0726 (0.160)
							0.0744 (0.160)
						50	0.0716 (0.160)
							0.0743 (0.161)
50	0.0398 (0.048)	0.0638 (0.128)	0.0637 (0.126)	0.1	*0.0390* (0.047)	5	0.0512 (0.083)
	0.0439 (0.049)	0.0671 (0.126)	0.0653 (0.125)		*0.0429* (0.048)		0.0551 (0.082)
						20	0.0436 (0.054)
							0.0471 (0.055)
						50	0.0418 (0.057)
							0.0461 (0.057)
				0.3	0.0393 (0.047)		
					0.0432 (0.048)		
				0.5	0.0398 (0.048)	5	0.0536 (0.089)
					0.0439 (0.049)		0.0563 (0.088)
						20	0.0442 (0.056)
							0.0485 (0.056)
						50	0.0439 (0.061)
							0.0475 (0.063)
				0.7	0.0410 (0.051)		
					0.0451 (0.052)		
				0.9	0.0451 (0.061)	5	0.0597 (0.105)
					0.0490 (0.062)		0.0622 (0.105)
						20	0.0522 (0.075)
							0.0552 (0.077)
						50	0.0492 (0.078)
							0.0531 (0.079)
100	0.0284 (0.026)	0.0467 (0.068)	0.0470 (0.067)	0.1	*0.0277* (0.026)	5	0.0383 (0.048)
	0.0339 (0.026)	0.0531 (0.067)	0.0501 (0.066)		*0.0324* (0.027)		0.0448 (0.048)
						20	0.0331 (0.034)
							0.0369 (0.035)
						50	0.0337 (0.043)
							0.0382 (0.043)
				0.3	0.0279 (0.025)		
					0.0330 (0.026)		
				0.5	0.0284 (0.026)	5	0.0407 (0.049)
					0.0339 (0.026)		0.0458 (0.049)
						20	0.0339 (0.034)
							0.0395 (0.034)
						50	0.0350 (0.043)
							0.0400 (0.043)
				0.7	0.0296 (0.028)		
					0.0355 (0.028)		
				0.9	0.0336 (0.036)	5	0.0451 (0.058)
					0.0395 (0.036)		0.0496 (0.058)
						20	0.0429 (0.048)
							0.0459 (0.048)
						50	0.0397 (0.051)
							0.0443 (0.051)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.0194 (0.015)	0.0329 (0.034)	0.0327 (0.034)	0.1	0.0193 (0.017)	5	0.0264 (0.029)
	0.0243 (0.015)	0.0377 (0.034)	0.0367 (0.034)		*0.0228* (0.018)		0.0349 (0.030)
						20	0.0236 (0.026)
							0.0294 (0.027)
						50	0.0275 (0.035)
							0.0313 (0.035)
				0.3	*0.0189* (0.015)		
					0.0234 (0.015)		
				0.5	0.0194 (0.015)	5	0.0277 (0.027)
					0.0243 (0.015)		0.0352 (0.027)
						20	0.0252 (0.022)
							0.0316 (0.022)
						50	0.0273 (0.029)
							0.0309 (0.027)
				0.7	0.0208 (0.016)		
					0.0261 (0.017)		
				0.9	0.0250 (0.022)	5	0.0324 (0.031)
					0.0301 (0.022)		0.0382 (0.032)
						20	0.0340 (0.029)
							0.0371 (0.030)
						50	0.0319 (0.030)
							0.0356 (0.030)
300	**0.0153** (0.011)	**0.0259** (0.022)	**0.0243** (0.022)	0.1	0.0165 (0.015)	5	0.0212 (0.023)
	0.0203 (0.011)	**0.0309** (0.023)	**0.0302** (0.023)		***0.0191*** (0.017)		0.0285 (0.025)
						20	0.0210 (0.024)
							0.0263 (0.025)
						50	0.0255 (0.034)
							0.0282 (0.034)
				0.3	***0.0151*** (0.012)		
					0.0195 (0.012)		
				0.5	0.0153 (0.011)	5	**0.0203** (0.019)
					0.0203 (0.011)		0.0285 (0.019)
						20	0.0205 (0.017)
							0.0272 (0.018)
						50	0.0232 (0.022)
							0.0269 (0.022)
				0.7	0.0166 (0.012)		
					0.0222 (0.013)		
				0.9	0.0200 (0.016)	5	0.0244 (0.022)
					0.0268 (0.017)		0.0318 (0.023)
						20	0.0276 (0.022)
							0.0324 (0.022)
						50	0.0262 (0.023)
							0.0324 (0.023)



(a) Ursprüngliche Methode *ILDA* bzw. *OLDC* mit $\lambda = 0.5$.



(b) Erweiterung durch lokale lineare Regressionsmodelle mit $N_{\text{trend}} = 50$.

Abbildung 9.41: Veränderung der Klassifikationsgrenze der LDA im Laufe der Zeit durch Aktualisierung des Modells mit neuen Beobachtungen aus Verteilungen mit verschobenem Erwartungswertvektor und Auswirkung auf den Prognosefehler für die Datensituation **„Vorbeilaufen“**.
Schwarze ∘ und △ veranschaulichen Erwartungswertvektoren der Verteilungen von Klasse 1 bzw. Klasse 2, aus denen Beobachtungen im Datenstrom realisiert werden, die in (a) für die Anpassung der LDA herangezogen werden.

Datensituation „Vorbeilaufen“ (gerade) ($p = 2$) Bei der Datensituation „Vorbeilaufen“ (gerade) laufen die Verteilungen beider Klassen auch mit der Zeit aufeinander zu und aneinander vorbei, im Gegensatz zur vorherigen Situation im zweidimensionalen Fall jedoch parallel zur x-Achse (vgl. Seite 222 f.). Da die parallelen Geraden, auf denen sich die Erwartungswertvektoren aneinander vorbei bewegen, nur noch um den Wert 3 auseinander liegen, hat dies zur Folge, dass der Bayesfehler in der Mitte des Datenstroms etwas stärker ansteigt als in der vorherigen Datensituation, da die Verteilungen der beiden Klassen sich stärker überlappen. Dies zieht auch einen erhöhten durchschnittlichen Bayesfehler über die Zeit von etwa 0.0222 nach sich.

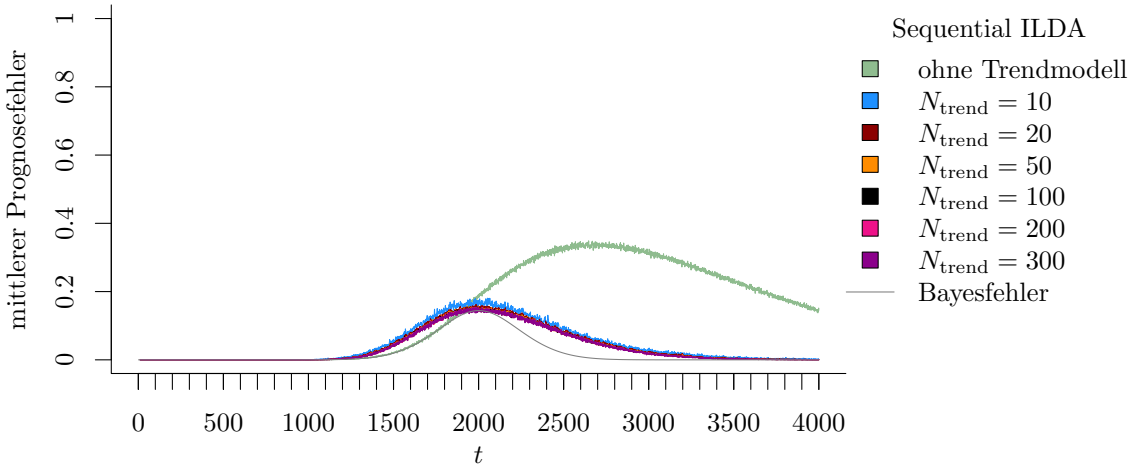
Die Resultate sind aufgrund der ähnlichen Datensituation stark vergleichbar mit jenen der vorherigen Datensituation. Anhand der Abbildungen 9.42–9.46 wird deutlich, dass die Verläufe der mittleren Prognosefehler für alle Methoden eine ähnliche Struktur aufweisen wie in der vorherigen Datensituation „Vorbeilaufen“ (vgl. Abbildungen 9.35–9.39).

Auch die durchschnittlichen mittleren Prognosefehler über die Zeit lassen dieselben Schlüsse zu wie bei der Datensituation „Vorbeilaufen“ (vgl. Tabelle 9.11). Der einzige Unterschied ist, dass die Fehler insgesamt alle etwas höher sind als in der vorherigen Datensituation – aufgrund des oben genannten Grundes der Lage der Verteilungen.

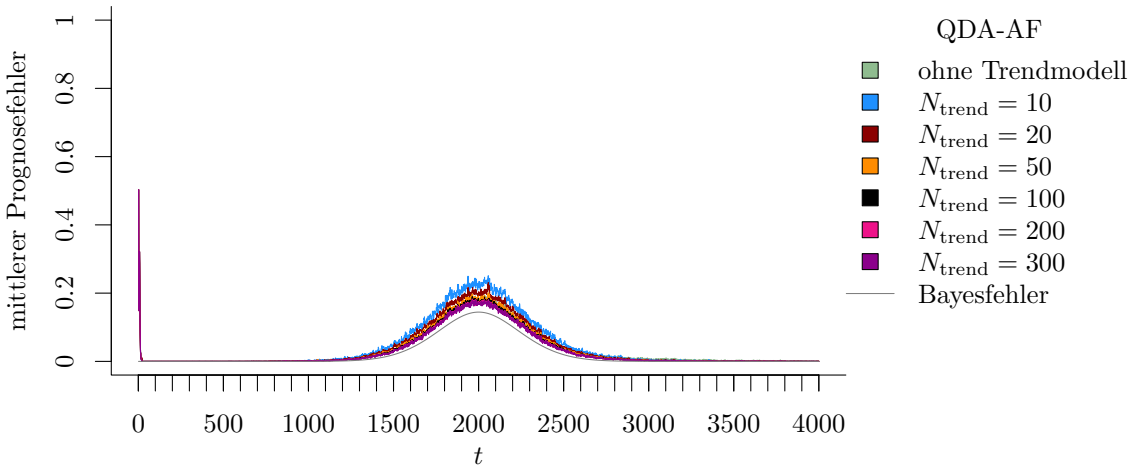
In Abbildung 9.49 auf Seite 315 am Ende des Abschnittes sind auch für diese Datensituation die Klassifikationsgrenzen bzw. die Veränderung der Trenngerade und des Prognosefehlers bei Anpassung durch *ILDA* für verschiedene Zeitpunkte mit und ohne Erweiterung durch lokale lineare Regressionsmodelle auf Fenstern der Breite $N_{\text{trend}} = 50$ veranschaulicht.

Die wichtigsten Erkenntnisse bezüglich des Prognosefehlers sind die Folgenden:

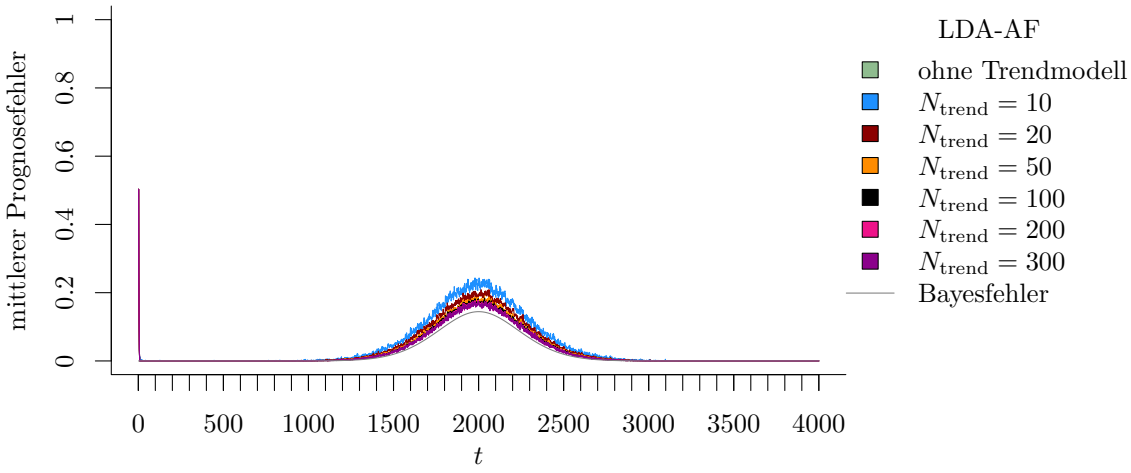
- Der durchschnittliche Bayesfehler über den gesamten Datenstrom für diese Datensituation beträgt 0.0222 mit einer Standardabweichung von 0.042.
- Für *ILDA* und *OLDC* wird direkt durch Integration von Regressionsmodellen eine Reduzierung des Prognosefehlers erzielt. Zudem sinkt dieser für steigendes N_{trend} immer weiter aufgrund der erfüllten Voraussetzung eines linearen Trends.
- Für *QDA-AF* und *LDA-AF* steigt der durchschnittliche mittlere Prognosefehler über die Zeit zunächst bei zusätzlicher Betrachtung von Trendmodellen im Vergleich zu den ursprünglichen Methoden leicht an. Für größere Fenster N_{trend} für die lokalen linearen Regressionsmodelle kann jedoch auch hier der durchschnittliche mittlere Prognosefehler über die Zeit reduziert werden.
- Bei *OLDC* mit fester Lernrate ist in allen Fällen eine hohe Lernrate λ am besten.
- *OLDC adaptive*: Für steigendes N_{trend} kann der durchschnittliche mittlere Prognosefehler über die Zeit minimal verringert werden. Die niedrigen Werte des ursprünglichen Modells werden für großes N_{trend} und $\lambda_{\text{start}} = 0.9$ leicht unterschritten.
- Der minimale durchschnittliche mittlere Prognosefehler über die Zeit von 0.0233 wird durch *OLDC* mit fester Lernrate von $\lambda = 0.9$ und der Integration lokaler linearer Regressionsmodelle auf Fenstern der Breite $N_{\text{trend}} = 300$ erzielt.



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

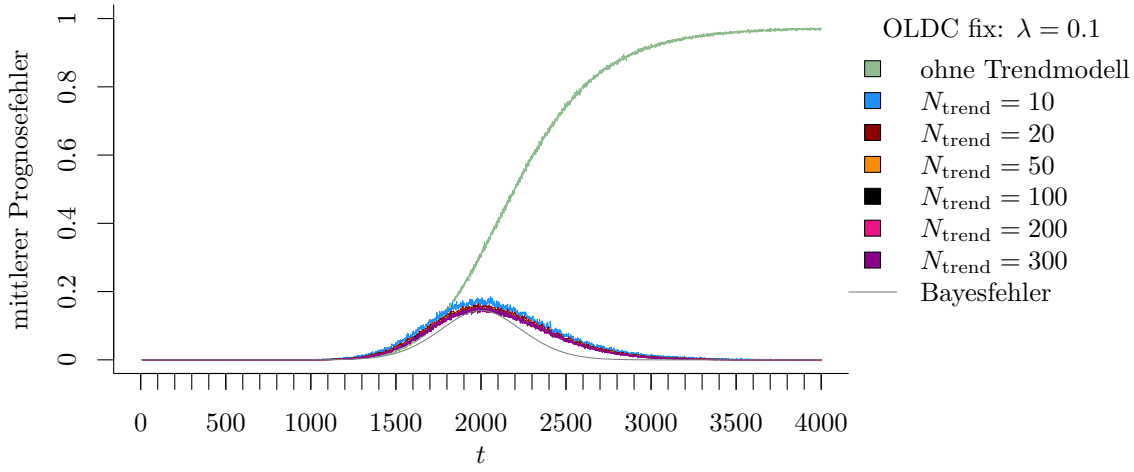


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

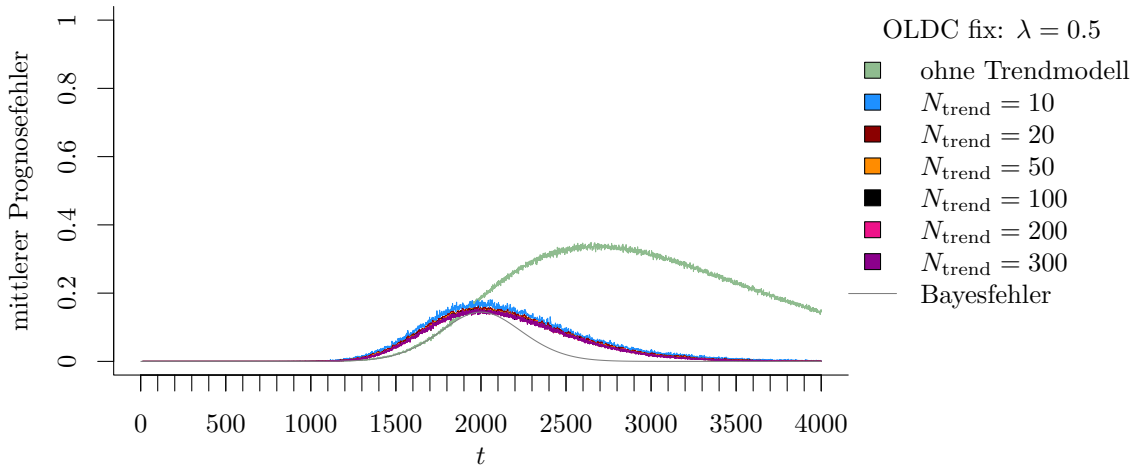


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

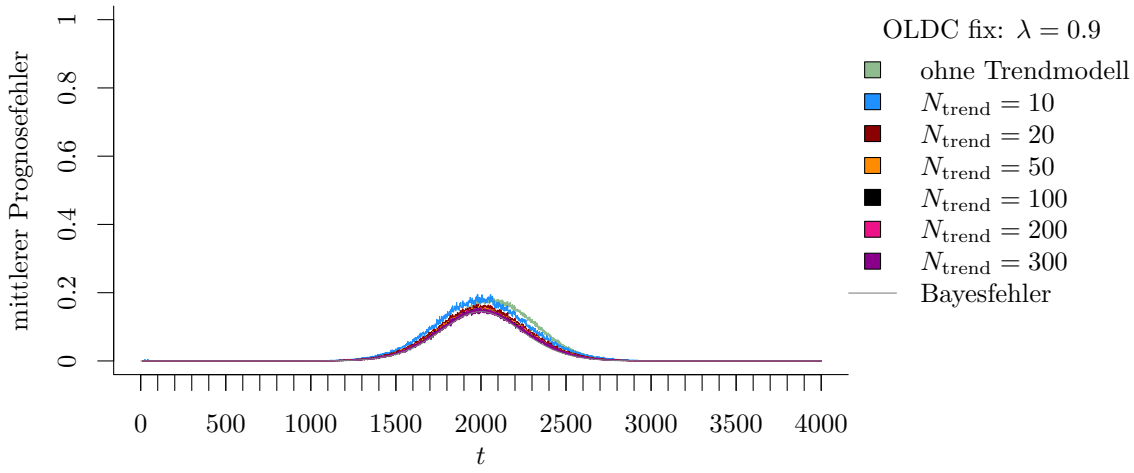
Abbildung 9.42: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“ (gerade)** im zweidimensionalen Raum.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.

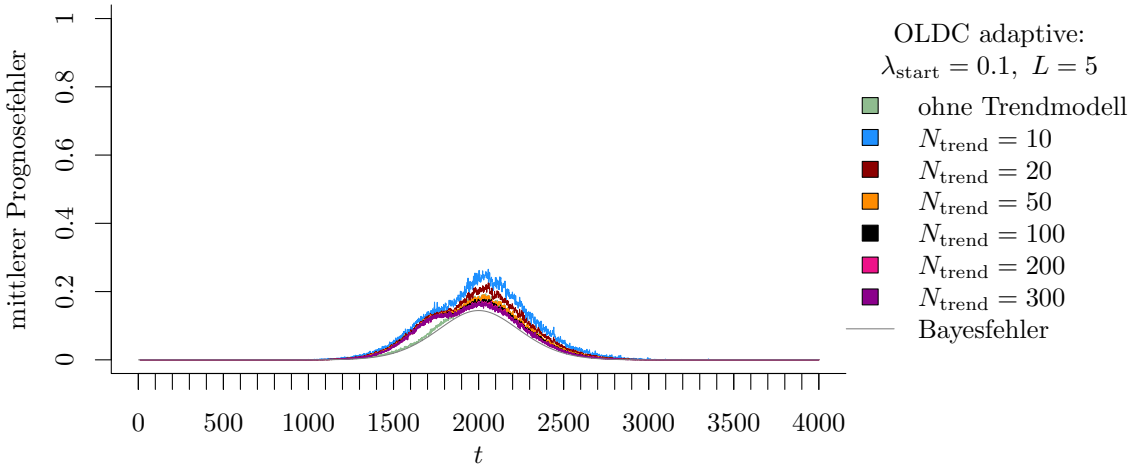


(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.

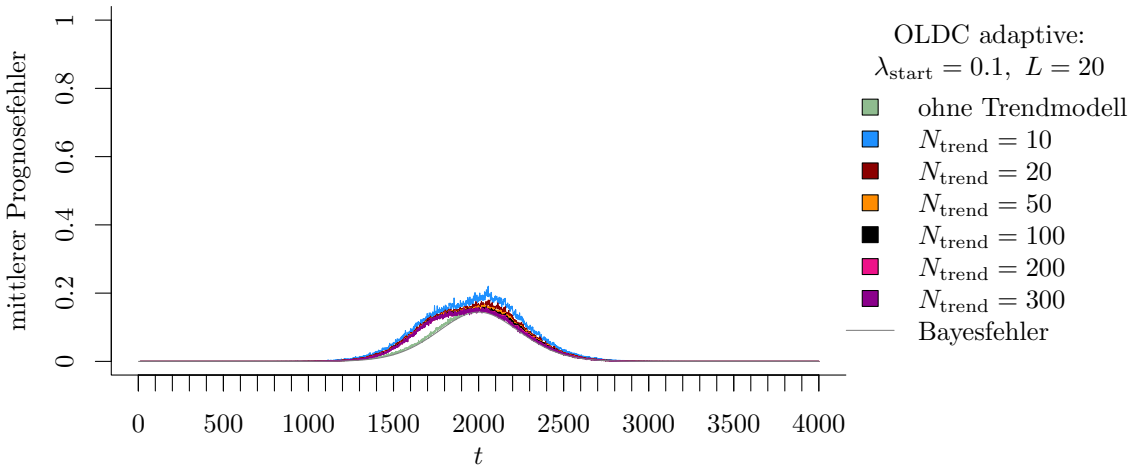


(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

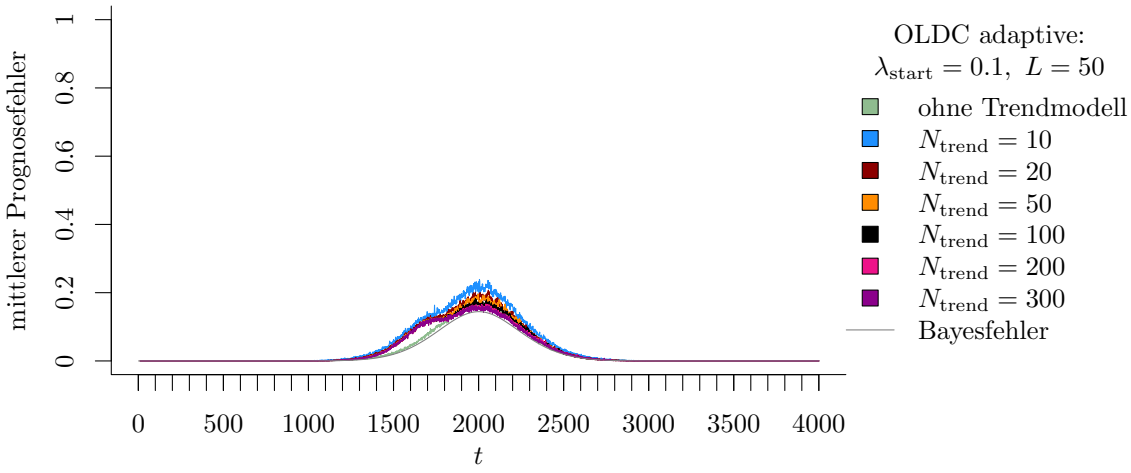
Abbildung 9.43: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“ (gerade)** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

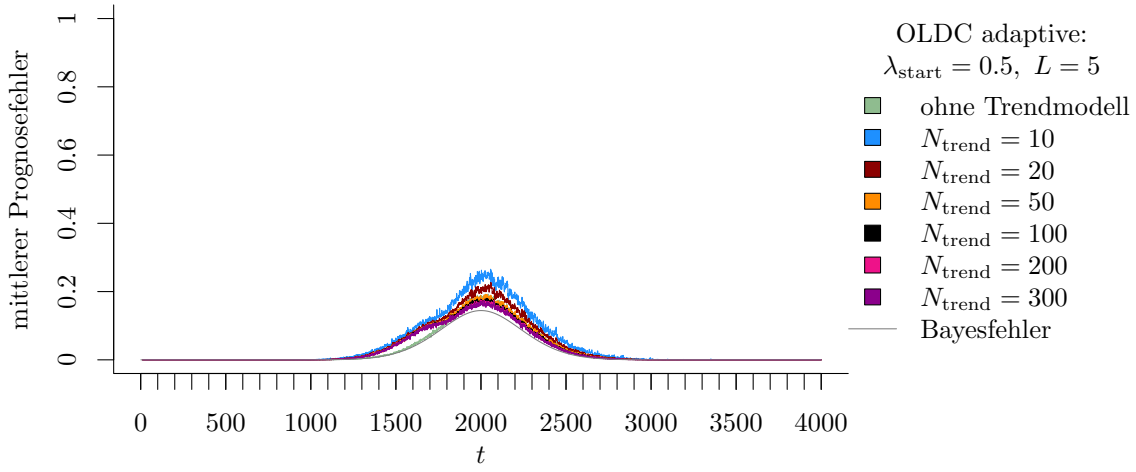


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

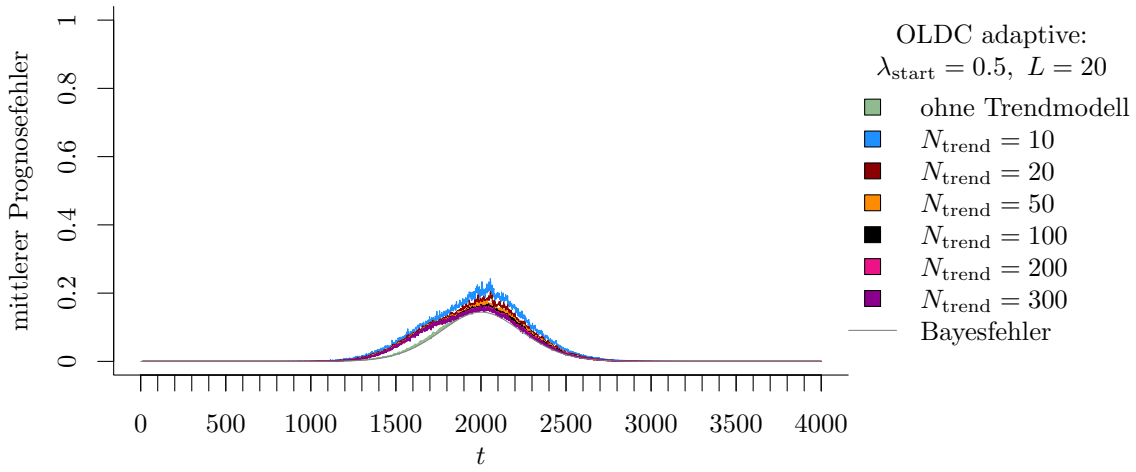


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

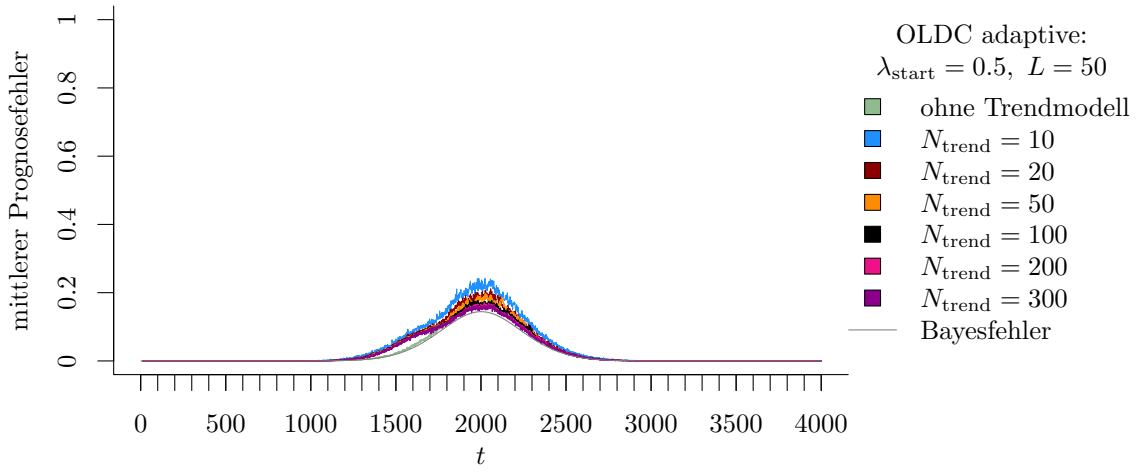
Abbildung 9.44: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“ (gerade)** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

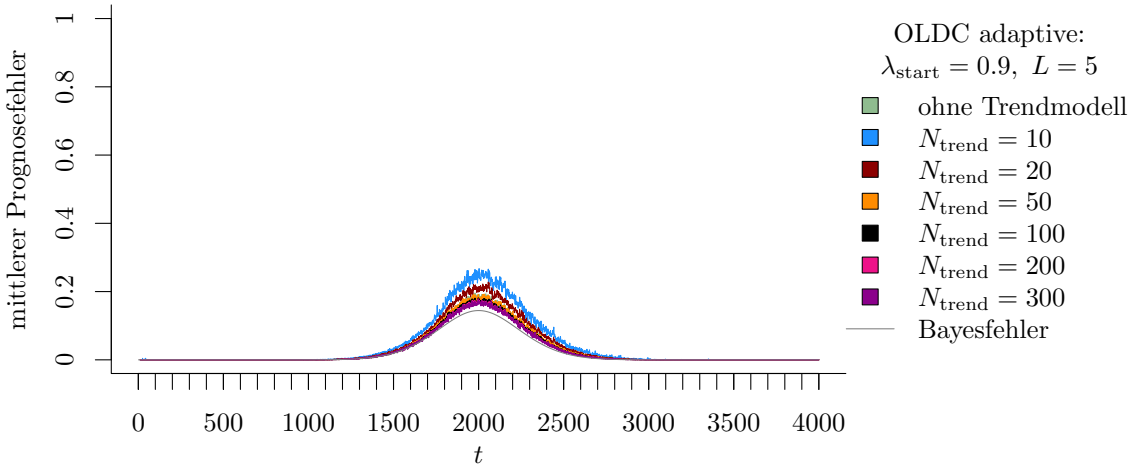


(b) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

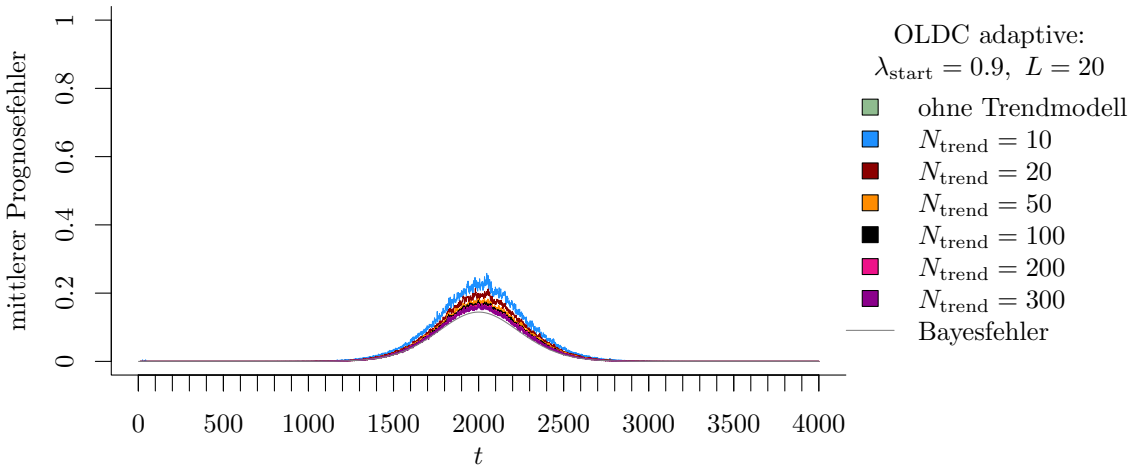


(c) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

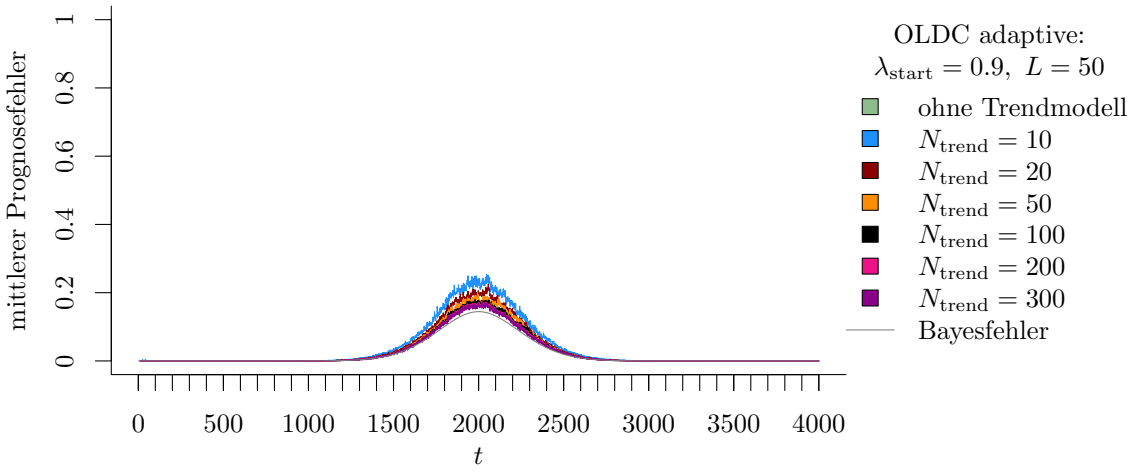
Abbildung 9.45: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“ (gerade)** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.46: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **„Vorbeilaufen“ (gerade)** im zweidimensionalen Raum.

Tabelle 9.11: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittl. Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ (gerade)** ($p = 2$) getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.1449 (0.001)	0.0332 (0.001)	0.0285 (0.000)	0.1	0.4311 (0.003)	5	0.0276 (0.000)
						20	*0.0244* (0.000)
						50	0.0256 (0.000)
				0.3	0.2526 (0.002)		
				0.5	0.1449 (0.001)	5	0.0276 (0.000)
						20	0.0245 (0.000)
						50	0.0256 (0.000)
				0.7	0.0707 (0.001)		
				0.9	0.0290 (0.000)	5	0.0279 (0.000)
						20	0.0254 (0.000)
						50	0.0262 (0.000)
10	0.0439 (0.001)	0.0419 (0.002)	0.0395 (0.002)	0.1	0.0378 (0.001)	5	0.0410 (0.002)
						20	0.0345 (0.001)
						50	0.0367 (0.001)
				0.3	0.0448 (0.001)		
				0.5	0.0439 (0.001)	5	0.0413 (0.002)
						20	0.0358 (0.001)
						50	0.0368 (0.001)
				0.7	0.0371 (0.001)		
				0.9	*0.0305* (0.001)	5	0.0395 (0.002)
						20	0.0354 (0.001)
						50	0.0361 (0.002)
20	0.0389 (0.000)	0.0362 (0.001)	0.0330 (0.001)	0.1	0.0331 (0.000)	5	0.0345 (0.001)
						20	0.0301 (0.000)
						50	0.0317 (0.001)
				0.3	0.0398 (0.000)		
				0.5	0.0389 (0.000)	5	0.0345 (0.001)
						20	0.0309 (0.001)
						50	0.0315 (0.001)
				0.7	0.0325 (0.000)		
				0.9	*0.0259* (0.000)	5	0.0325 (0.001)
						20	0.0298 (0.001)
						50	0.0303 (0.001)
50	0.0370 (0.000)	0.0325 (0.001)	0.0294 (0.001)	0.1	0.0314 (0.000)	5	0.0308 (0.000)
						20	0.0283 (0.000)
						50	0.0303 (0.001)
				0.3	0.0379 (0.000)		
				0.5	0.0370 (0.000)	5	0.0305 (0.000)
						20	0.0287 (0.000)
						50	0.0297 (0.001)
				0.7	0.0308 (0.000)		
				0.9	*0.0242* (0.000)	5	0.0283 (0.000)
						20	0.0268 (0.000)
						50	0.0279 (0.001)
100	0.0365 (0.000)	0.0308 (0.001)	0.0280 (0.000)	0.1	0.0309 (0.000)	5	0.0294 (0.000)
						20	0.0277 (0.000)
						50	0.0291 (0.000)
				0.3	0.0373 (0.000)		
				0.5	0.0365 (0.000)	5	0.0290 (0.000)
						20	0.0278 (0.000)
						50	0.0282 (0.000)
				0.7	0.0304 (0.000)		
				0.9	*0.0237* (0.000)	5	0.0267 (0.000)
						20	0.0256 (0.000)
						50	0.0263 (0.000)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	0.0362 (0.000)	0.0297 (0.001)	0.0271 (0.000)	0.1	0.0307 (0.000)	5	0.0286 (0.000)
						20	0.0272 (0.000)
						50	0.0280 (0.000)
				0.3	0.0371 (0.000)		
				0.5	0.0362 (0.000)	5	0.0282 (0.000)
						20	0.0270 (0.000)
						50	0.0271 (0.000)
				0.7	0.0301 (0.000)		
				0.9	*0.0234* (0.000)	5	0.0257 (0.000)
						20	0.0246 (0.000)
						50	0.0252 (0.000)
300	**0.0361** (0.000)	**0.0293** (0.001)	**0.0268** (0.000)	0.1	0.0306 (0.000)	5	0.0283 (0.000)
						20	0.0269 (0.000)
						50	0.0276 (0.000)
				0.3	0.0370 (0.000)		
				0.5	0.0361 (0.000)	5	0.0279 (0.000)
						20	0.0267 (0.000)
						50	0.0267 (0.000)
				0.7	0.0300 (0.000)		
				0.9	***0.0233*** (0.000)	5	0.0254 (0.000)
						20	**0.0243** (0.000)
						50	0.0248 (0.000)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0222 (Standardabweichung 0.042)

Der Verlauf der adaptiven Lernrate bei *OLDC adaptive* ist ebenfalls sehr ähnlich zu jener der Datensituation „Vorbeilaufen“ wie anhand der Abbildung 9.47 zu erkennen ist. Die adaptive Lernrate steigt für alle Startwerte λ_{start} und Fenstergrößen L kurz vor dem Zeitpunkt $t = 2000$ bis auf (annähernd) den maximalen Wert an. Für $L = 50$ erfolgt danach ein deutlicher Abfall der Lernrate, während die Lernrate bei schmaleren Fenstern L zur Adaption auf einem ähnlich hohen Niveau bis zum Ende des Datenstroms bleibt.

Die Abbildung 9.48 lässt dieselben Schlüsse bezüglich der Schätzer der Erwartungswertvektoren im Laufe des Datenstroms (beispielhaft) bei *ILDA* und der Erweiterung zu wie bei der vorherigen Datensituation „Vorbeilaufen“. Mit der ursprünglichen Methode werden die Erwartungswertvektoren im Datenstrom zeitverzögert repräsentativ geschätzt (vgl. linke obere Grafik). Durch die Erweiterung der Update-Methode wird dieses Problem behoben. Mit steigender Fenstergröße N_{trend} für die einzelnen Regressionsmodelle zur Modellierung des Trends sinkt dabei zusätzlich die Varianz der Schätzer.

Die durchschnittlichen euklidischen Abstände zwischen wahren und mittleren geschätzten (bei ursprünglichen Methoden) bzw. mittleren prognostizierten Erwartungswertvektoren (bei erweiterten Methoden) über die Zeit in Tabelle 9.12 sind ähnlich zu jenen der Datensituationen „Kreuzen“ (vgl. Tabelle 9.8) und „Vorbeilaufen“ (vgl. Tabelle 9.10), bei denen die Annahme eines linearen Trends der Erwartungswertvektoren ebenfalls erfüllt ist. Für die ursprünglichen Methoden unterscheiden sich die euklidischen Abstände noch aufgrund unterschiedlicher Verteilungen in beiden Klassen (vgl. Seite 220 ff.). Die Resultate bei Erweiterung von *ILDA* und *OLDC* mit fester Lernrate sind jedoch für beide Klassen identisch wie bei den vorherigen Datensituationen „Kreuzen“ und „Vorbeilaufen“. Dies liegt auch hier daran, dass die „Stärke“ bzw. „Geschwindigkeit“ des Drifts für beide Klassen in allen drei Datensituationen identisch ist und sich die Erwartungswertvektoren beider

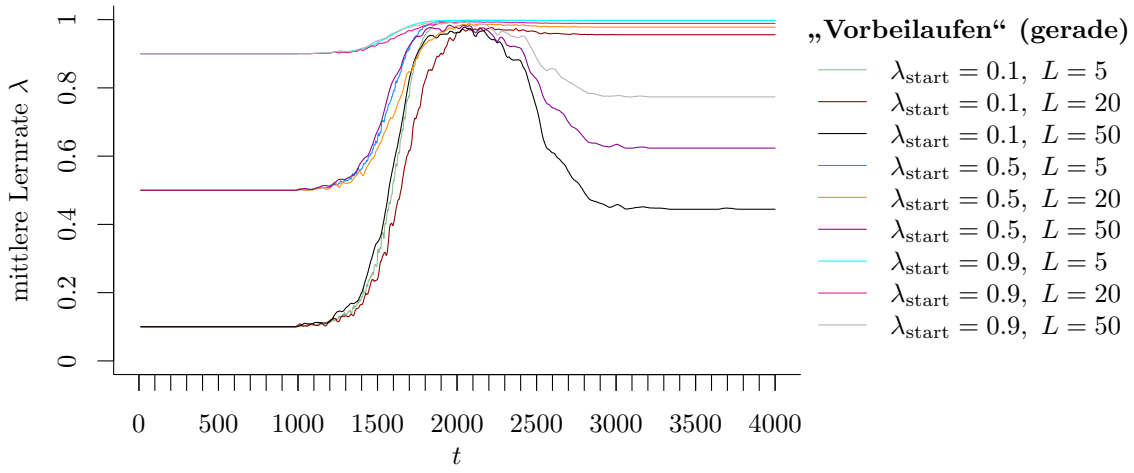


Abbildung 9.47: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **„Vorbeilaufen“ (gerade)**.

Klassen zu jedem Zeitpunkt um 0.005 in beiden bzw. hier in einer Dimension bewegen. Der „Verlauf“ des linearen Trends für alle drei Datensituationen ist somit ähnlich und die Erwartungswerte werden durch die lokalen linearen Regressionsmodelle gleich gut prognostiziert, sodass schließlich bezüglich des durchschnittlichen euklidischen Abstandes in allen drei Datensituationen dieselben Ergebnisse resultieren. Für die erweiterte Methode *OLDC adaptive* unterscheiden sich hingegen auch hier die Ergebnisse leicht gegenüber jenen der beiden vorherigen Datensituationen, da die adaptive Lernrate vom Zusammenspiel der Verteilungen beider Klassen abhängt (vgl. Seite 296). Die Ergebnisse der erweiterten Methoden *QDA-AF* und *LDA-AF* sind ähnlich zu jenen der Datensituationen „Kreuzen“ und „Vorbeilaufen“, allerdings nicht komplett identisch. Dies liegt daran, dass bei der Online Diskriminanzanalyse mit adaptivem Vergessen die Form der aktuellen klassenspezifischen Kovarianzmatrix $\tilde{\mathbf{\Sigma}}_t^{(c)}$ (*QDA-AF*) bzw. gepoolten Kovarianzmatrix $\tilde{\mathbf{\Sigma}}_t^{(P)}$ (*LDA-AF*) durch den Faktor $\lambda^{(c)}_{(n_t^{(c)})}$ in die Update-Formel des Schätzers für den Erwartungswertvektor einfließt (vgl. Abschnitt 4.4). Dies wirkt sich in allen drei Datensituationen qualitativ unterschiedlich auf die Schätzer der Erwartungswertvektoren aus, die dann bei der Erweiterung der jeweiligen Methode für die lokalen linearen Regressionsmodelle und die Prognose der Erwartungswertvektoren herangezogen werden (vgl. auch Erklärung auf Seite 295 ff.).

Der lineare Trend der Erwartungswertvektoren führt dazu, dass diese mit steigendem N_{trend} immer besser geschätzt bzw. prognostiziert werden können. Die lokalen linearen Regressionsmodelle basieren auf mehr aktualisierten Mittelwerten, was eine stabilere Prognose der zukünftigen Erwartungswerte gewährleistet. Dadurch kann auch hier der durchschnittliche euklidische Abstand über die Zeit in beiden Klassen durch Erweiterung der Methoden in allen Fällen verbessert werden. Zudem sinkt der durchschnittliche euklidische Abstand mit steigendem N_{trend} in allen Methoden (außer bei *OLDC adaptive*) immer weiter, wobei sich die Ergebnisse der verschiedenen Methoden immer weiter annähern.

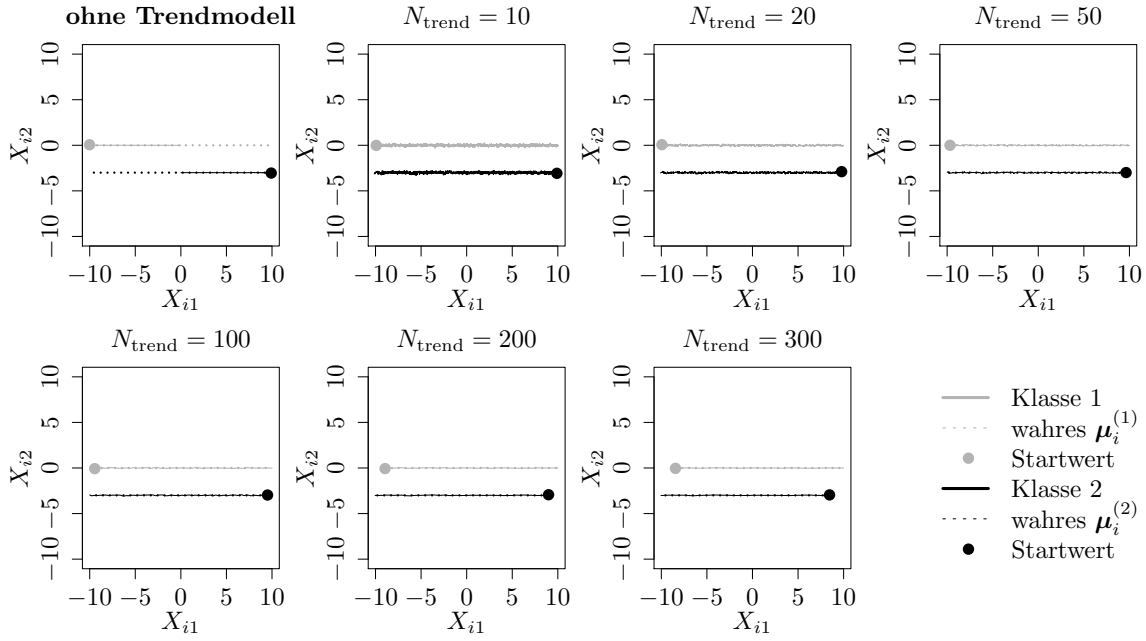


Abbildung 9.48: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren auf der Datensituation **„Vorbeilaufen" (gerade)** für den gesamten Datenstrom. Ergebnisse aus *ILDA* und Erweiterung mit verschiedenen N_{trend}.

Bei *OLDC* ist zudem zu sehen, dass bei der ursprünglichen Methode der Parameter λ einen großen Einfluss hat. Die durchschnittlichen euklidischen Abstände zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren über die Zeit hängen sehr stark von der Lernrate ab, wobei sie mit steigender Lernrate λ (*OLDC fix*) bzw. λ_{start} (*OLDC adaptive*) sinken. Zudem wachsen die durchschnittlichen euklidischen Abstände mit steigender Fenstergröße L (*OLDC adaptive*). Beide Effekte sind nach Erweiterung der Methoden stark reduziert (vgl. Spalten „OLDC fix" und „OLDC adaptive" in Tabelle 9.12). Die durchschnittlichen euklidischen Abstände sinken dann sogar teilweise mit steigendem Parameter L.

Fazit: Auch hier ist die Annahme eines linearen Trends der Erwartungswertvektoren in beiden Klassen erfüllt. Die Prognosegüte fast aller Methoden für Online Diskriminanzanalyse kann daher auch in dieser Datensituation durch die Erweiterung deutlich verbessert werden. Die Prognosegüte wird dabei tendenziell mit steigendem N_{trend} besser. Nach Erweiterung der Methoden spielt bei genügend großem Fenster N_{trend} für die einzelnen lokalen linearen Regressionsmodelle sowohl die Wahl der ursprünglichen Methode zur Aktualisierung der LDA (bzw. QDA) sowie auch die Wahl eventueller Parameter keine allzu große Rolle mehr.

Tabelle 9.12: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ (gerade)** ($p = 2$) getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

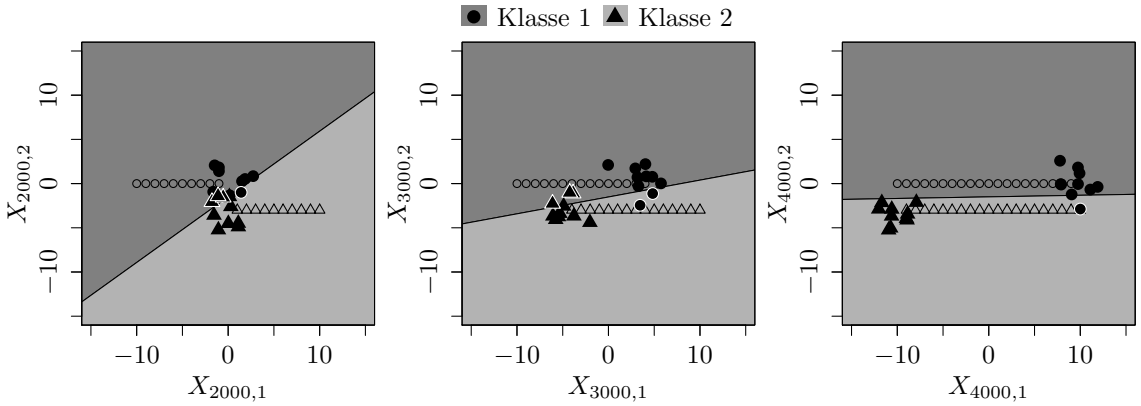
	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	5.0205 (0.008)	0.2962 (0.744)	*0.1465* (0.110)	0.1	9.0399 (0.114)	5	1.7851 (0.229)
	5.0130 (0.009)	0.2983 (0.738)	*0.1454* (0.115)		8.9901 (0.202)		1.7652 (0.264)
						20	2.1544 (0.294)
							2.1314 (0.329)
						50	3.0579 (0.453)
							3.0293 (0.461)
				0.3	7.0292 (0.012)		
					7.0083 (0.019)		
				0.5	5.0205 (0.008)	5	0.9427 (0.106)
					5.0130 (0.009)		0.9340 (0.104)
						20	1.0933 (0.074)
							1.0853 (0.072)
						50	1.9791 (0.443)
							1.9682 (0.428)
				0.7	3.0135 (0.011)		
					3.0140 (0.010)		
				0.9	1.0105 (0.019)	5	0.2128 (0.079)
					1.0125 (0.018)		0.2097 (0.079)
						20	0.2613 (0.048)
							0.2582 (0.047)
						50	0.9459 (0.570)
							0.9411 (0.575)
10	0.1010 (0.345)	0.1322 (0.620)	0.1329 (0.635)	0.1	*0.1005* (0.341)	5	0.1243 (0.556)
	0.1021 (0.345)	0.1314 (0.616)	0.1326 (0.632)		*0.1016* (0.342)		0.1244 (0.560)
						20	0.1044 (0.370)
							0.1056 (0.371)
						50	0.1038 (0.375)
							0.1052 (0.378)
				0.3	0.1007 (0.342)		
					0.1018 (0.343)		
				0.5	0.1010 (0.345)	5	0.1250 (0.564)
					0.1021 (0.345)		0.1255 (0.571)
						20	0.1072 (0.393)
							0.1078 (0.396)
						50	0.1056 (0.387)
							0.1068 (0.392)
				0.7	0.1017 (0.351)		
					0.1029 (0.351)		
				0.9	0.1049 (0.382)	5	0.1292 (0.605)
					0.1060 (0.378)		0.1295 (0.613)
						20	0.1148 (0.466)
							0.1157 (0.468)
						50	0.1113 (0.444)
							0.1120 (0.446)

Fortsetzung auf der nächsten Seite

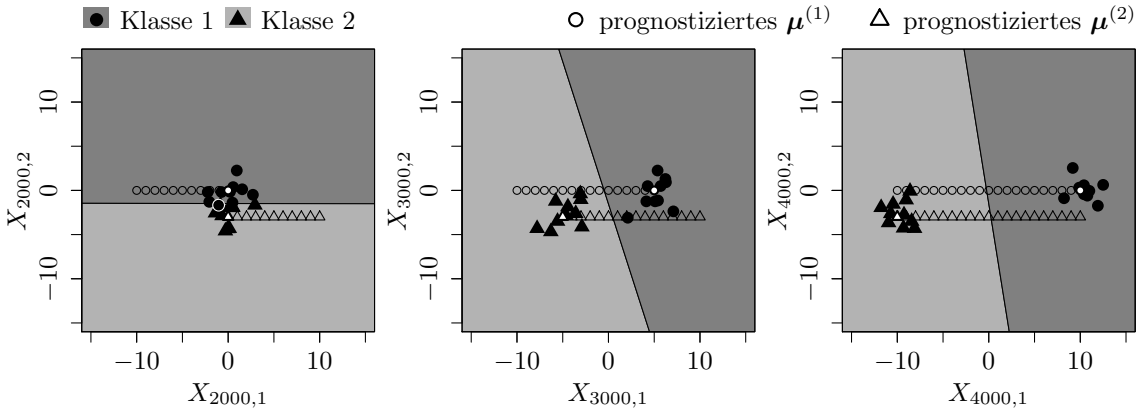
	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
20	0.0646 (0.126)	0.0928 (0.271)	0.0931 (0.273)	0.1	*0.0640* (0.124)	5	0.0857 (0.231)
	0.0666 (0.126)	0.0940 (0.269)	0.0941 (0.271)		*0.0659* (0.124)		0.0875 (0.231)
						20	0.0704 (0.151)
							0.0725 (0.150)
						50	0.0684 (0.153)
							0.0710 (0.151)
				0.3	0.0642 (0.124)		
					0.0661 (0.124)		
				0.5	0.0646 (0.126)	5	0.0860 (0.235)
					0.0666 (0.126)		0.0877 (0.234)
						20	0.0733 (0.167)
							0.0748 (0.167)
						50	0.0698 (0.159)
							0.0722 (0.159)
				0.7	0.0655 (0.129)		
					0.0675 (0.129)		
				0.9	0.0687 (0.144)	5	0.0899 (0.253)
					0.0710 (0.144)		0.0915 (0.254)
						20	0.0801 (0.204)
							0.0823 (0.204)
						50	0.0750 (0.183)
							0.0774 (0.185)
50	0.0398 (0.048)	0.0639 (0.127)	0.0634 (0.125)	0.1	*0.0390* (0.047)	5	0.0568 (0.101)
	0.0439 (0.049)	0.0671 (0.125)	0.0650 (0.124)		*0.0429* (0.048)		0.0614 (0.100)
						20	0.0528 (0.072)
							0.0567 (0.073)
						50	0.0491 (0.106)
							0.0532 (0.095)
				0.3	0.0393 (0.047)		
					0.0432 (0.048)		
				0.5	0.0398 (0.048)	5	0.0576 (0.102)
					0.0439 (0.049)		0.0610 (0.101)
						20	0.0544 (0.080)
							0.0588 (0.082)
						50	0.0499 (0.104)
							0.0547 (0.096)
				0.7	0.0410 (0.051)		
					0.0451 (0.052)		
				0.9	0.0451 (0.061)	5	0.0620 (0.113)
					0.0490 (0.062)		0.0653 (0.113)
						20	0.0594 (0.100)
							0.0644 (0.100)
						50	0.0546 (0.107)
							0.0574 (0.105)
100	0.0284 (0.026)	0.0470 (0.067)	0.0468 (0.067)	0.1	*0.0277* (0.026)	5	0.0420 (0.054)
	0.0339 (0.026)	0.0534 (0.067)	0.0498 (0.066)		*0.0324* (0.027)		0.0473 (0.054)
						20	0.0423 (0.047)
							0.0481 (0.048)
						50	0.0416 (0.080)
							0.0447 (0.070)
				0.3	0.0279 (0.025)		
					0.0330 (0.026)		
				0.5	0.0284 (0.026)	5	0.0428 (0.054)
					0.0339 (0.026)		0.0475 (0.053)
						20	0.0440 (0.049)
							0.0497 (0.050)
						50	0.0404 (0.071)
							0.0454 (0.066)
				0.7	0.0296 (0.028)		
					0.0355 (0.028)		
				0.9	0.0336 (0.036)	5	0.0461 (0.062)
					0.0395 (0.036)		0.0514 (0.061)
						20	0.0473 (0.060)
							0.0539 (0.060)
						50	0.0423 (0.072)
							0.0463 (0.067)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.0194 (0.015)	0.0330 (0.034)	0.0327 (0.034)	0.1	0.0193 (0.017)	5	0.0295 (0.031)
	0.0243 (0.015)	0.0381 (0.034)	0.0366 (0.034)		*0.0228* (0.018)		0.0374 (0.032)
						20	0.0313 (0.032)
							0.0392 (0.033)
						50	0.0325 (0.048)
							0.0366 (0.046)
				0.3	*0.0189* (0.015)		
					0.0234 (0.015)		
				0.5	0.0194 (0.015)	5	0.0302 (0.028)
					0.0243 (0.015)		0.0375 (0.028)
						20	0.0329 (0.029)
							0.0421 (0.030)
						50	0.0293 (0.040)
							0.0368 (0.040)
				0.7	0.0208 (0.016)		
					0.0261 (0.017)		
				0.9	0.0250 (0.022)	5	0.0332 (0.032)
					0.0301 (0.022)		0.0403 (0.033)
						20	0.0358 (0.033)
							0.0437 (0.034)
						50	0.0319 (0.040)
							0.0379 (0.039)
300	**0.0153** (0.011)	**0.0261** (0.022)	**0.0243** (0.022)	0.1	0.0165 (0.015)	5	**0.0239** (0.024)
	0.0203 (0.011)	**0.0310** (0.023)	**0.0301** (0.023)		***0.0191*** (0.017)		0.0314 (0.026)
						20	0.0277 (0.027)
							0.0339 (0.029)
						50	0.0296 (0.047)
							0.0342 (0.047)
				0.3	***0.0151*** (0.012)		
					0.0195 (0.012)		
				0.5	0.0153 (0.011)	5	0.0240 (0.020)
					0.0203 (0.011)		**0.0312** (0.020)
						20	0.0264 (0.022)
							0.0368 (0.023)
						50	0.0249 (0.037)
							0.0318 (0.035)
				0.7	0.0166 (0.012)		
					0.0222 (0.013)		
				0.9	0.0200 (0.016)	5	0.0265 (0.022)
					0.0268 (0.017)		0.0342 (0.023)
						20	0.0290 (0.023)
							0.0374 (0.025)
						50	0.0256 (0.033)
							0.0331 (0.031)



(a) Ursprüngliche Methode *ILDA* bzw. *OLDC* mit $\lambda = 0.5$.



(b) Erweiterung durch lokale lineare Regressionsmodelle mit $N_{\text{trend}} = 50$.

Abbildung 9.49: Veränderung der Klassifikationsgrenze der LDA im Laufe der Zeit durch Aktualisierung des Modells mit neuen Beobachtungen aus Verteilungen mit verschobenem Erwartungswertvektor und Auswirkung auf den Prognosefehler für die Datensituation **„Vorbeilaufen“ (gerade)**.
Schwarze ∘ und △ veranschaulichen Erwartungswertvektoren der Verteilungen von Klasse 1 bzw. Klasse 2, aus denen Beobachtungen im Datenstrom realisiert werden, die in (a) für die Anpassung der LDA herangezogen werden.

Datensituation Gradual Drift mit „Kreuzen“ ($p = 2$) In dieser Datensituation erfolgt ein „weicher“ Übergang von einer Verteilung zur anderen. Die Beobachtungen aus Klasse 1 resultieren zunächst aus einer multivariaten Normalverteilung mit Erwartungswertvektor $(-10, -10)^T$ und im Laufe des Datenstroms mit immer größerer Wahrscheinlichkeit aus der multivariaten Normalverteilung mit Erwartungswertvektor $(10, 10)^T$. Für die zweite Klasse wird der Erwartungswertvektor $(10, -10)^T$ nach und nach durch $(-10, 10)^T$ ersetzt (vgl. Seite 223). Die Kovarianzmatrix ist in beiden Klassen und zu allen Zeitpunkten identisch.

Der Bayesfehler zu jedem Zeitpunkt der Datensituation wird hier so bestimmt, dass die beiden Verteilungen der beiden Klassen entsprechend mit dem aktuellen Anteil (der „Wahrscheinlichkeit“, mit der eine Beobachtung jeweils aus der ersten oder zweiten Verteilung resultiert) gewichtet werden. Dies führt zu einem Bayesfehler, welcher durchgehend kleiner als 0.0001 ist (vgl. Abbildungen 9.52–9.56).

Bei den Methoden ohne Anpassung an einen concept drift steigt der Prognosefehler etwa linear von 0 bis 1 im Laufe der Zeit an, bevor er kurz vor Ende des Datenstroms wieder auf 0.5 abfällt (vgl. grüne Kurven in Abbildungen 9.52 (a) und 9.53 (b)). Dieser Verlauf kann durch die Einführung einer adaptiven oder hohen festen Lernrate von $\lambda = 0.9$ bei *OLDC* etwas verbessert werden (vgl. Abbildungen 9.53 (c) und 9.54–9.56). Der Prognosefehler fällt ab Mitte des Datenstroms und einem Wert von etwa 0.5 wieder ab. Der Bayesfehler, welcher durchgehend fast 0 ist, wird jedoch bei Weitem nicht approximiert. Generell können demnach die meisten der betrachteten Methoden für Online Diskriminanzanalyse nicht gut mit dieser Art von concept drift bzw. dieser speziellen Datensituation umgehen.

Die einzige Ausnahme stellt die Methode *QDA-AF* dar (vgl. Abbildung 9.52 (b)), welche sich bereits in ihrer ursprünglichen Variante sehr gut an den gradual drift anpassen kann, sodass der Bayesfehler zu jedem Zeitpunkt approximiert wird. Dieses Ergebnis wird auch anhand der durchschnittlichen mittleren Prognosefehler über die Zeit in Tabelle 9.13 deutlich. Für *QDA-AF* schwankt dieser zwischen 0.0040 und 0.0046. Die Prognosegüte kann hier durch die Integration lokaler linearer Trendmodelle nur noch leicht verbessert werden. Dies liegt daran, dass *QDA-AF* im Gegensatz zu allen anderen Methoden eine Update-Methode für die Quadratische Diskriminanzanalyse ist und demnach nicht-lineare Trennungen möglich sind. Die Datensituation kann bereits von einer einfachen QDA (ohne Gewichtungen) sehr gut angepasst werden, da sich lediglich der Anteil an Beobachtungen in den beiden Konzepten pro Klasse verschiebt, die Klassen jedoch zu jedem Zeitpunkt nicht-linear fast perfekt voneinander trennbar sind. Dies ist in Abbildung 9.50 für drei verschiedene Zeitpunkte veranschaulicht. Es ist jedoch hervorzuheben, dass *QDA-AF* nicht generell unschlagbar bei einem gradual drift ist. Die Datensituation repräsentiert eine spezielle Struktur eines gradual drifts, sodass die Verteilungen beider Klassen zu jedem Zeitpunkt durch quadratische Diskriminanzfunktionen getrennt werden können.

Die Methode *LDA-AF* kann vergleichbar nicht so gut mit diesem gradual drift umgehen, da nur lineare Trennungen möglich sind. Bei den durchschnittlichen mittleren Prognosefehlern

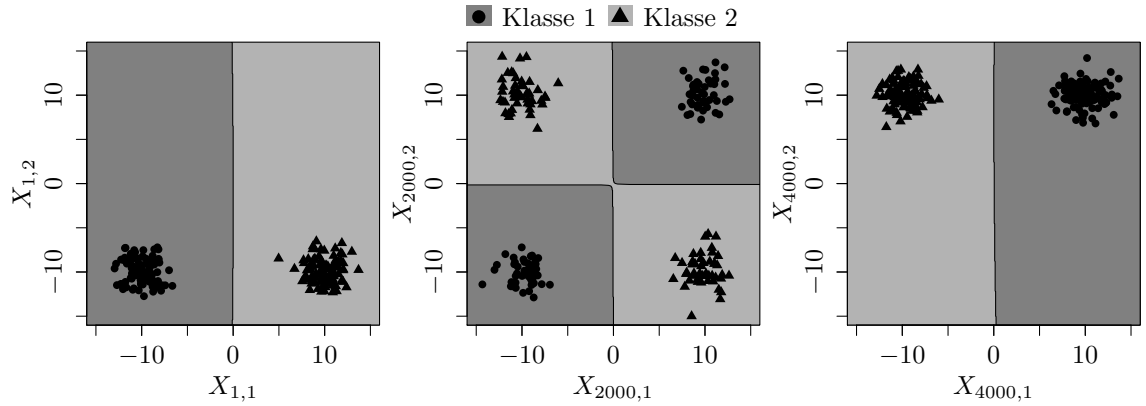


Abbildung 9.50: Veränderung der Klassifikationsgrenze der QDA im Laufe der Zeit durch Aktualisierung des Modells mit neuen Beobachtungen aus den Verteilungen der Klassen bei **Gradual Drift mit „Kreuzen“**.

bei zusätzlicher Betrachtung von Regressionsmodellen zur Modellierung des Trends und Prognose der Erwartungswertvektoren ist zu erkennen, dass der Fehler mit wachsendem Fenster $N_{\text{trend}} > 20$ für die Regressionsmodelle sogar leicht größer wird (vgl. Spalte „LDA-AF“). In Abbildung 9.52 (c) wird die grüne Kurve von den anderen überlagert.

Bei *ILDA* und *OLDC fix* bewirkt die Erweiterung der Methoden eine Verbesserung bezüglich der Prognosegüte. Am besten ist hier eine relativ kleine Fensterbreite von $N_{\text{trend}} = 10$ für die Anpassung der Regressionsmodelle. Für größere Werte von $N_{\text{trend}} > 10$ wird der Prognosefehler wieder größer, wobei jedoch bis zu $N_{\text{trend}} = 300$ der durchschnittliche mittlere Prognosefehler über die Zeit der ursprünglichen Methode nicht überschritten wird.

Bei *OLDC fix* sinkt der Prognosefehler mit steigender Lernrate λ. Dieser Einfluss wird durch die Erweiterung ausgeschaltet. Zwar ist für alle betrachteten N_{trend} immer $\lambda = 0.9$ am besten, allerdings sind die durchschnittlichen mittleren Prognosefehler für die anderen Lernraten nicht mehr deutlich höher als beim Vergleich der Ergebnisse der ursprünglichen Methode. Insgesamt sind die durchschnittlichen mittleren Prognosefehler bei Integration von Trendmodellen auf einem ähnlichen Niveau wie bei der ursprünglichen Methode mit $\lambda = 0.9$. Dies wird auch an den Verläufen der Prognosefehler in Abbildung 9.53 deutlich.

Wird die Methode *OLDC* mit adaptiver Lernrate zur Klassifikation herangezogen, so wird in dieser Datensituation wie bei *LDA-AF* die Prognosegüte durch die Erweiterung mit Regressionsmodellen etwas schlechter. Die durchschnittlichen mittleren Prognosefehler über die Zeit schwanken dabei unabhängig von der Wahl für N_{trend} zwischen 0.21 und 0.23. Dabei resultieren für $N_{\text{trend}} \leq 100$ bei festem λ_{start} die höchsten Prognosefehler wie bei den meisten anderen Datensituationen wieder bei Verwendung des großen Fensters $L = 50$ zur Adaption der Lernrate. Auch hier sollte dieses Fenster somit nicht zu groß gewählt werden. In Abbildung 9.51 wird deutlich, dass bei $L = 50$ die Lernrate λ für jeden Startwert über die Zeit stark schwankt und zum Ende des Datenstroms abfällt. Da sich die Verteilung jedoch stetig ändert, ist eine kleine Lernrate ungeeignet und die Adaption misslingt bei großem

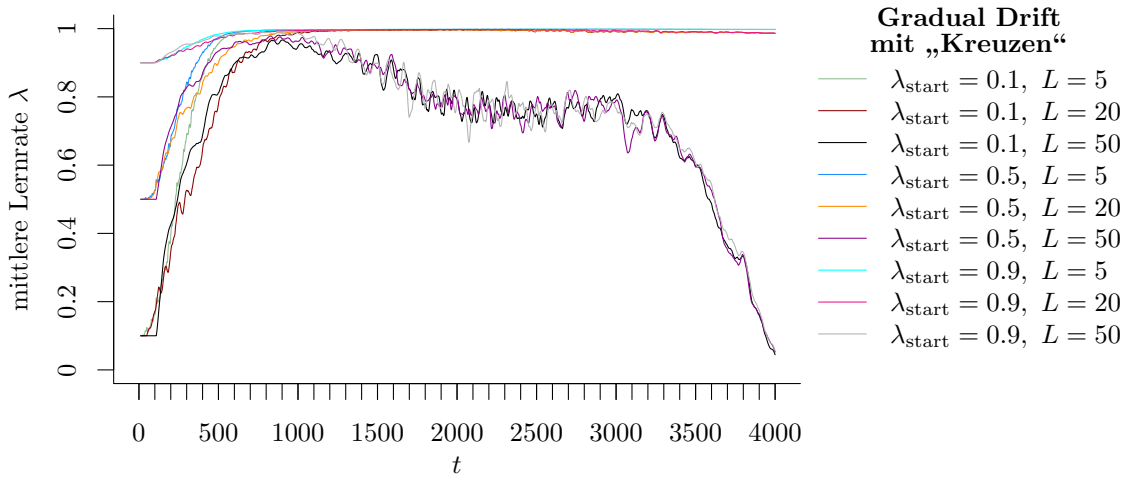
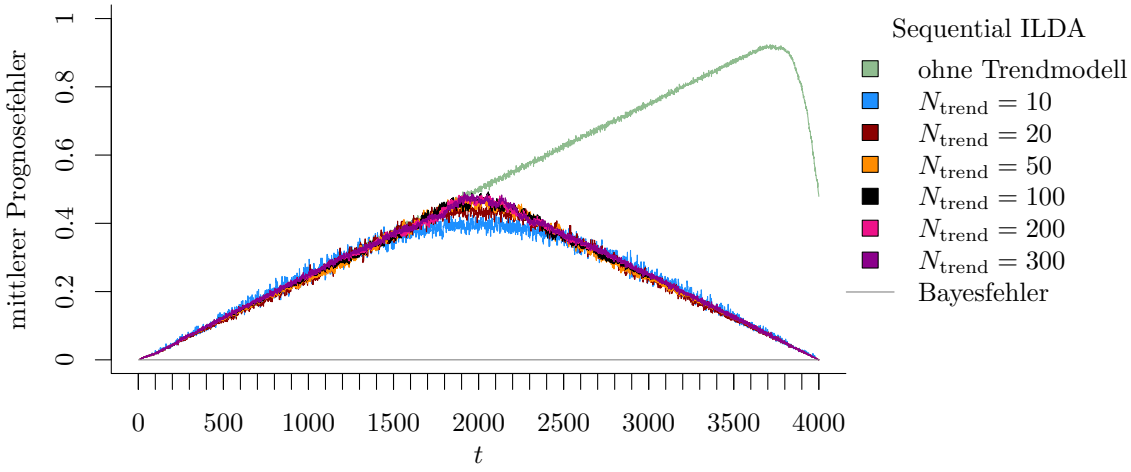


Abbildung 9.51: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **Gradual Drift mit „Kreuzen“**.

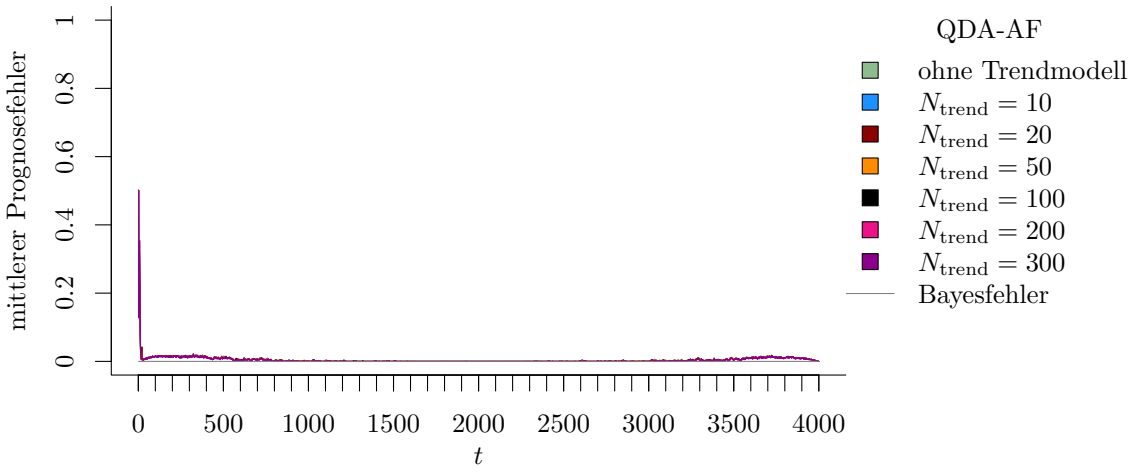
L. Die durchschnittlichen mittleren Prognosefehler über die Zeit der Methode *OLDC* mit adaptiver Lernrate ohne Erweiterung sind geringer als bei allen anderen Methoden (außer *QDA-AF*) und schwanken für die verschiedenen Kombinationen aus λ_{start} und L zwischen 0.2106 und 0.2182 (vgl. erste „Zeile“).

Zusammenfassung der wichtigsten Resultate:

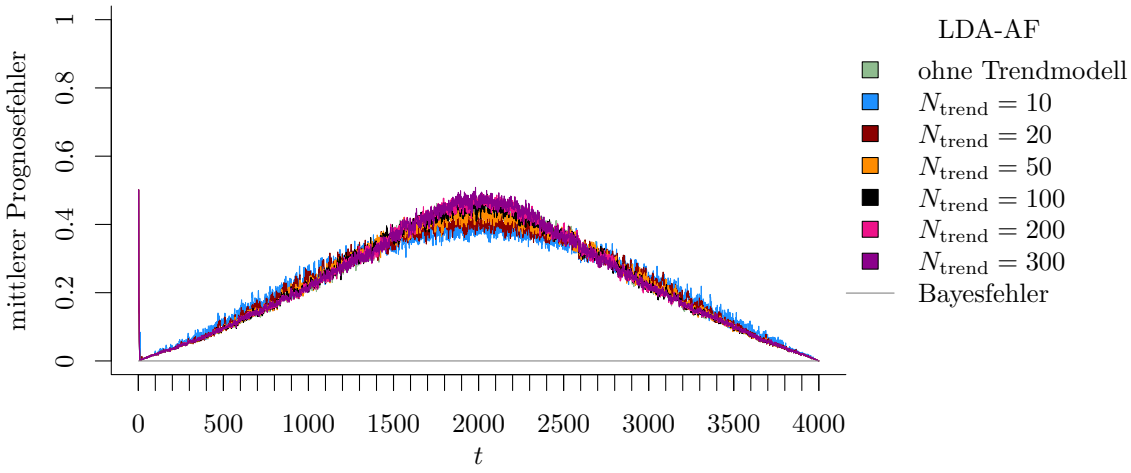
- Der durchschnittliche Bayesfehler über den gesamten Datenstrom für diese Datensituation beträgt < 0.0001 mit einer Standardabweichung von < 0.001.
- *QDA-AF* ist am besten zur Klassifikation auf dieser Datensituation.
- Die Erweiterung von *QDA-AF* führt nur noch zu einer leichten Verbesserung bezüglich der Prognosegüte.
- Bei *LDA-AF* steigt der durchschnittliche mittlere Prognosefehler über die Zeit leicht mit steigendem N_{trend} für $N_{\text{trend}} > 20$.
- Die Prognosegüte bei *ILDA* kann durch die Erweiterung direkt deutlich verbessert werden. Steigende Werte von $N_{\text{trend}} > 10$ vergrößern den Prognosefehler wieder, bis zum betrachteten Wert von $N_{\text{trend}} = 300$ bleiben die durchschnittlichen mittleren Prognosefehler über die Zeit jedoch unterhalb dem der ursprünglichen Methode.
- Bei *OLDC* mit fester Lernrate sinkt der durchschnittliche Prognosefehler mit steigender Fehlerrate λ. Durch Integration der Regressionsmodelle hat die Lernrate keinen deutlichen Einfluss mehr auf den Prognosefehler. Die Wahl von $\lambda = 0.9$ ist immer noch am besten, aber die durchschnittlichen Prognosefehler sind nun alle auf einem ähnlichen Niveau wie bei der ursprünglichen Methode mit $\lambda = 0.9$.
- *OLDC* mit adaptiver Lernrate führt hier zu den niedrigsten durchschnittlichen mittleren Prognosefehlern über die Zeit (abgesehen von *QDA-AF*), wobei der Startwert λ_{start} keinen deutlichen Einfluss hat und nach Erweiterung $L < 50$ gewählt werden sollte. Die Prognosegüte wird durch die Erweiterung jedoch leicht verschlechtert.



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

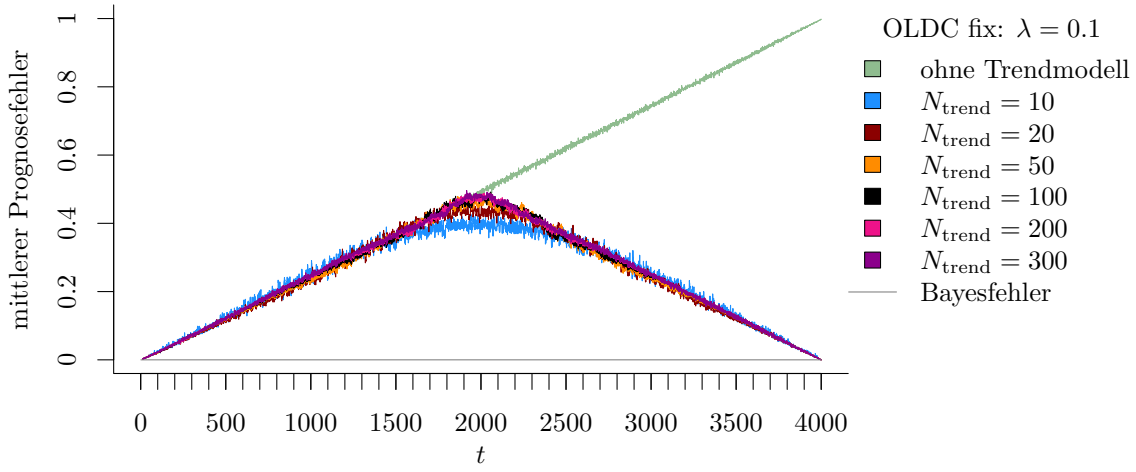


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

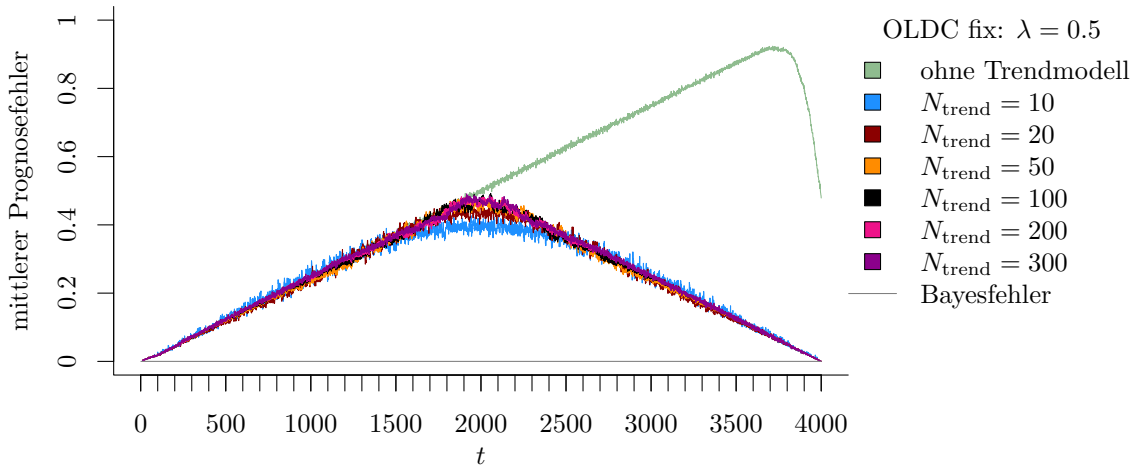


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

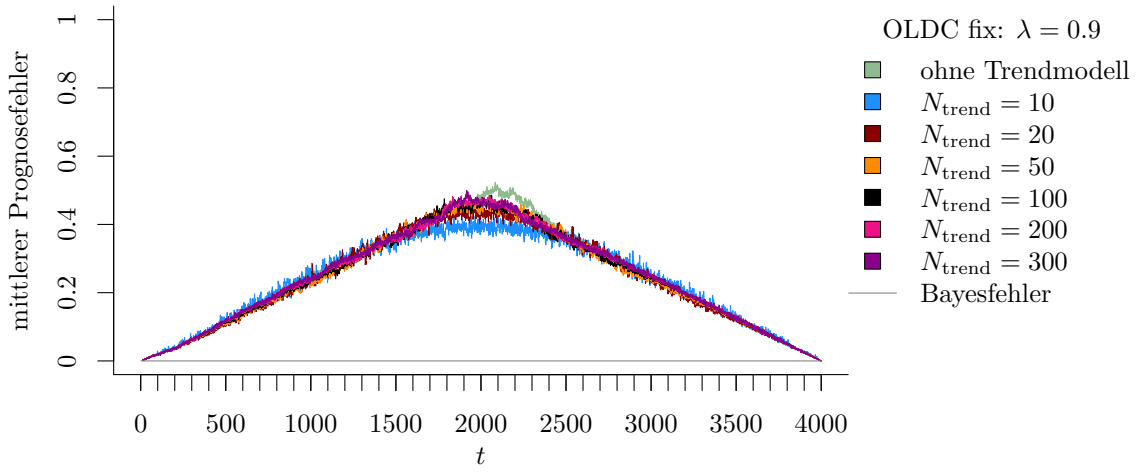
Abbildung 9.52: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Kreuzen“** im zweidimensionalen Raum.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.

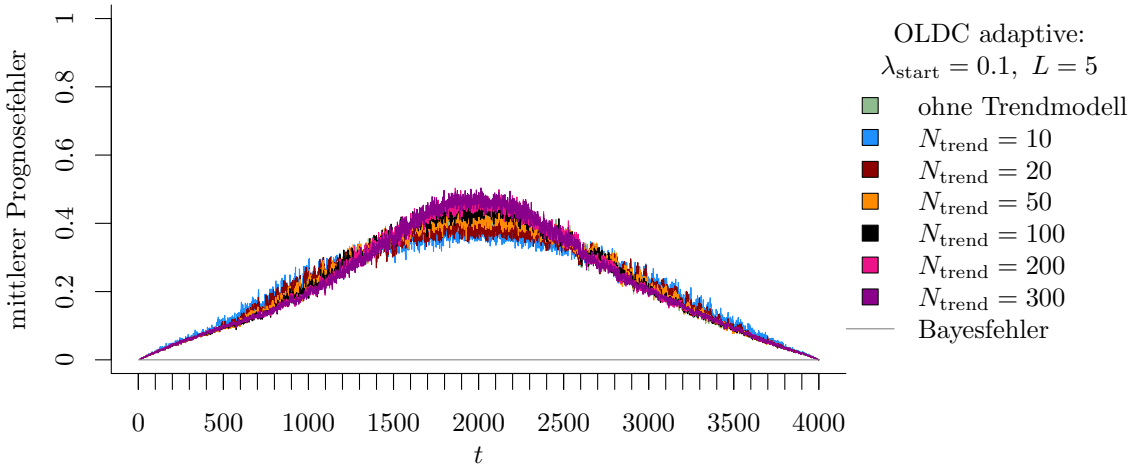


(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.

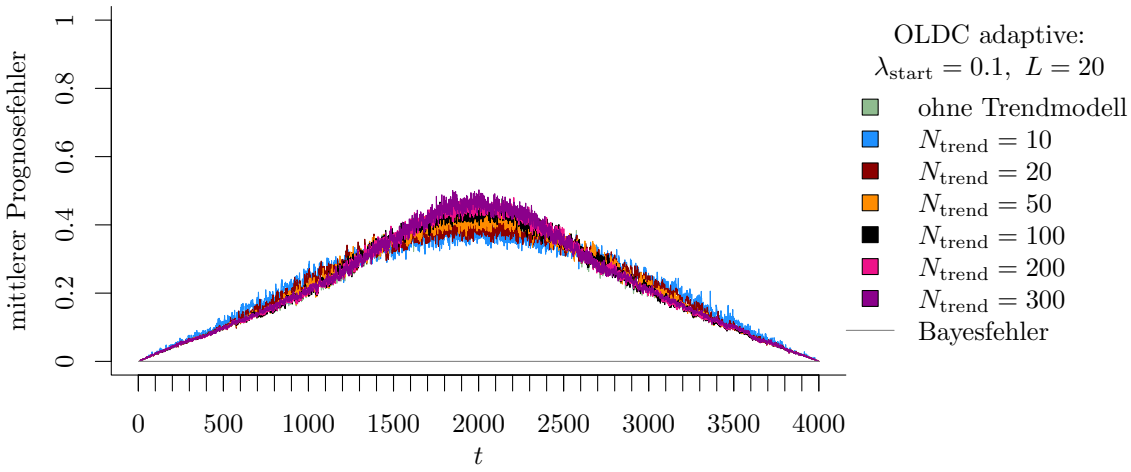


(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

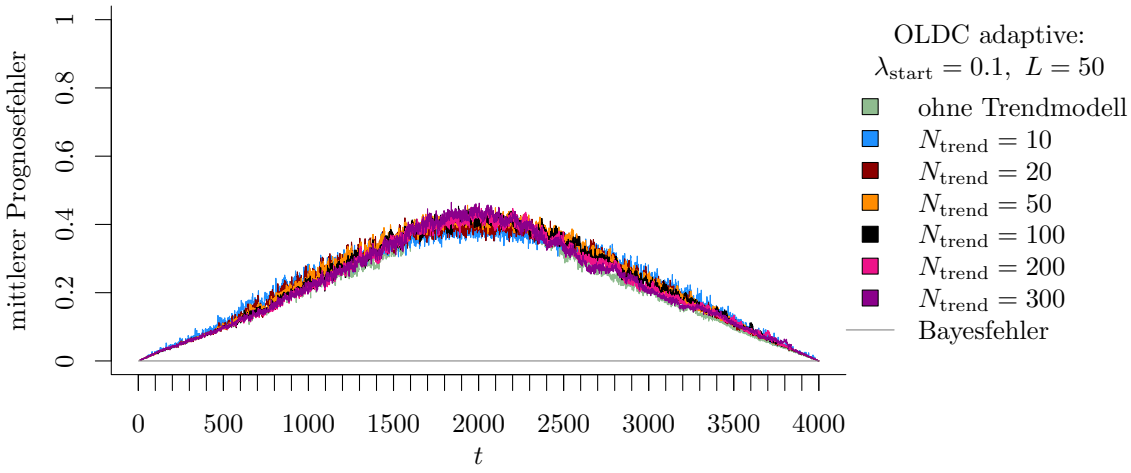
Abbildung 9.53: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Kreuzen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

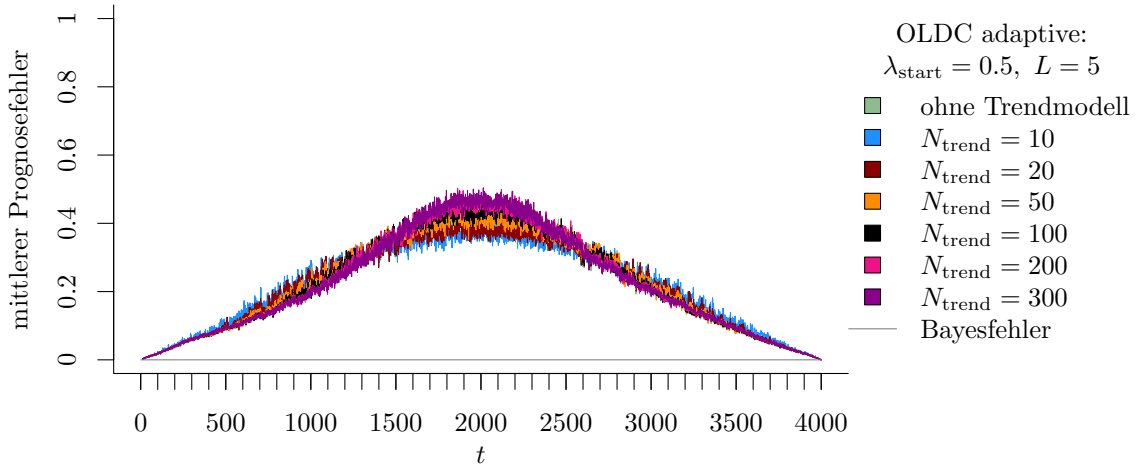


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

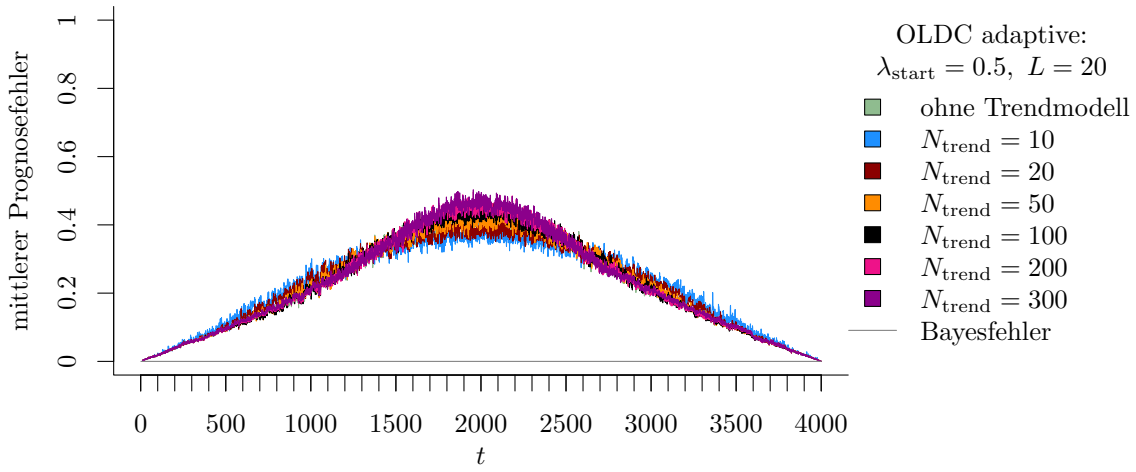


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

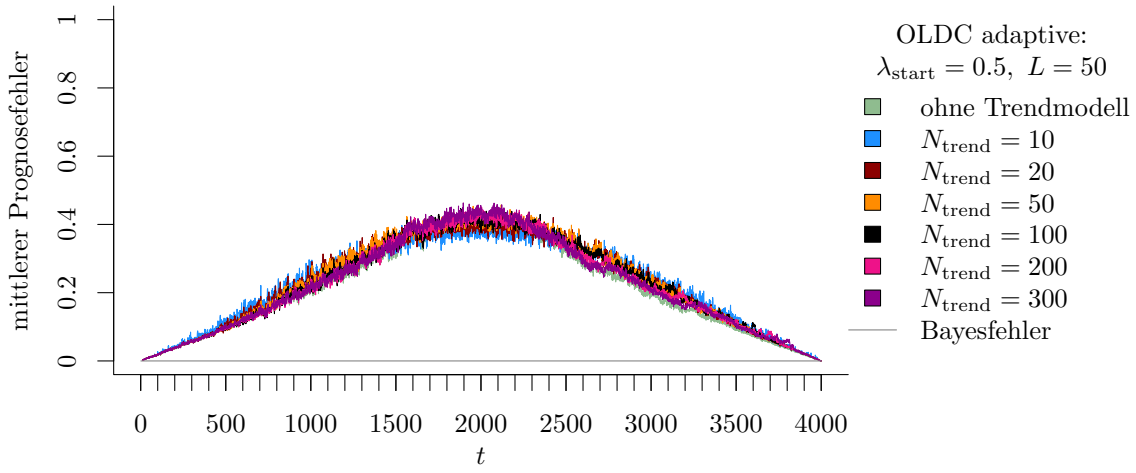
Abbildung 9.54: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Kreuzen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

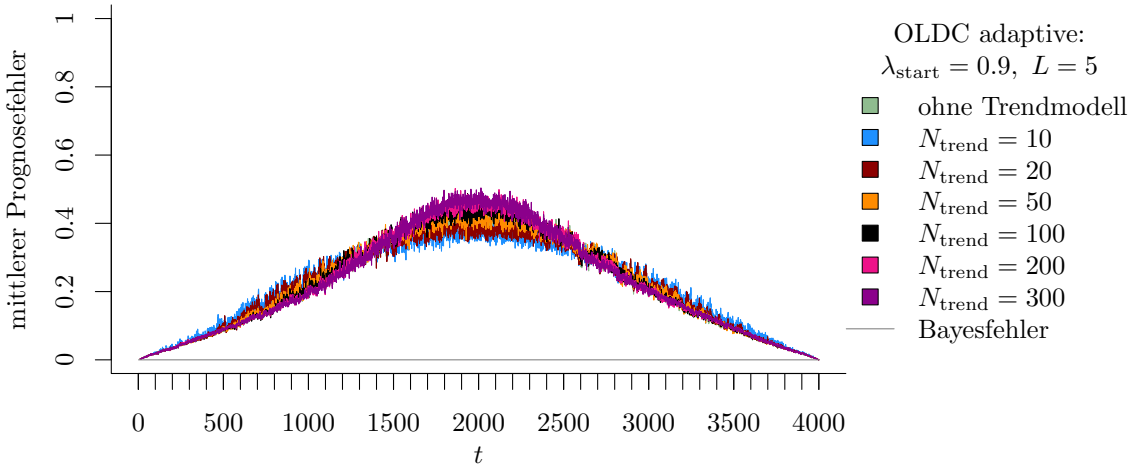


(b) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

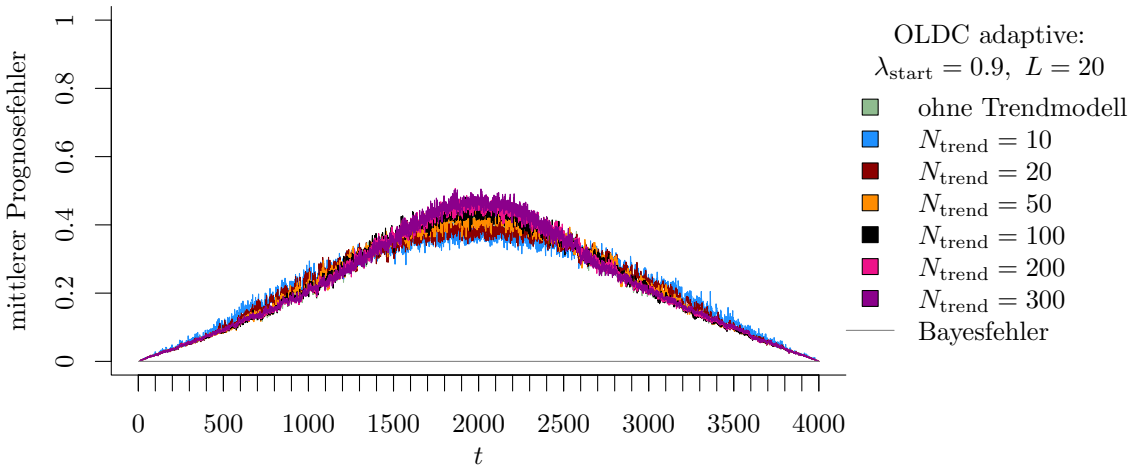


(c) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

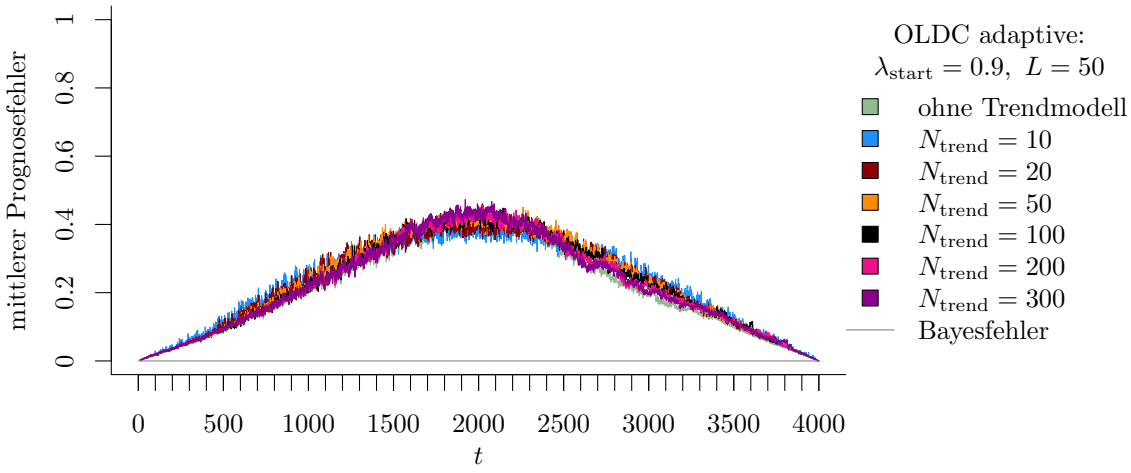
Abbildung 9.55: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Kreuzen“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.56: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Kreuzen“** im zweidimensionalen Raum.

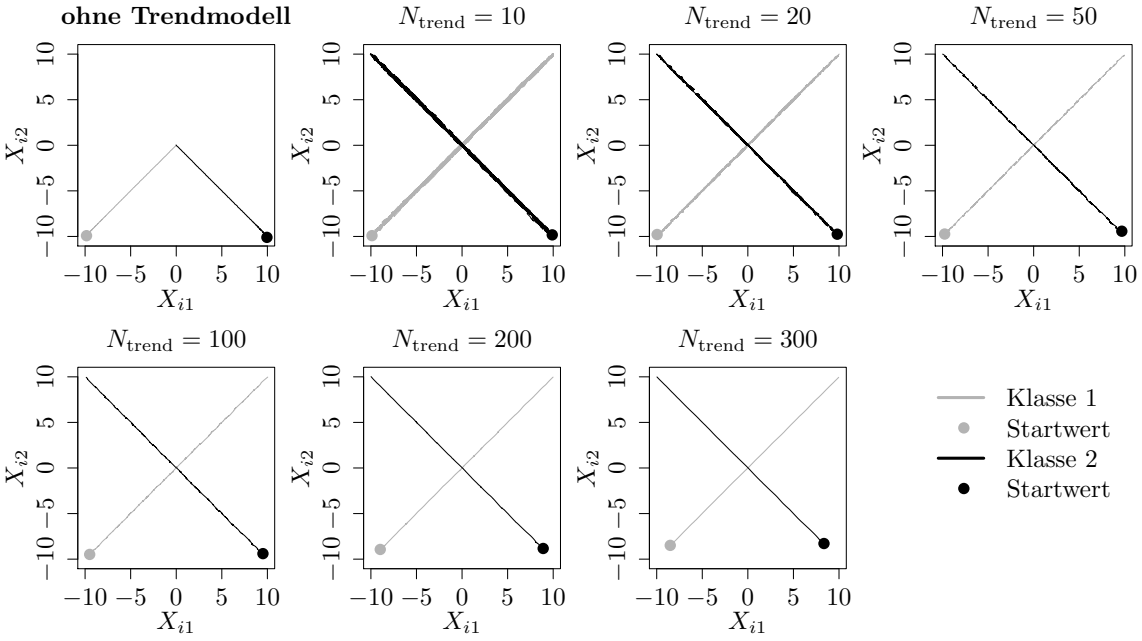


Abbildung 9.57: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren auf der Datensituation **Gradual Drift mit „Kreuzen“** für den gesamten Datenstrom. Ergebnisse aus *ILDA* und Erweiterung mit verschiedenen N_{trend}.

Für diese Datensituation kann die Erwartungstreue der Schätzer für die Erwartungswertvektoren nicht anhand der Betrachtung euklidischer Abstände analysiert werden, da keine eindeutigen „wahren“ Erwartungswertvektoren der Verteilungen zu jedem Zeitpunkt vorliegen, mit denen die Schätzer verglichen werden können. Vielmehr existieren lediglich zwei Konzepte für jede Klasse, zwischen denen ein langsamer Übergang erfolgt.

In Abbildung 9.57 sind erneut die Verläufe der herangezogenen Schätzer für die Erwartungswertvektoren beider Klassen bei der Methode *ILDA* für den gesamten Datensatz visualisiert. In der oberen linken Grafik wird die zeitliche Verzögerung der Schätzer der ursprünglichen Methode deutlich. Die Erwartungswertvektoren $(10, 10)^T$ (Klasse 1) bzw. $(-10, 10)^T$ (Klasse 2) des jeweils zweiten Konzepts werden zu keinem Zeitpunkt im Datenstrom durch die Mittelwertvektoren repräsentativ geschätzt. Dieses Verhalten wird durch die Erweiterung der Methode verbessert. Es ist zu erkennen, dass sowohl die Erwartungswertvektoren des jeweils ersten Konzepts zu Beginn des Datenstroms sowie jene des jeweils zweiten Konzepts zu Ende des Datenstroms repräsentativ geschätzt werden. In der Zwischenzeit läuft der prognostizierte Erwartungswertvektor aus Klasse 1 linear von $(-10, -10)^T$ nach $(10, 10)^T$ und nähert sich dem zweiten Erwartungswertvektor immer weiter an, je größer die Wahrscheinlichkeit für das zweite Konzept ist. Dieser Verlauf ist darauf zurückzuführen, dass vom Modell ein linearer Trend der Erwartungswertvektoren unterstellt wird. Für die prognostizierten Erwartungswertvektoren aus Klasse 2 passiert entsprechend dasselbe.

Tabelle 9.13: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittl. Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **Gradual Drift mit „Kreuzen“ ($p = 2$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4894 (0.005)	*0.0044* (0.000)	0.2272 (0.012)	0.1	0.4964 (0.002)	5	0.2182 (0.015)
						20	0.2170 (0.013)
						50	**0.2106** (0.010)
				0.3	0.5009 (0.002)		
				0.5	0.4894 (0.005)	5	0.2178 (0.015)
						20	0.2168 (0.013)
						50	0.2107 (0.010)
				0.7	0.2943 (0.004)		
				0.9	0.2492 (0.003)	5	0.2174 (0.015)
						20	0.2173 (0.014)
						50	0.2117 (0.011)
10	**0.2360** (0.013)	*0.0045* (0.000)	0.2272 (0.018)	0.1	0.2357 (0.013)	5	0.2187 (0.018)
						20	0.2219 (0.016)
						50	0.2279 (0.015)
				0.3	0.2360 (0.013)		
				0.5	0.2360 (0.013)	5	0.2185 (0.018)
						20	0.2214 (0.017)
						50	0.2276 (0.016)
				0.7	0.2358 (0.013)		
				0.9	**0.2344** (0.013)	5	0.2183 (0.018)
						20	0.2207 (0.017)
						50	0.2284 (0.016)
20	0.2386 (0.008)	*0.0046* (0.000)	**0.2267** (0.015)	0.1	0.2382 (0.008)	5	0.2181 (0.017)
						20	0.2206 (0.015)
						50	0.2270 (0.013)
				0.3	0.2387 (0.008)		
				0.5	0.2386 (0.008)	5	0.2179 (0.017)
						20	0.2201 (0.015)
						50	0.2265 (0.014)
				0.7	0.2382 (0.008)		
				0.9	0.2365 (0.009)	5	0.2176 (0.017)
						20	0.2195 (0.016)
						50	0.2269 (0.014)
50	0.2409 (0.005)	*0.0045* (0.000)	0.2273 (0.013)	0.1	0.2407 (0.004)	5	0.2211 (0.015)
						20	0.2205 (0.014)
						50	0.2308 (0.013)
				0.3	0.2411 (0.005)		
				0.5	0.2409 (0.005)	5	0.2208 (0.015)
						20	0.2201 (0.014)
						50	0.2294 (0.013)
				0.7	0.2403 (0.005)		
				0.9	0.2377 (0.006)	5	0.2204 (0.015)
						20	0.2207 (0.014)
						50	0.2290 (0.013)
100	0.2436 (0.004)	*0.0042* (0.000)	0.2288 (0.011)	0.1	0.2435 (0.003)	5	0.2224 (0.014)
						20	0.2218 (0.012)
						50	0.2261 (0.012)
				0.3	0.2440 (0.003)		
				0.5	0.2436 (0.004)	5	0.2220 (0.014)
						20	0.2219 (0.013)
						50	0.2246 (0.012)
				0.7	0.2429 (0.004)		
				0.9	0.2400 (0.004)	5	0.2216 (0.014)
						20	0.2226 (0.013)
						50	0.2250 (0.012)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.2455 (0.003)	***0.0040*** (0.000)	0.2308 (0.010)	0.1	0.2457 (0.003)	5	0.2239 (0.012)
						20	0.2244 (0.011)
						50	0.2211 (0.011)
				0.3	0.2459 (0.003)		
				0.5	0.2455 (0.003)	5	0.2235 (0.012)
						20	0.2243 (0.011)
						50	0.2213 (0.011)
				0.7	0.2449 (0.003)		
				0.9	0.2424 (0.004)	5	0.2230 (0.012)
						20	0.2251 (0.012)
						50	0.2217 (0.011)
300	0.2463 (0.003)	***0.0040*** (0.000)	0.2317 (0.009)	0.1	0.2467 (0.002)	5	0.2244 (0.011)
						20	0.2267 (0.010)
						50	0.2209 (0.010)
				0.3	0.2467 (0.003)		
				0.5	0.2463 (0.003)	5	0.2240 (0.011)
						20	0.2263 (0.010)
						50	0.2214 (0.010)
				0.7	0.2455 (0.003)		
				0.9	0.2433 (0.004)	5	0.2234 (0.011)
						20	0.2264 (0.011)
						50	0.2219 (0.010)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: < 0.0001 (Standardabw. < 0.001)

An den Verläufen ist zudem zu erkennen, dass zunächst die Varianz der Schätzer durch Erweiterung der Methode etwas vergrößert wird. Dies wird auch in Tabelle 9.13 deutlich und ist für alle Methoden gleich. Ab $N_{\text{trend}} = 50$ ist die durchschnittliche Varianz des Prognosefehlers über die Zeit jedoch jeweils wieder genauso gering wie bei der ursprünglichen Methode. Der Verlauf in Abbildung 9.57 wird wieder „glatt“.

Fazit: Diese spezielle Situation eines gradual drifts führt dazu, dass die Klassen zu jedem Zeitpunkt fast perfekt nicht-linear trennbar sind. Daher funktioniert die Methode *QDA-AF* bereits ohne Erweiterung durch Modellierung und Prognose eines Trends sehr gut. Durch die eingeführte Erweiterung ist es kaum noch möglich die Prognosegüte zu verbessern.

Alle anderen Methoden für eine Online Variante der Linearen Diskriminanzanalyse produzieren deutlich höhere Prognosefehler. Eine adaptive oder hohe feste Lernrate $\lambda = 0.9$ bei *OLDC* oder Gewichtung der Likelihood Terme bei *LDA-AF* kann den Verlauf des Prognosefehlers auch ohne Erweiterung der Methode etwas minimieren. Der Bayesfehler wird zu einem Großteil der Zeitpunkte jedoch bei Weitem nicht approximiert.

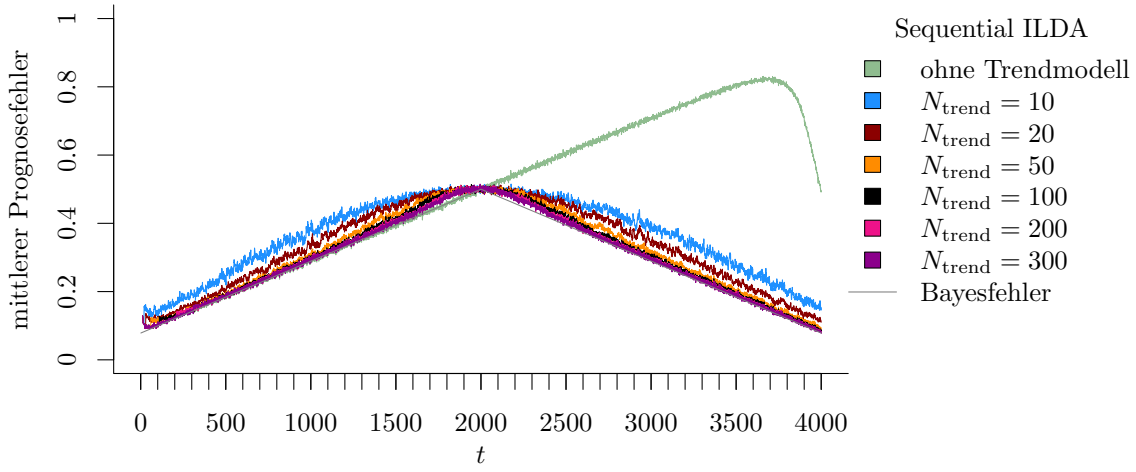
Trotz des gradual drifts anstelle eines linearen Trends bei einem incremental drift kann die Erweiterung der Methoden durch Integration lokaler linearer Regressionsmodelle eine Verbesserung der Prognosegüte für einige Methoden (außer *OLDC adaptive*) bewirken. Bei *ILDA* und *OLDC* mit fester Lernrate ist ein kleines Fenster von $N_{\text{trend}} = 10$ am besten. Für größere Fenster steigt der durchschnittliche Prognosefehler über die Zeit wieder an. Bis $N_{\text{trend}} = 300$ bleiben die Fehler jedoch unterhalb derer der ursprünglichen Methoden. Zudem wird der Einfluss der Lernrate λ bei *OLDC* größtenteils ausgeschaltet. Bei *LDA-AF* ist ein Fenster von $N_{\text{trend}} = 20$ optimal, bevor der Prognosefehler wieder steigt.

Datensituation Gradual Drift mit „Austausch“ ($p = 2$) In dieser Datensituation erfolgt ein „weicher“ Übergang von der multivariaten Normalverteilung mit Erwartungswertvektor $(-2, 0)^T$ zu jener mit Erwartungswertvektor $(2, 0)^T$ in Klasse 1 und in Klasse 2 umgekehrt (vgl. Seite 224). Anschaulich erfolgt demnach ein stetiger „Austausch“ der Verteilungen beider Klassen über die Zeit. Dies erklärt den Verlauf des Bayesfehlers, welcher in den Abbildungen 9.58–9.62 mit eingezeichnet ist. Zu Beginn und am Ende des betrachteten Datenstroms resultieren die Beobachtungen mit Wahrscheinlichkeit 1 aus dem ersten (Klasse 1) bzw. zweiten (Klasse 2) Konzept. Dadurch ergibt sich ein Bayesfehler von 0.0786 durch die Überschneidung der Verteilungen wie in der Datensituation „Kreisen“ (vgl. Seite 255 f.). Der Bayesfehler steigt bis zur Mitte des Datenstroms auf etwa 0.5 an, da zu diesem Zeitpunkt die Beobachtungen beider Klassen mit Wahrscheinlichkeit 0.5 aus $\mathcal{N}((-2, 0)^T, \mathbf{\Sigma})$ und mit Wahrscheinlichkeit 0.5 aus $\mathcal{N}((2, 0)^T, \mathbf{\Sigma})$ erzeugt werden. Die Kovarianzmatrix ist in beiden Klassen und zu allen Zeitpunkten identisch. Die Klassifikation kann demnach nur zufällig erfolgen. Durch weiteren Übergang zugunsten des jeweils zweiten Konzepts fällt danach der Bayesfehler wieder nach und nach ab.

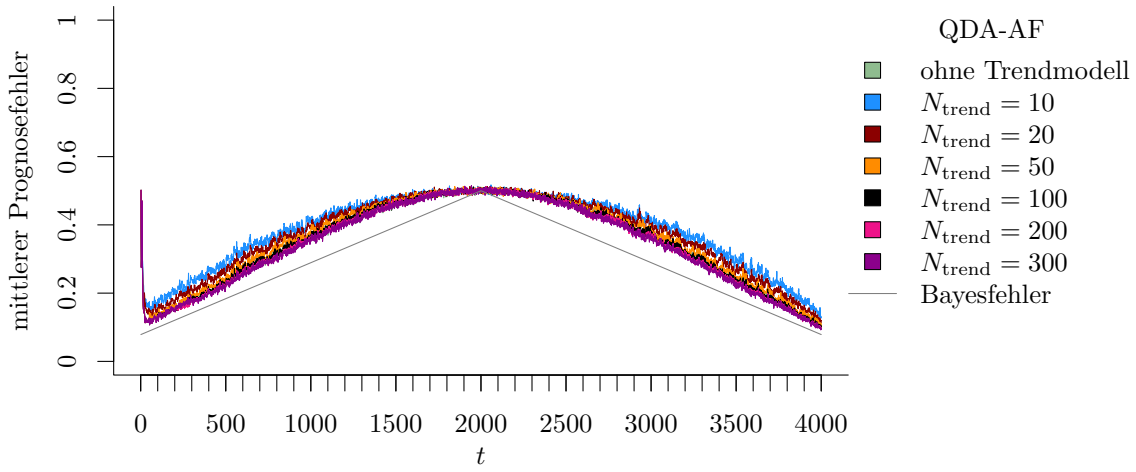
Für die Methoden ohne Anpassung an einen möglichen concept drift führt dies dazu, dass bei stetiger Aktualisierung des Modells durch neue Beobachtungen der Prognosefehler ähnlich wie in der vorherigen Situation immer weiter bis etwa 0.8 ansteigt und erst gegen Ende wieder auf 0.5 abfällt (vgl. grüne Kurven in Abbildungen 9.58 (a) und 9.59 (b)).

Durch die Einführung einer Lernrate bei *OLDC* zur stärkeren Gewichtung aktueller Beobachtungen bei den Updates kann die Prognosegüte bereits etwas verbessert werden (vgl. grüne Kurve in Abbildung 9.59 (c)). Für *OLDC* mit adaptiver Lernrate sehen die Kurven für die Kombinationen aus verschiedenen Startwerten λ_{start} und Fenster L qualitativ ähnlich aus (vgl. Abbildungen 9.60–9.62). Der Bayesfehler wird allerdings, insbesondere für Zeitpunkte $t \leq 2000$, nicht ganz so gut approximiert. Auch bei *QDA-AF* und *LDA-AF* (vgl. Abbildung 9.58 (b) und (c)) wird der Anstieg des Prognosefehlers minimiert, wobei die grüne Kurve hier durch jene der Erweiterungen überlagert wird.

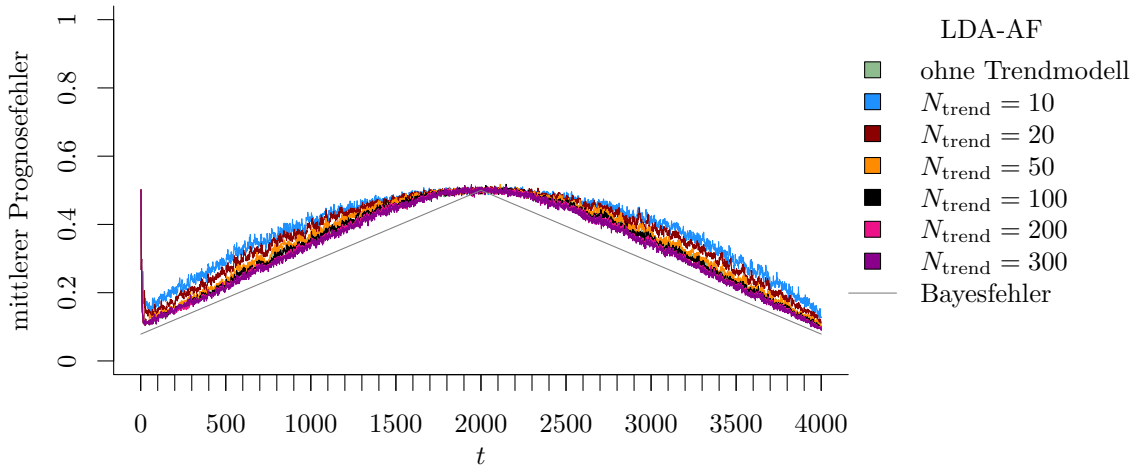
Generell ist zu erkennen, dass die Erweiterung der Methoden durch Einführung lokaler linearer Regressionsmodelle zur Modellierung des Trends und Prognose der Erwartungswertvektoren zu einer Verbesserung der Prognosegüte der Diskriminanzanalyse durch die jeweilige Methode führen kann. Die Kurven des Prognosefehlers nehmen einen „kurvenartigen“ Verlauf für den Datenstrom an und nähern sich dem Bayesfehler an. Insbesondere breite Fenster für die Regressionsmodelle, d. h. große Werte für N_{trend} scheinen besonders gut zu sein. Zu vielen Zeitpunkten im Datenstrom liegen bei diesen Werten die Kurven der Erweiterungen unterhalb jener der ursprünglichen Methoden. Es fällt jedoch auf, dass anders als bei den meisten Datensituationen hier die Erweiterung insbesondere bei *ILDA* und *OLDC* mit fester Lernrate gut funktioniert. Bei *OLDC* mit adaptiver Lernrate sowie *QDA-AF* und *LDA-AF* ist der „kurvenförmige“ Verlauf auch für große N_{trend} stärker ausgeprägt und der Bayesfehler wird weniger gut approximiert.



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

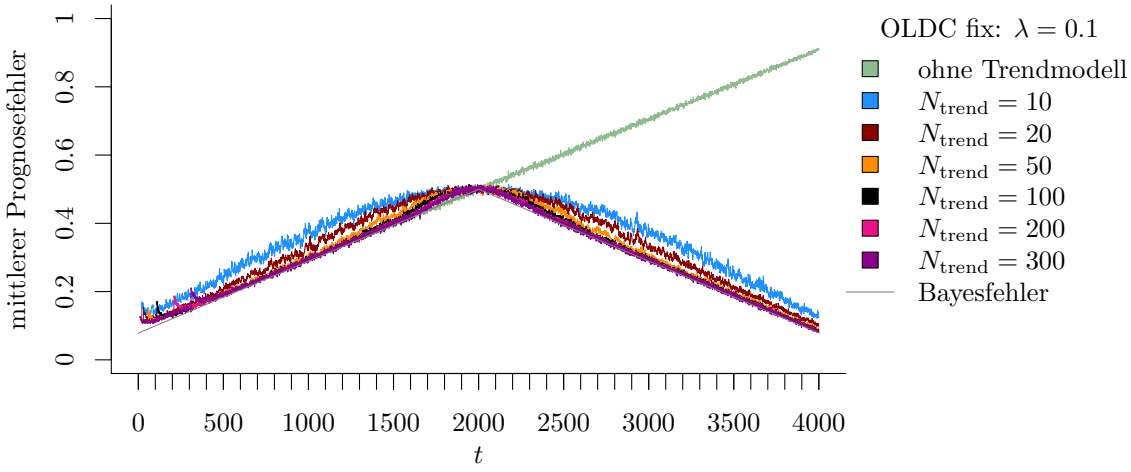


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

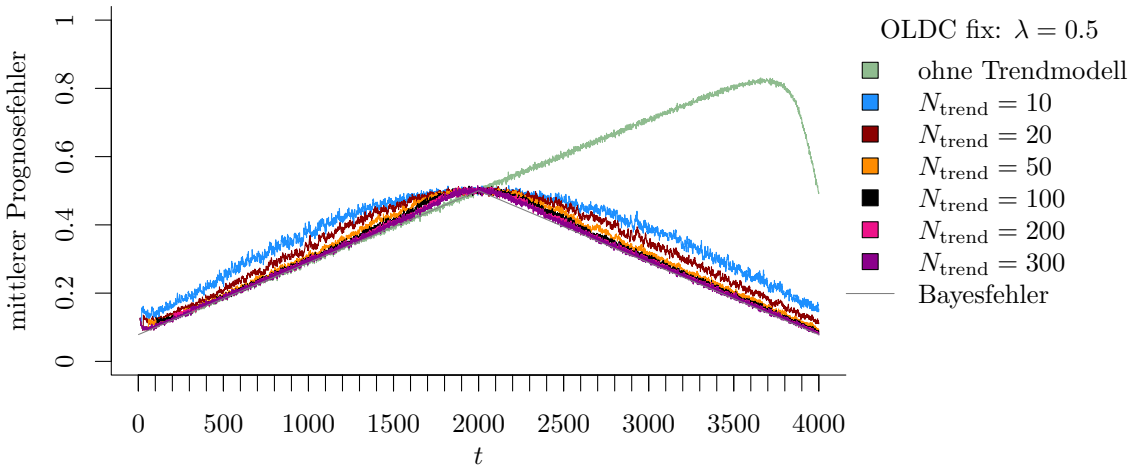


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

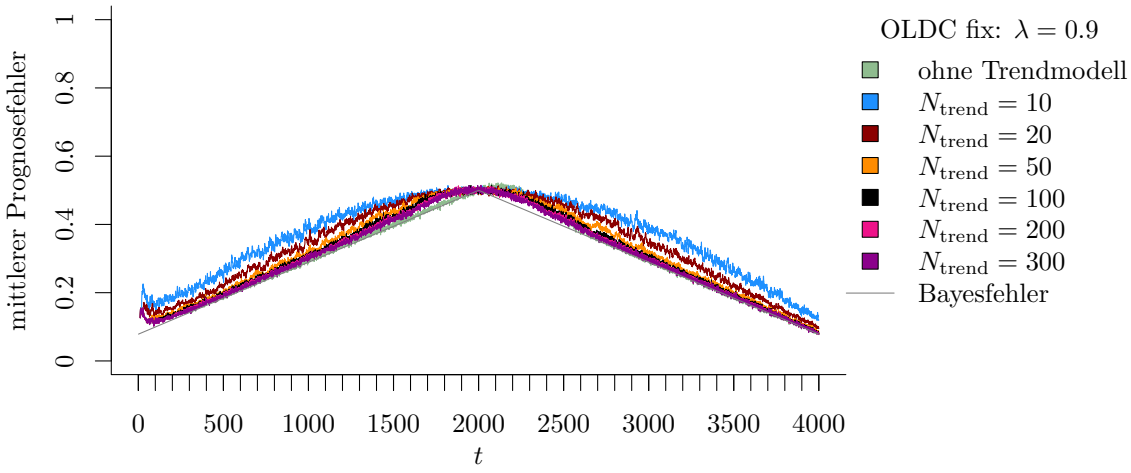
Abbildung 9.58: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Austausch“** im zweidimensionalen Raum.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.

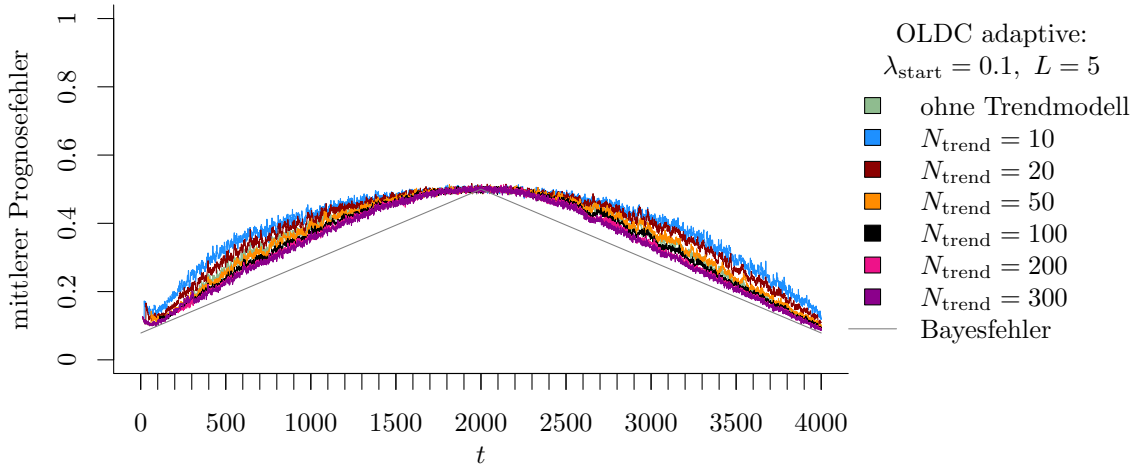


(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.

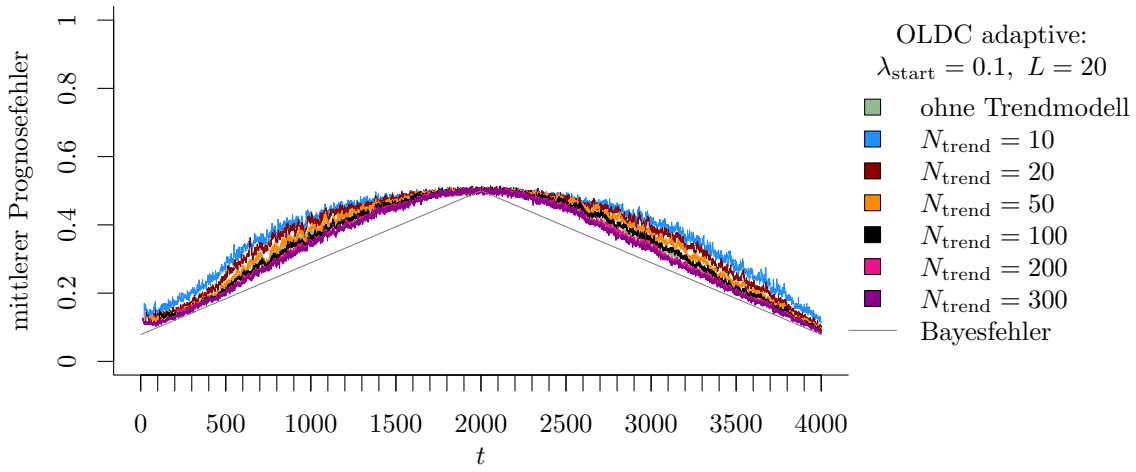


(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

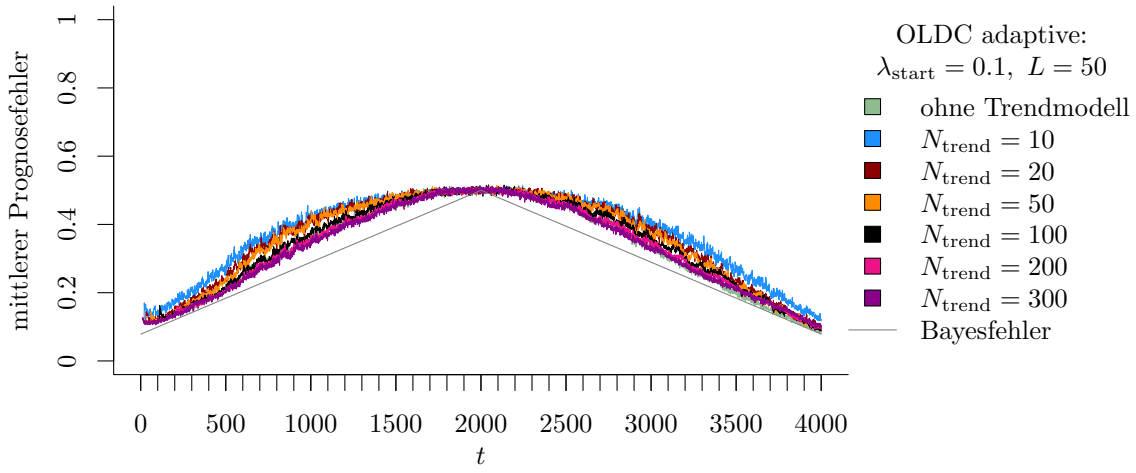
Abbildung 9.59: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Austausch“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

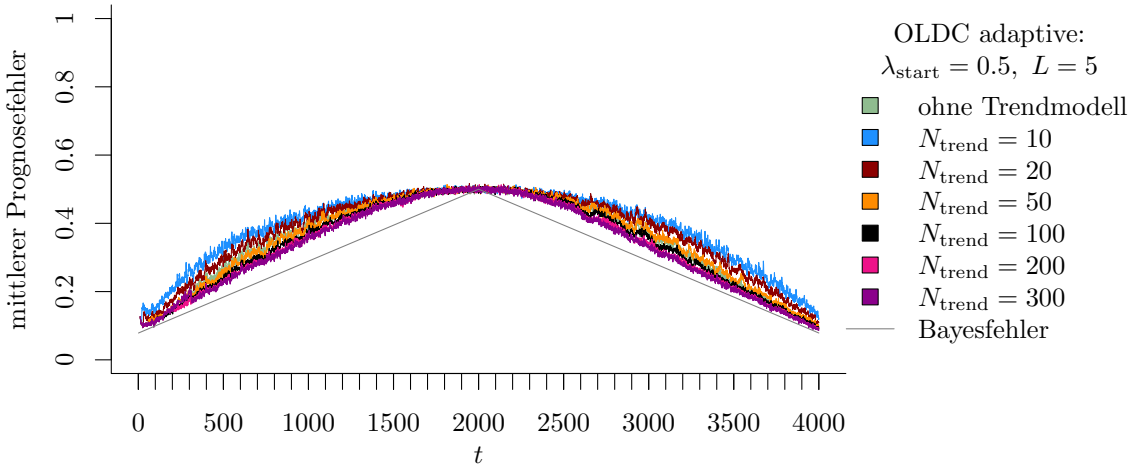


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

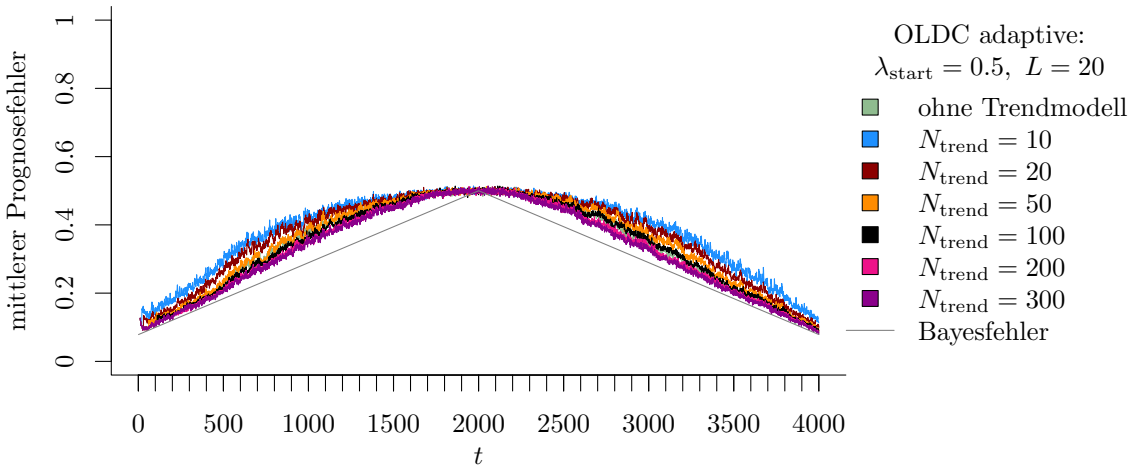


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

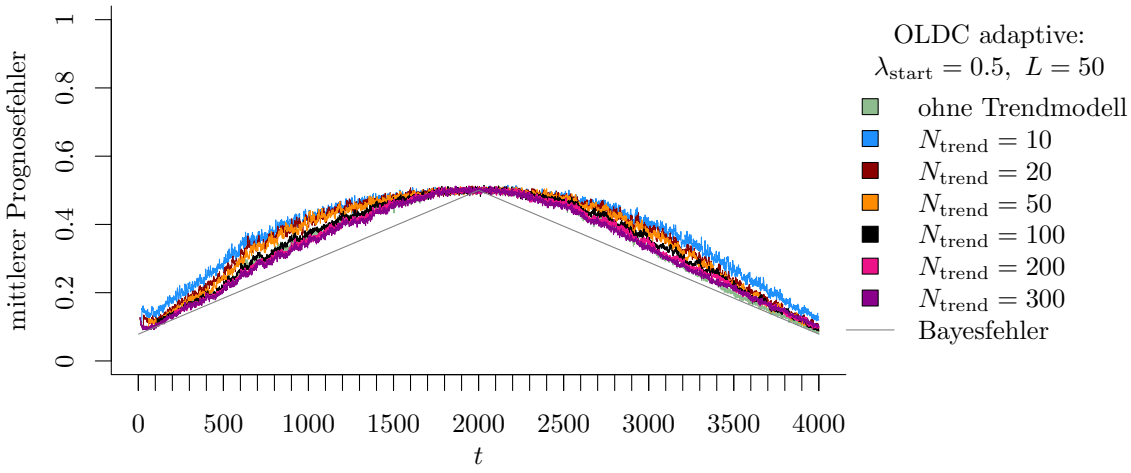
Abbildung 9.60: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Austausch“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

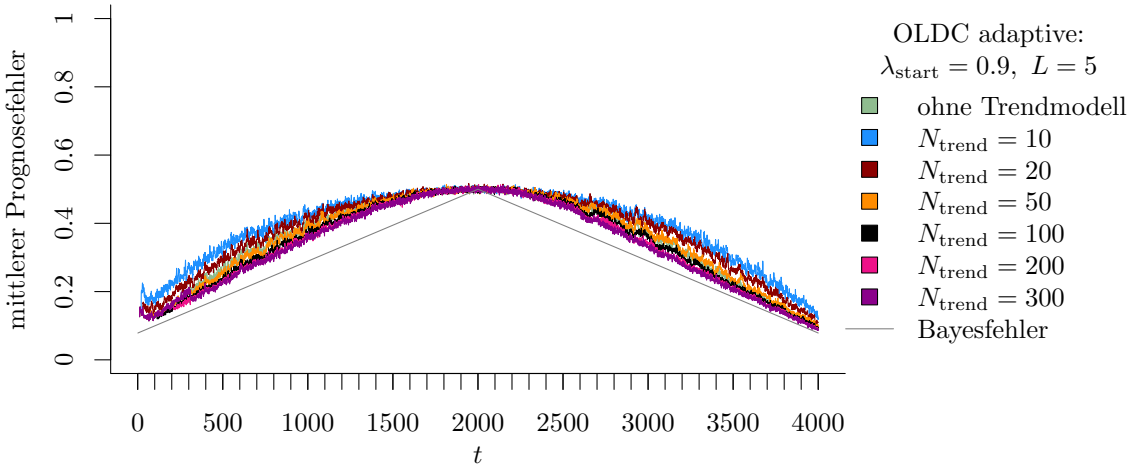


(b) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

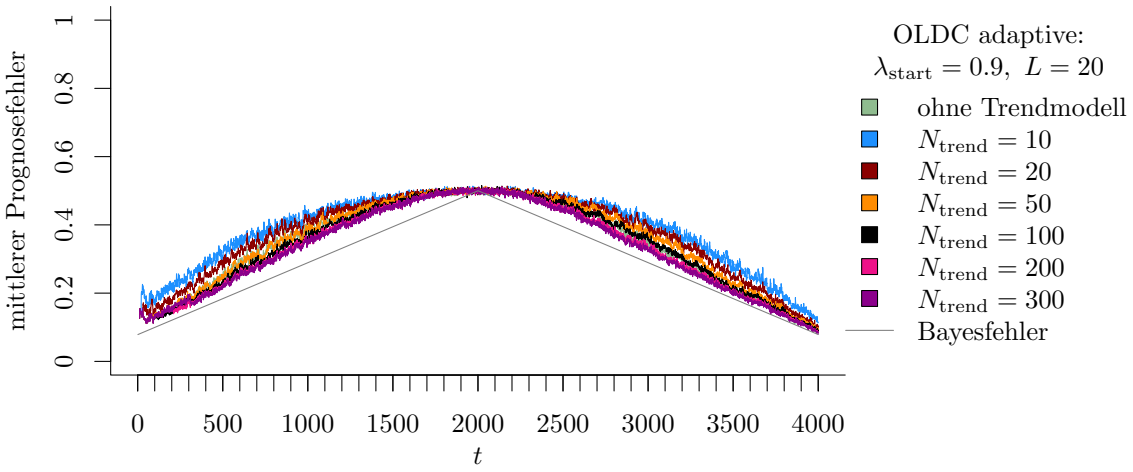


(c) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

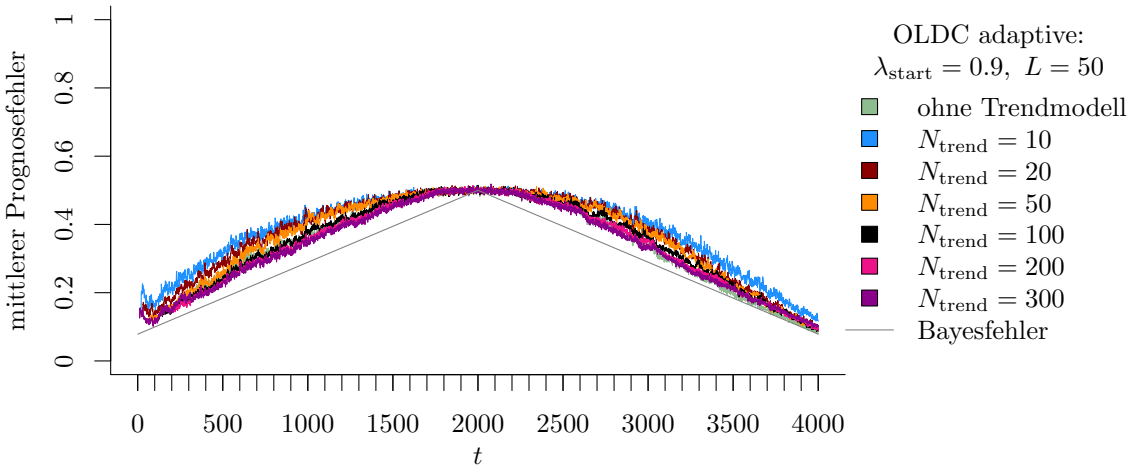
Abbildung 9.61: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Austausch“** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.62: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Gradual Drift mit „Austausch“** im zweidimensionalen Raum.

Die Resultate zeigen sich auch durch Konzentration auf ein einzelnes Maß der Prognosegüte und den Vergleich der durchschnittlichen mittleren Prognosefehler über die Zeit in Tabelle 9.14. Für alle erweiterten Methoden (Spalten) sinkt der durchschnittliche Prognosefehler mit steigendem N_{trend} und ist bei den betrachteten Werten minimal für $N_{\text{trend}} = 300$. Die Ergebnisse für *ILDA* und *OLDC fix* sind dabei für alle N_{trend} vergleichsweise etwas geringer als jene der anderen erweiterten Methoden.

Bei *QDA-AF* und *LDA-AF* erfolgt eine Verbesserung der Prognosegüte durch die Erweiterung erst bei genügend großem Wert für N_{trend}. Im Speziellen ab $N_{\text{trend}} = 200$ für *QDA-AF* und ab $N_{\text{trend}} = 100$ für *LDA-AF*. Bei *ILDA* und *OLDC fix* mit kleinen Lernraten $\lambda \in \{0.1, 0.3, 0.5\}$ wird der durchschnittliche Prognosefehler bereits durch Anpassung der Regressionsmodelle auf verhältnismäßig kleinen Fenstern N_{trend} verbessert. Bei großen Lernraten λ erfolgt eine Verbesserung der Prognosegüte erst bei verhältnismäßig größeren Fenstern N_{trend} für die Trendmodelle.

Bei *OLDC adaptive* fallen anhand des Verlaufs des Prognosefehlers in den Abbildungen 9.60–9.62 kaum Unterschiede für die verschiedenen Parameterkombinationen aus λ_{start} und L auf. Bei Konzentration auf den durchschnittlichen mittleren Prognosefehler über die Zeit sind diese jedoch ersichtlich (vgl. Spalte „OLDC adaptive“ in Tabelle 9.14). In der ursprünglichen Variante wird der durchschnittliche mittlere Prognosefehler über die Zeit durch ein größeres Fenster L zur Adaption der Fehlerrate geringer. Nach Erweiterung der Methode ist ab einem genügend großen Wert für N_{trend} von den betrachteten Werten ein mittleres Fenster $L = 20$ optimal. Es resultieren die geringsten durchschnittlichen Prognosefehler. Bei kleinem Fenster $L = 5$ erfolgt zusätzlich eine Verbesserung durch lineare Regressionsmodelle auf Fenstern der Breite $N_{\text{trend}} \geq 100$. Je größer L ist, desto größer muss auch N_{trend} sein, damit der durchschnittliche mittlere Prognosefehler über die Zeit kleiner als bei der ursprünglichen Methode ist (vgl. Spalte „OLDC adaptive“). Der Startwert λ_{start} ist hingegen weniger wichtig. Kleinere Startwerte $\lambda_{\text{start}} \in \{0.1, 0.5\}$ resultieren zwar in geringeren durchschnittlichen Prognosefehlern, die Unterschiede zu den Ergebnissen mit Startwert $\lambda_{\text{start}} = 0.9$ sind jedoch nicht so stark wie bei unterschiedlichen Fensterbreiten L. Dieses Muster lässt sich durch den Verlauf der adaptiven Lernrate erklären (vgl. Abbildung 9.63). Auch hier ist zu sehen, dass die Adaption relativ unabhängig vom Startwert λ_{start} ist. Die Kurven mit identischer Fenstergröße L zur Adaption liegen unabhängig vom Startwert übereinander. Während die Lernrate für Fenstergrößen $L \in \{5, 20\}$ relativ schnell (annähernd) den maximalen Wert annimmt und sich kaum noch verändert, ist der Verlauf für $L = 50$ sehr unruhig. Zum Ende des Datenstroms erfolgt zudem wieder eine Verringerung der Lernrate.

Während die Prognosegüte bei *OLDC* mit fester Lernrate ohne Erweiterung mit steigendem λ steigt (vgl. grüne Kurven in Abbildung 9.59), steigt sie nach der Erweiterung mit sinkendem λ (vgl. Spalte „OLDC fix“ in Tabelle 9.14), wobei sich die durchschnittlichen mittleren Prognosefehler über die Zeit jedoch nicht mehr allzu stark unterscheiden.

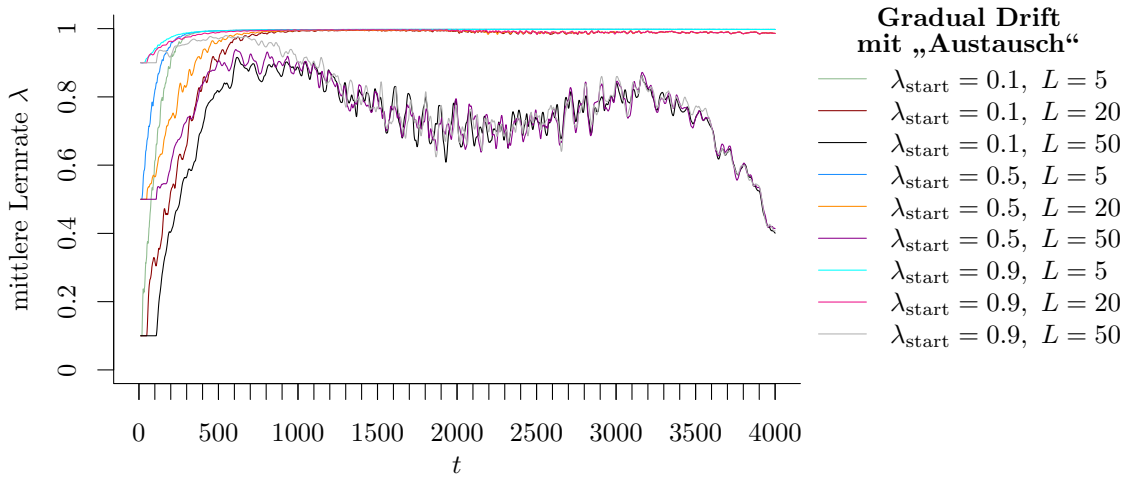


Abbildung 9.63: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **Gradual Drift mit „Austausch“**.

Die wichtigsten Ergebnisse bezüglich des Prognosefehlers für diese Datensituation sind die Folgenden:

- Der mittlere Bayesfehler über den gesamten Datenstrom für diese Datensituation beträgt 0.2893 mit einer Standardabweichung von 0.122.
- Die Erweiterung der Methoden führt tendenziell zu einer Verbesserung der Prognosegüte, wobei der durchschnittliche mittlere Prognosefehler über die Zeit mit steigendem N_{trend} immer weiter sinkt.
- Der minimale durchschnittliche Prognosefehler von 0.2958 wird durch *OLDC* mit $\lambda = 0.3$ und inkludierten Regressionsmodellen auf Fenstern der Breite $N_{\text{trend}} = 300$ erzielt. Dieser Fehler liegt nur etwas über dem durchschnittlichen Bayesfehler.
- Bei *OLDC fix* ohne Erweiterung ist eine höhere Lernrate besser, sobald die Methode erweitert wird, nimmt die Prognosegüte hingegen mit sinkendem λ leicht zu.
- Für *QDA-AF* wird erstmals der durchschnittliche mittlere Prognosefehler über die Zeit der ursprünglichen Methode bei $N_{\text{trend}} = 200$ unterboten, für *LDA-AF* bei $N_{\text{trend}} = 100$.
- *OLDC adaptive*: Bei kleinem Fenster $L = 5$ erfolgt eine Verbesserung der Prognosegüte durch lineare Regressionsmodelle auf Fenstern der Breite $N_{\text{trend}} \geq 100$. Je größer L ist, desto größer muss tendenziell auch N_{trend} sein, damit der durchschnittliche mittlere Prognosefehler über die Zeit kleiner als bei der ursprünglichen Methode ist. Bei genügend großem Fenster N_{trend} ist $L = 20$ am besten. Tendenziell sind kleinere Startwerte $\lambda_{\text{start}} \leq 0.5$ für die Lernrate besser, die Wahl des Fensters L hat jedoch einen größeren Einfluss auf die Prognosegüte.

Tabelle 9.14: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittl. Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **Gradual Drift mit „Austausch“ ($p = 2$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4874 (0.003)	0.3462 (0.004)	0.3397 (0.005)	0.1	0.5016 (0.002)	5	0.3511 (0.006)
						20	0.3361 (0.005)
						50	0.3293 (0.005)
				0.3	0.5013 (0.002)		
				0.5	0.4874 (0.003)	5	0.3518 (0.006)
						20	0.3382 (0.005)
						50	0.3302 (0.005)
				0.7	0.3311 (0.003)		
				0.9	*0.2985* (0.002)	5	0.3529 (0.006)
						20	0.3435 (0.005)
						50	0.3353 (0.005)
10	0.3619 (0.008)	0.3810 (0.008)	0.3809 (0.009)	0.1	*0.3534* (0.009)	5	0.3863 (0.010)
						20	0.3769 (0.009)
						50	0.3728 (0.009)
				0.3	0.3583 (0.009)		
				0.5	0.3619 (0.008)	5	0.3872 (0.010)
						20	0.3790 (0.009)
						50	0.3742 (0.009)
				0.7	0.3632 (0.008)		
				0.9	0.3639 (0.009)	5	0.3890 (0.010)
						20	0.3844 (0.010)
						50	0.3803 (0.010)
20	0.3331 (0.005)	0.3672 (0.006)	0.3643 (0.007)	0.1	*0.3260* (0.005)	5	0.3712 (0.008)
						20	0.3616 (0.007)
						50	0.3580 (0.007)
				0.3	0.3297 (0.005)		
				0.5	0.3331 (0.005)	5	0.3718 (0.008)
						20	0.3636 (0.007)
						50	0.3593 (0.007)
				0.7	0.3347 (0.005)		
				0.9	0.3367 (0.005)	5	0.3732 (0.008)
						20	0.3680 (0.008)
						50	0.3650 (0.008)
50	0.3122 (0.003)	0.3540 (0.005)	0.3477 (0.005)	0.1	*0.3076* (0.003)	5	0.3525 (0.005)
						20	0.3475 (0.005)
						50	0.3523 (0.006)
				0.3	0.3096 (0.003)		
				0.5	0.3122 (0.003)	5	0.3529 (0.005)
						20	0.3484 (0.005)
						50	0.3534 (0.006)
				0.7	0.3139 (0.003)		
				0.9	0.3171 (0.003)	5	0.3537 (0.006)
						20	0.3512 (0.006)
						50	0.3562 (0.006)
100	0.3035 (0.002)	0.3463 (0.004)	0.3382 (0.004)	0.1	*0.3009* (0.002)	5	0.3404 (0.004)
						20	0.3367 (0.004)
						50	0.3399 (0.005)
				0.3	0.3016 (0.002)		
				0.5	0.3035 (0.002)	5	0.3407 (0.004)
						20	0.3374 (0.004)
						50	0.3402 (0.005)
				0.7	0.3052 (0.002)		
				0.9	0.3088 (0.003)	5	0.3416 (0.004)
						20	0.3402 (0.004)
						50	0.3427 (0.005)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	0.2985 (0.002)	0.3411 (0.004)	0.3317 (0.004)	0.1	0.2977 (0.002)	5	0.3327 (0.003)
						20	0.3268 (0.003)
						50	0.3318 (0.004)
				0.3	*0.2972* (0.002)		
				0.5	0.2985 (0.002)	5	0.3331 (0.003)
						20	0.3275 (0.003)
						50	0.3320 (0.004)
				0.7	0.3001 (0.002)		
				0.9	0.3037 (0.002)	5	0.3341 (0.003)
						20	0.3309 (0.004)
						50	0.3348 (0.005)
300	**0.2968** (0.002)	**0.3393** (0.004)	**0.3293** (0.004)	0.1	0.2970 (0.002)	5	0.3303 (0.003)
						20	**0.3227** (0.003)
						50	0.3272 (0.004)
				0.3	***0.2958*** (0.002)		
				0.5	0.2968 (0.002)	5	0.3309 (0.003)
						20	0.3237 (0.003)
						50	0.3273 (0.004)
				0.7	0.2983 (0.002)		
				0.9	0.3017 (0.002)	5	0.3320 (0.003)
						20	0.3276 (0.003)
						50	0.3309 (0.004)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.2893 (Standardabweichung 0.122)

Die Grafiken in Abbildung 9.64 lassen eine ähnliche Schlussfolgerung zu wie in der vorherigen Datensituation mit Gradual Drift mit „Kreuzen“. Bei der ursprünglichen Methode (hier *ILDA*) hängen die Mittelwertvektoren als Schätzer für die Erwartungswertvektoren zeitlich hinterher, sodass beim gradual drift die Erwartungswertvektoren $(2,0)^T$ (Klasse 1) bzw. $(-2,0)^T$ (Klasse 2) des jeweils zweiten Konzepts im Laufe des gesamten Datenstroms nicht repräsentativ geschätzt werden (vgl. linke obere Grafik). Die zeitliche Verzögerung wird durch Einführung der lokalen linearen Regressionsmodelle zur Modellierung des Trends der Erwartungswerte und Prognose der zukünftigen Erwartungswertvektoren korrigiert. Auch hier bewegen sich die prognostizierten Erwartungswertvektoren im Laufe des Datenstroms „linear“ vom Erwartungswertvektor des jeweils ersten Konzepts zu dem des jeweils zweiten Konzepts, da ein linearer Trend unterstellt wird. Werden kleine Fenster N_{trend} für die lokalen linearen Regressionsmodelle betrachtet, sind die Modelle noch recht instabil, was anhand der unruhigen Verläufe deutlich wird. Für größere Fenster N_{trend} wird der Verlauf zunehmend ruhiger.

In Tabelle 9.14 wird deutlich, dass ebenfalls für die meisten Methoden die durchschnittliche Varianz des Prognosefehlers durch Erweiterung der Methoden zunächst gegenüber jener des Prognosefehlers der ursprünglichen Methode etwas erhöht wird. In den meisten Fällen wird ab einer Fensterbreite von $N_{\text{trend}} = 50$ für die Anpassung der lokalen linearen Regressionsmodelle wieder das ursprüngliche Niveau der Varianz erreicht und für wachsendes N_{trend} kann diese weiter sinken.

Fazit: Anders als in der vorherigen Datensituation Gradual Drift mit „Kreuzen“ (vgl. Seite 316 ff.) ist hier nicht die Methode *QDA-AF* allen anderen überlegen. Dies liegt daran, dass aufgrund des stetigen „Austausches“ der Verteilungen der Klassen auch keine

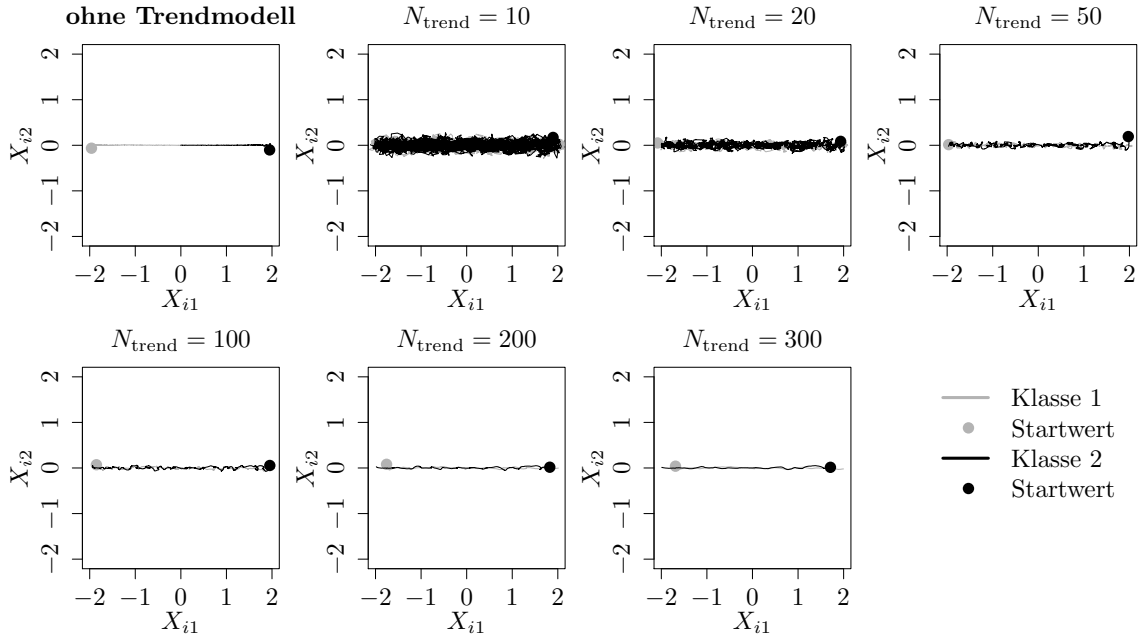


Abbildung 9.64: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren auf der Datensituation **Gradual Drift mit „Austausch“** für den gesamten Datenstrom. Ergebnisse aus *ILDA* und Erweiterung mit verschiedenen N_{trend}.

nicht-linearen Trenngeraden die Verteilungen voneinander trennen können. Obwohl die Erweiterung der Methoden nicht für gradual drifts entwickelt wurde, zeigt sich jedoch auch hier, dass vielfach trotzdem eine Verbesserung der Prognosegüte erfolgen kann.

Der Prognosefehler sinkt bei allen erweiterten Methoden mit steigender Fensterbreite N_{trend} für die lokalen linearen Regressionsmodelle und nähert sich dem Bayesfehler an. Zudem hängt der Prognosefehler nicht mehr so stark von der Lernrate λ bei *OLDC* ab. Generell resultieren in dieser Situation durch die Erweiterung von *ILDA* und *OLDC* mit fester Lernrate die geringsten durchschnittlichen Prognosefehler über die Zeit.

Datensituation Sudden Drift ($p = 2$) Alle betrachteten Update-Methoden für Diskriminanzanalyse und auch die Erweiterungen wurden nicht für den Umgang mit sudden drifts bzw. Strukturbrüchen entwickelt. Trotzdem sind auch solche Datensituationen interessant, da auch sie eine Form von concept drift darstellen und in der Praxis die Form der Veränderung nicht unbedingt bekannt ist. Zudem ist der Übergang von einem incremental drift zu einem sudden drift fließend, da sich nur die Stärke bzw. das Ausmaß des Drifts unterscheidet und nicht klar definiert ist, ab welcher Stärke ein sudden drift „beginnt“. Daher wird auch das Verhalten der Methoden und ihrer Erweiterungen auf einer Datensituation mit sudden drifts untersucht.

Dazu werden drei plötzliche Strukturbrüche der Verteilungen betrachtet (vgl. Seite 224). Der Bayesfehler beträgt trotzdem konstant 0.0786, da die Abstände der Verteilungen beider Klassen durch gleichmäßige Rotation der Erwartungswertvektoren auf dem Kreis gleich bleiben und sich die Überlappung der Verteilungen durch identische und symmetrische Kovarianzmatrizen nicht ändert. Dieser Bayesfehler kann durch keine der Update-Methoden zu jedem Zeitpunkt im Datenstrom approximiert werden. Bei allen Methoden sind die Strukturbrüche durch den Prognosefehler erkennbar (vgl. Abbildungen 9.66–9.70). Ohne Lernrate bzw. Anpassung der LDA an einen concept drift bleibt der Prognosefehler zudem nach einem starken Strukturbruch lange auf dem gleich hohen Niveau des Anstiegs, bevor er wieder abflacht (vgl. grüne Kurven in Abbildungen 9.66 (a) und 9.67 (b)).

Durch die Methoden für den Umgang mit concept drift kann diese Verzögerung vermindert werden, was daran zu sehen ist, dass die grünen Kurven in Abbildungen 9.66 (b) und (c), 9.67 (c) und 9.68–9.70 nach dem starken Anstieg schneller wieder auf das ursprüngliche Niveau vor dem Strukturbruch abfallen. In den Grafiken zu *QDA-AF* und *LDA-AF* (vgl. Abbildung 9.66 (b) und (c)) ist bereits zu sehen, dass die Methoden relativ gut auf der Datensituation funktionieren. Die grüne Kurve wird jeweils von den Kurven der erweiterten Methode überlagert. Dieses Resultat bestätigt sich auch durch die Konzentration auf den durchschnittlichen mittleren Prognosefehler über die Zeit in Tabelle 9.15. Die Prognosefehler von 0.1124 (*QDA-AF*) bzw. 0.1090 (*LDA-AF*) sind im Vergleich aller Methoden bereits gering und können durch die Erweiterung der Methoden in dieser Datensituation auch nicht mehr verringert werden. Mit steigendem N_{trend} sinkt der durchschnittliche mittlere Prognosefehler über die Zeit zwar, der Wert der Ausgangsmethode wird jedoch für beide Methoden nicht unterschritten. Dies lässt den Schluss zu, dass *QDA-AF* und *LDA-AF* verhältnismäßig gut auch mit manchen sudden drifts umgehen können.

Ein ähnliches Fazit lässt sich für *OLDC* mit adaptiver Lernrate schließen. Die Prognosefehler der ursprünglichen Methode sind bereits sehr gering (vgl. Spalte „OLDC adaptive“). Durch $\lambda_{\text{start}} = 0.1$ und $L = 20$ resultiert sogar der global minimale durchschnittliche mittlere Prognosefehler über die Zeit von 0.0938. Für vereinzelte Parameterkombinationen mit $N_{\text{trend}} = 100$ und $L = 5$ können die Prognosefehler der entsprechenden ursprünglichen Methode etwas verringert werden. Generell können die Resultate durch die Erweiterung für die meisten Parameterkombinationen in dieser Datensituation plötzlicher Drifts jedoch

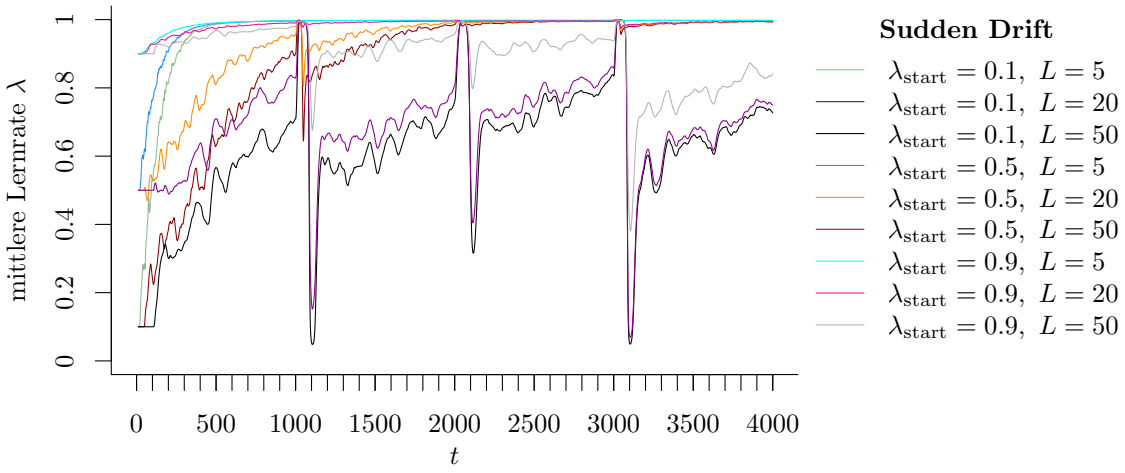


Abbildung 9.65: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **Sudden Drift**.

nicht verbessert werden. Die Ergebnisse werden jedoch durch die Erweiterung auch nicht auffallend stark verschlechtert. Zudem schwanken die Fehler pro N_{trend} nicht allzu stark.

Abbildung 9.65 verdeutlicht, dass die adaptive Lernrate auf die Strukturbrüche reagiert. Abhängig von der Wahl von λ_{start} und L verkleinert sich die Lernrate nach einem Strukturbruch unterschiedlich stark. Dies wirkt sich insbesondere bei der erweiterten Methode unterschiedlich auf die Prognosefehler aus. In den Abbildungen 9.68–9.70 ist zu sehen, dass ein großes Fenster $L = 50$ bei Modellierung des Trends auf Fenstern verschiedener Größe (insbesondere großem N_{trend}) zu unterschiedlich starken Ausschlägen nach den Strukturbrüchen führt. Die Kurven des Prognosefehlers für kleinere Fenster L zur Adaption der Lernrate liegen zu den meisten Zeiten vergleichsweise höher. Die Kurven nähern sich mit steigendem L zu Zeitpunkten stabiler Verteilungen dem Bayesfehler an. Aufgrund der Ausschläge bei $L = 50$ ist bei Konzentration auf einen einzelnen Wert der durchschnittliche mittlere Prognosefehler über die Zeit dennoch für $N_{\text{trend}} \geq 50$ für $L = 20$ minimal.

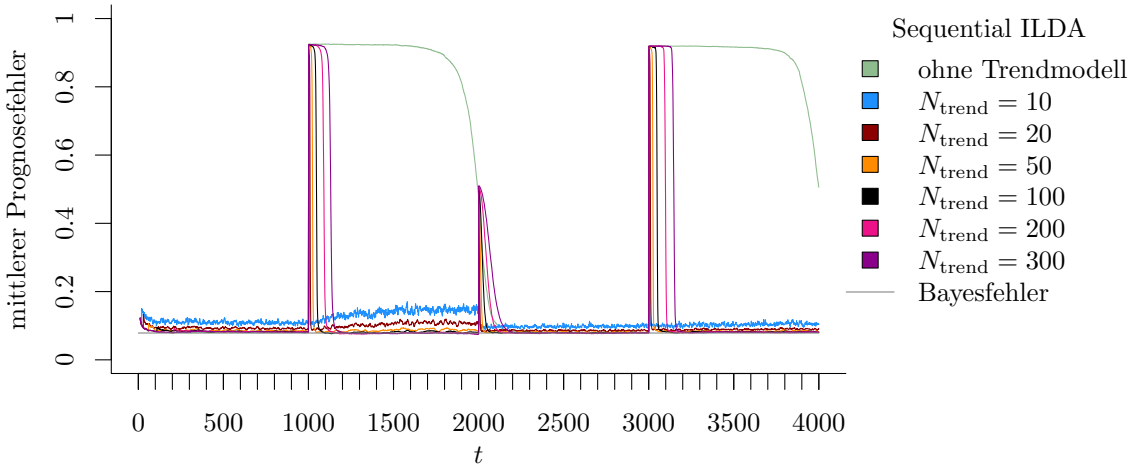
Für *ILDA* und *OLDC fix* kann die Prognosegüte durch die Einführung lokaler linearer Regressionsmodelle verbessert werden. Bis zu $N_{\text{trend}} = 50$ sinkt der durchschnittliche mittlere Prognosefehler über die Zeit immer weiter, bevor er für größere Fenster wieder ansteigt. Die Verbesserung im Durchschnitt über die Zeit ist dabei auf die schnellere Regeneration des Prognosefehlers direkt nach den sudden drifts zurückzuführen. Anhand der Abbildungen 9.66 (a) und 9.67 ist zu sehen, dass der Prognosefehler der Erweiterungen zwischen den Drifts, in Zeitabschnitten stabiler Konzepte, teilweise etwas höher liegt als mit den ursprünglichen Methoden und zudem auch eine höhere Varianz aufweist. Große Werte von N_{trend}, also mehr Beobachtungen für die einzelnen Regressionsmodelle, führen zu einem ruhigeren Verlauf des Prognosefehlers und niedrigeren Fehlerraten in Zeiträumen stabiler Konzepte nahe am Bayesfehler (vgl. z. B. pinke und lila Kurve in Abbildung 9.66 (a)). Kleine Werte von N_{trend} ziehen hingegen unruhigere Kurven des Prognosefehlers nach sich. Der Prognosefehler weist eine höhere Varianz auf und liegt in Zeiträumen stabiler Konzepte

deutlich höher und sichtbar über dem Bayesfehler. Auf der anderen Seite erfolgt jedoch ein schnellerer Abfall nach dem starken Anstieg des Prognosefehlers nach einem Strukturbruch (vgl. z. B. blaue und dunkelrote Kurve in Abbildung 9.66 (a)). Werden die Ergebnisse auf einen einzelnen Wert beschränkt, so ist der durchschnittliche mittlere Prognosefehler über die Zeit in den meisten Fällen für einen mittleren Wert von $N_{\text{trend}} = 50$ minimal.

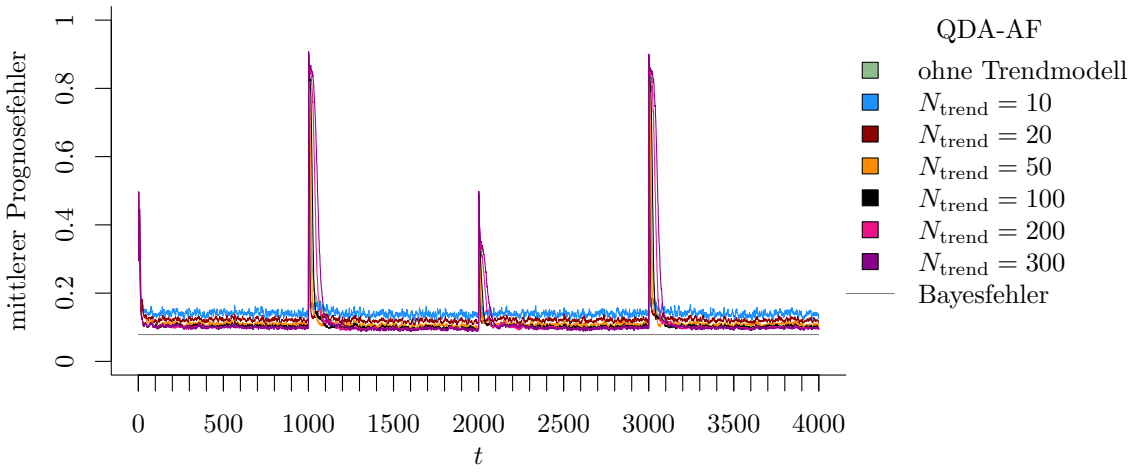
Für die Erweiterung von *OLDC* mit fester Lernrate gibt es folgende Auffälligkeit: Während bei der ursprünglichen Methode der durchschnittliche mittlere Prognosefehler über die Zeit mit steigendem λ sinkt, ist es nach Einführung lokaler linearer Regressionsmodelle genau andersrum. Für schmale Fenster N_{trend} sinkt der durchschnittliche Prognosefehler über die Zeit mit sinkendem λ. Je größer das Fenster für die Regressionsmodelle durch N_{trend} ist, desto mehr verschiebt sich dieses Minimum wieder zugunsten der hohen Lernrate, sodass für $N_{\text{trend}} = 200$ wieder der durchschnittliche Prognosefehler über die Zeit mit wachsendem λ sinkt (vgl. Spalte „OLDC fix“ in Tabelle 9.15).

Zusammengefasst lässt sich sagen:

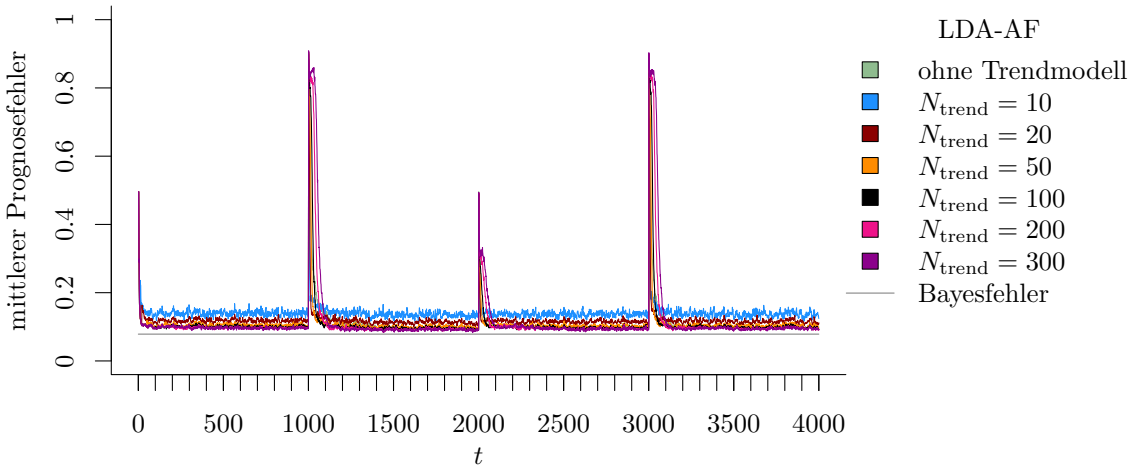
- Der durchschnittliche Bayesfehler über den gesamten Datenstrom für diese Datensituation beträgt 0.0786 mit einer Standardabweichung von 0.
- Bei allen Methoden und Erweiterungen sind die Strukturbrüche durch (teils kurzzeitige) starke Anstiege des Prognosefehlers zu erkennen.
- *QDA-AF* und *LDA-AF* funktionieren gut auf der Datensituation. Es erfolgt nur ein kurzer Anstieg des Prognosefehlers nach den Strukturbrüchen, der Prognosefehler sinkt jedoch sehr schnell wieder auf das Niveau des Bayesfehlers ab. Die Prognosegüte kann für diese Methoden durch die Erweiterung nicht verbessert werden.
- Für *ILDA* und *OLDC fix* kann der durchschnittliche Prognosefehler über die Zeit durch die Erweiterung verbessert werden, da ein schnellerer Abfall des Prognosefehlers nach dem Anstieg erzielt werden kann. Der durchschnittliche Prognosefehler ist minimal für mittlere Werte von N_{trend}, da zwei gegenläufige Effekte wirken:
  1. Große Werte von N_{trend}: Ruhigerer Verlauf des Prognosefehlers und niedrige Fehlerraten in Zeiträumen stabiler Konzepte nahe am Bayesfehler. Allerdings bleibt der Prognosefehler nach einem Strukturbruch länger erhöht.
  2. Kleine Werte von N_{trend}: Der Prognosefehler weist eine höhere Varianz auf und liegt in Zeiträumen stabiler Konzepte sichtbar über dem Bayesfehler. Auf der anderen Seite erfolgt jedoch ein schnellerer Abfall nach dem starken Anstieg des Prognosefehlers nach einem Strukturbruch.
- Bei *OLDC fix* sinkt der durchschnittliche Prognosefehler mit wachsender Lernrate. Nach Erweiterung ist der Effekt für kleine N_{trend} zunächst gegenläufig, bis mit wachsendem N_{trend} der durchschnittliche Prognosefehler wieder für große λ abnimmt.
- Bei *OLDC adaptive* können die bereits geringen durchschnittlichen Prognosefehler durch die Erweiterung in dieser Datensituation plötzlicher Drifts generell nicht verbessert werden. Die Ergebnisse werden jedoch auch nicht auffallend verschlechtert.



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

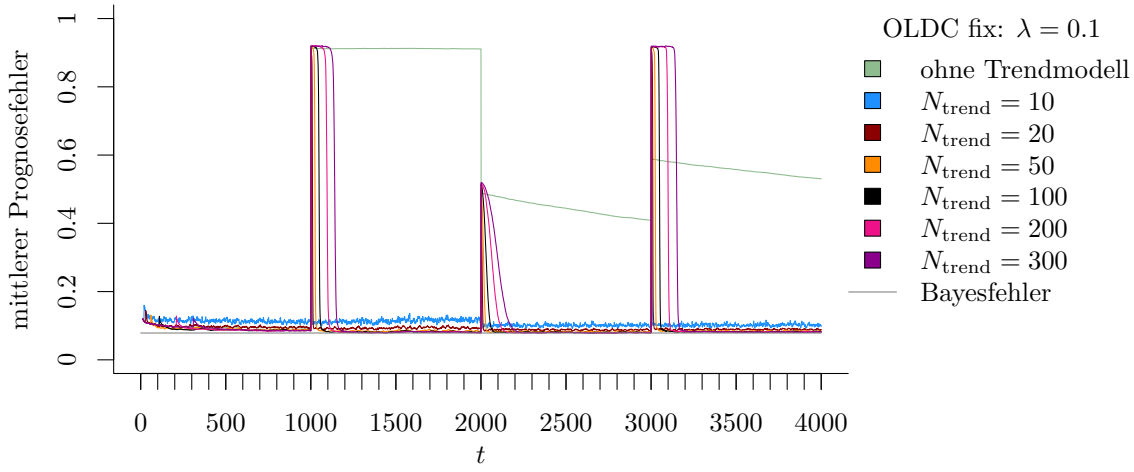


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

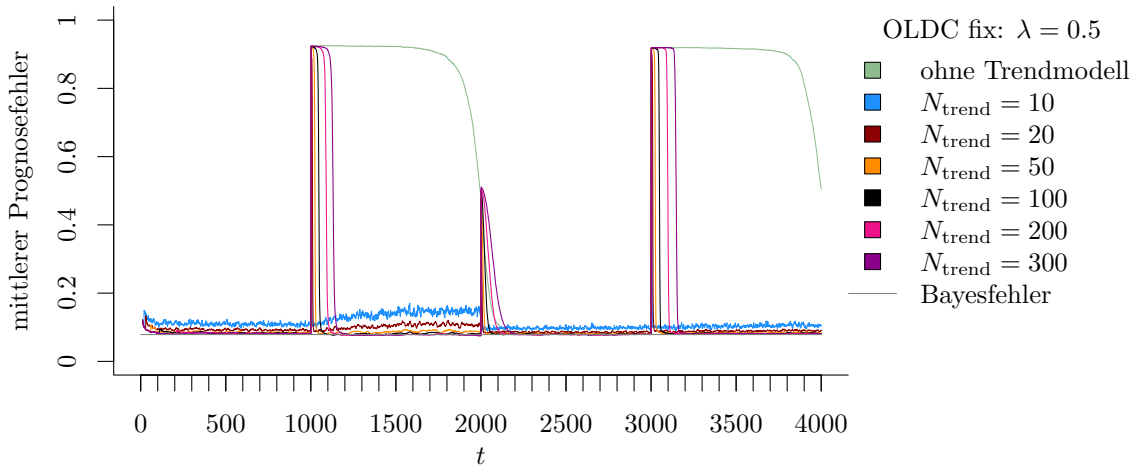


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

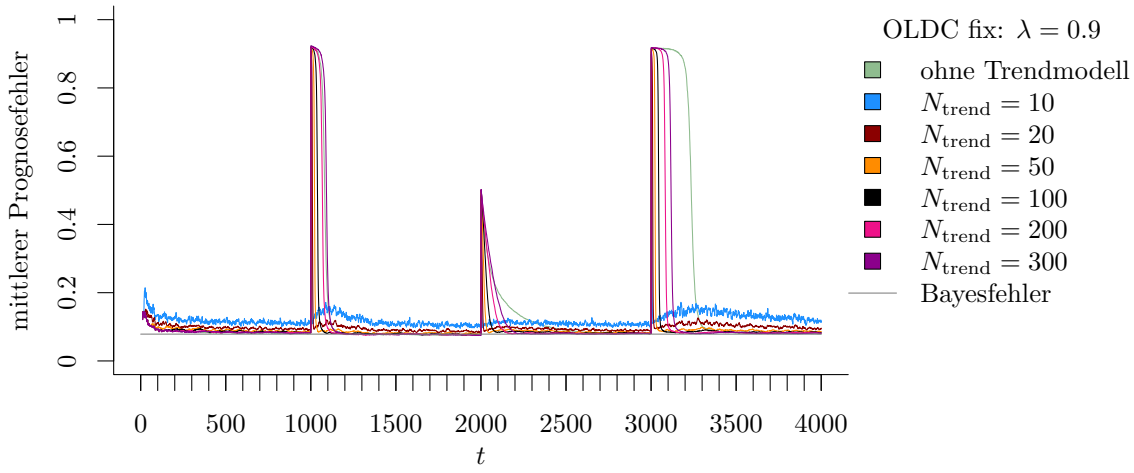
Abbildung 9.66: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Sudden Drift** im zweidimensionalen Raum.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.

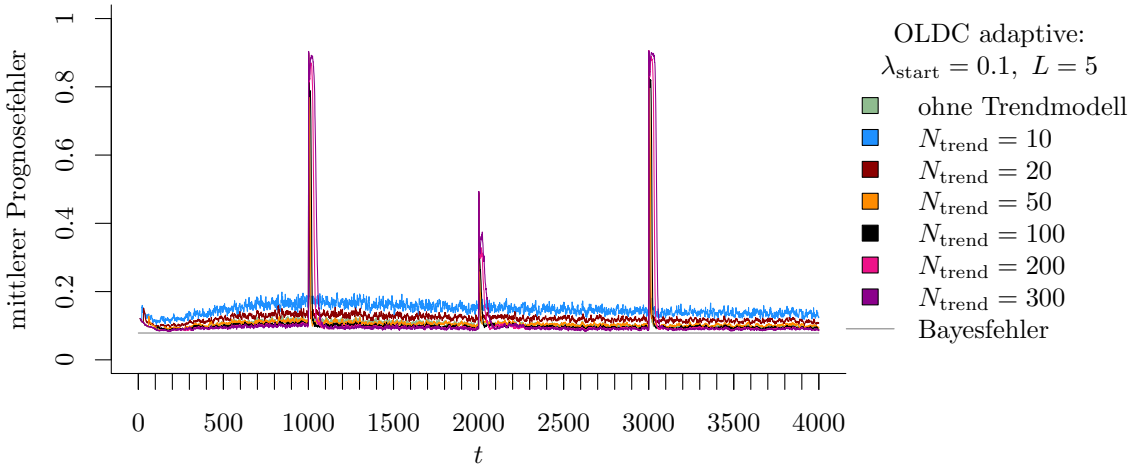


(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.

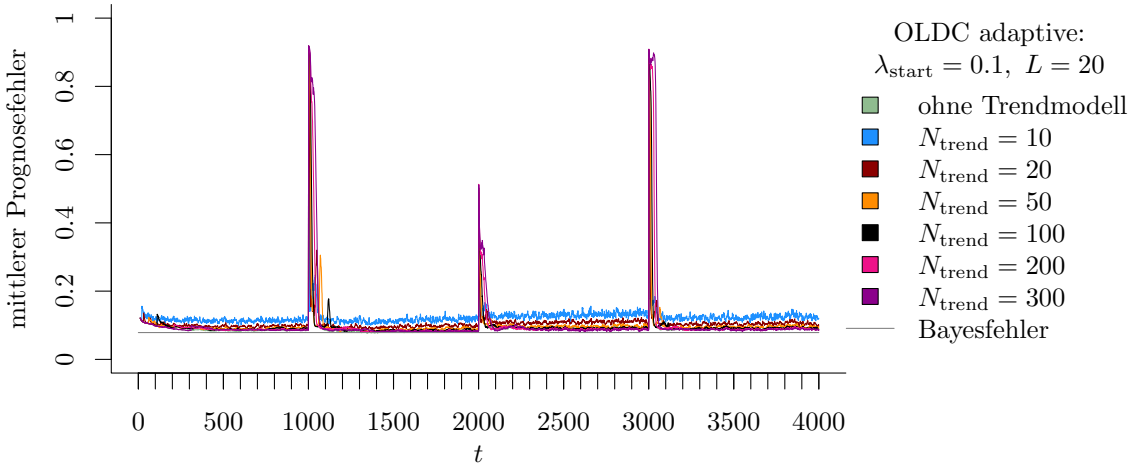


(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

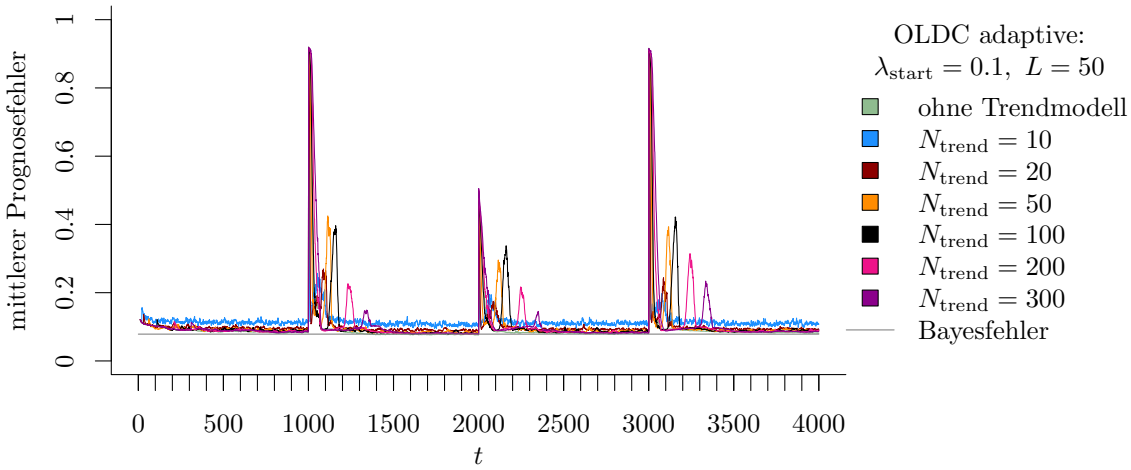
Abbildung 9.67: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Sudden Drift** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

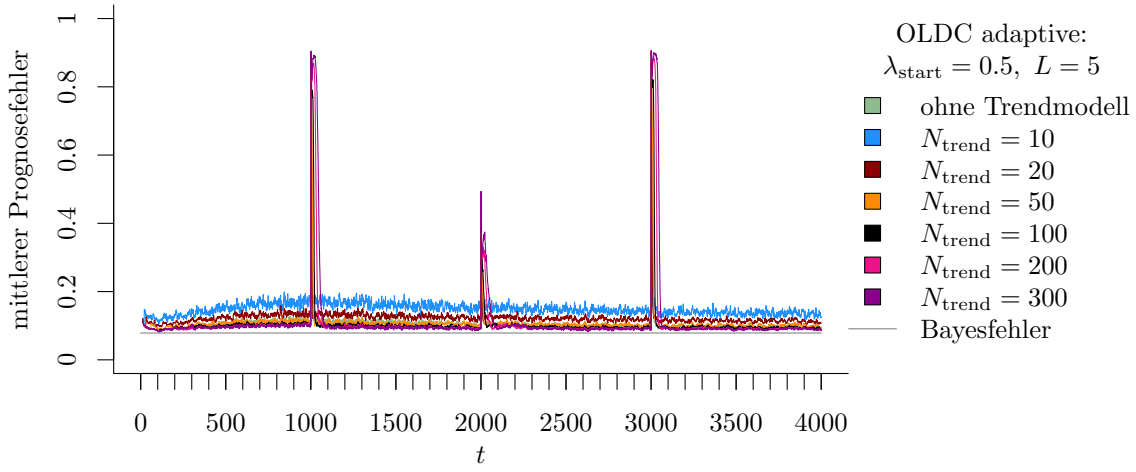


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

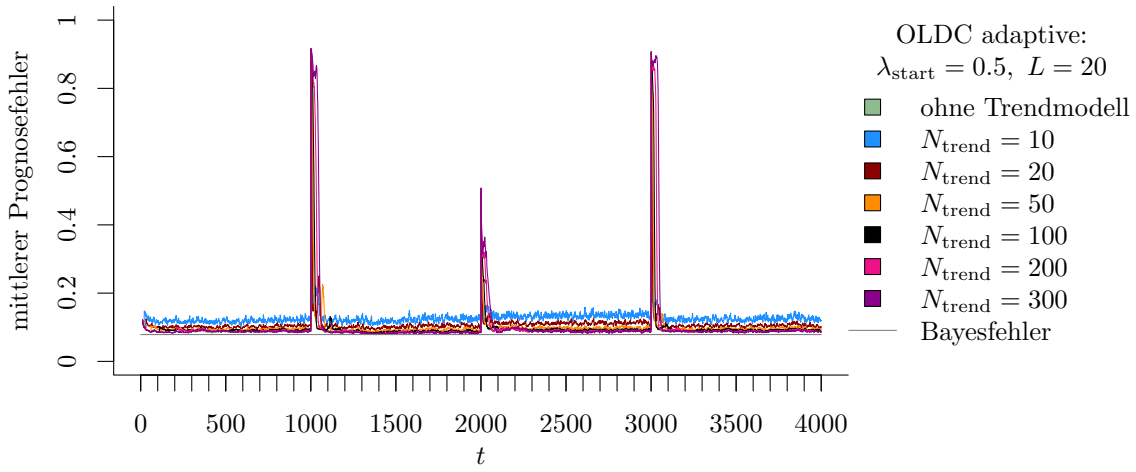


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

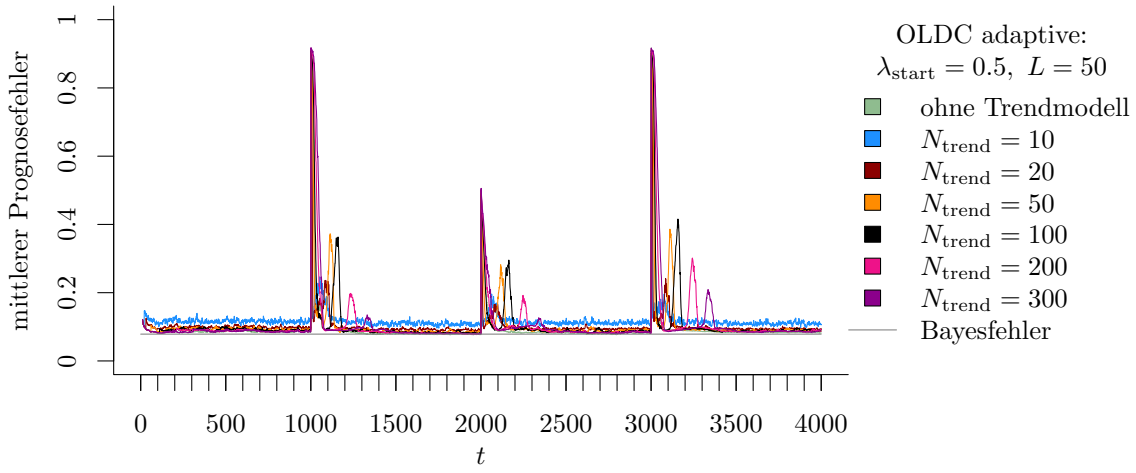
Abbildung 9.68: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Sudden Drift** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

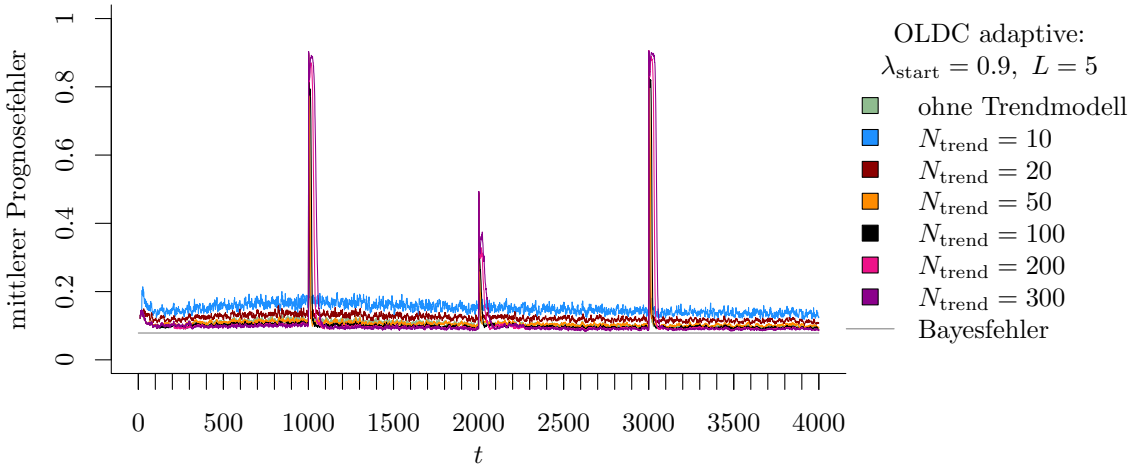


(b) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

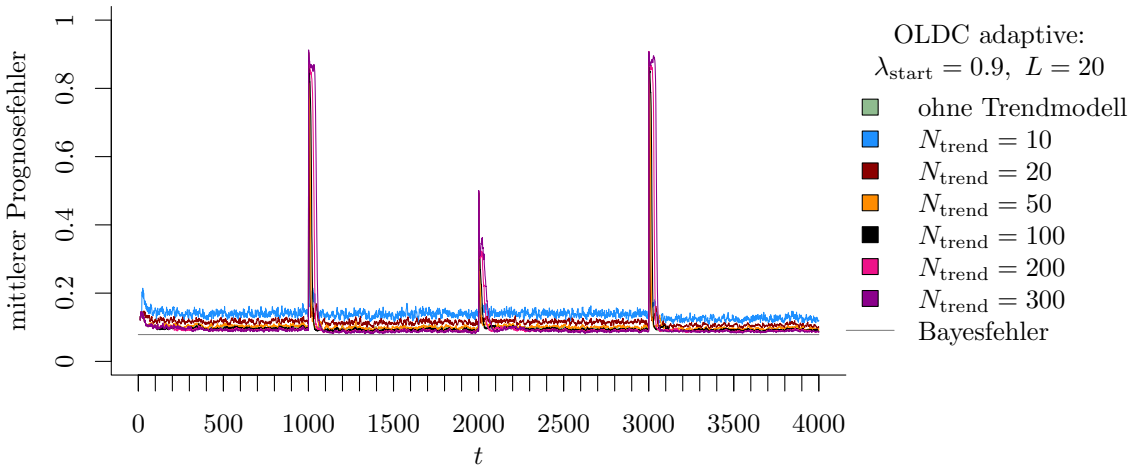


(c) **OLDC** mit $\lambda_{\text{start}} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

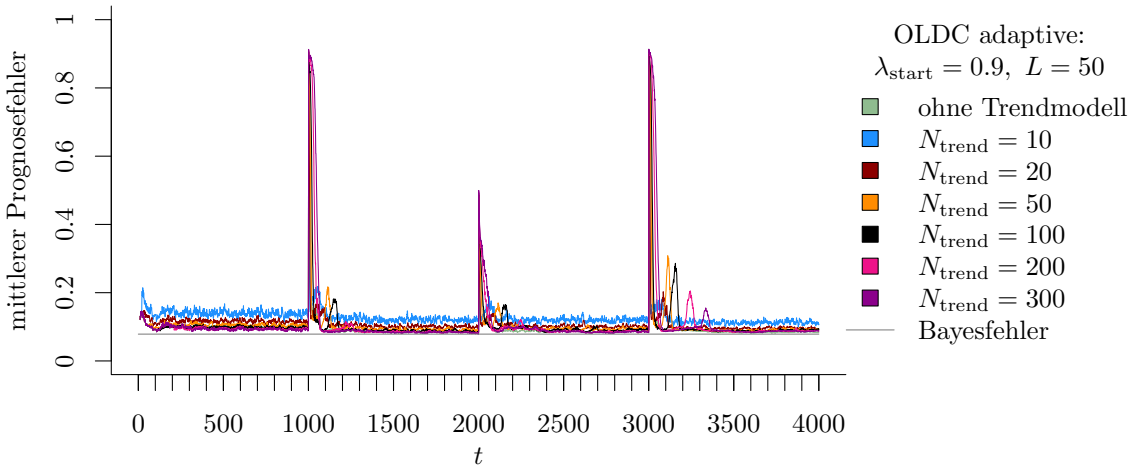
Abbildung 9.69: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Sudden Drift** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.70: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **Sudden Drift** im zweidimensionalen Raum.

Tabelle 9.15: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **Sudden Drift** ($p = 2$) getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4896 (0.004)	**0.1124** (0.003)	**0.1090** (0.003)	0.1	0.5037 (0.009)	5	0.1057 (0.002)
						20	***0.0938*** (0.001)
						50	0.0974 (0.002)
				0.3	0.5009 (0.002)		
				0.5	0.4896 (0.004)	5	0.1060 (0.002)
						20	0.0946 (0.001)
						50	0.0956 (0.002)
				0.7	0.3134 (0.002)		
				0.9	0.1582 (0.001)	5	0.1074 (0.002)
						20	0.0997 (0.002)
						50	0.0976 (0.002)
10	0.1144 (0.003)	0.1432 (0.006)	0.1410 (0.007)	0.1	*0.1101* (0.002)	5	0.1493 (0.008)
						20	0.1242 (0.004)
						50	0.1176 (0.004)
				0.3	0.1103 (0.003)		
				0.5	0.1144 (0.003)	5	0.1501 (0.008)
						20	0.1277 (0.005)
						50	0.1184 (0.004)
				0.7	0.1202 (0.003)		
				0.9	0.1218 (0.003)	5	0.1524 (0.009)
						20	0.1387 (0.007)
						50	0.1287 (0.005)
20	0.0973 (0.001)	0.1254 (0.003)	0.1202 (0.003)	0.1	0.0965 (0.001)	5	0.1232 (0.004)
						20	0.1072 (0.002)
						50	0.1036 (0.002)
				0.3	*0.0953* (0.001)		
				0.5	0.0973 (0.001)	5	0.1237 (0.004)
						20	0.1094 (0.002)
						50	0.1038 (0.002)
				0.7	0.1002 (0.001)		
				0.9	0.1017 (0.001)	5	0.1253 (0.004)
						20	0.1167 (0.003)
						50	0.1102 (0.003)
50	**0.0956** (0.001)	0.1155 (0.002)	0.1104 (0.002)	0.1	0.0973 (0.001)	5	0.1084 (0.002)
						20	0.1009 (0.002)
						50	0.1055 (0.003)
				0.3	***0.0950*** (0.001)		
				0.5	0.0956 (0.001)	5	0.1085 (0.002)
						20	0.1017 (0.002)
						50	0.1050 (0.003)
				0.7	0.0973 (0.001)		
				0.9	0.0982 (0.001)	5	0.1096 (0.002)
						20	0.1055 (0.002)
						50	0.1062 (0.003)
100	0.1042 (0.001)	0.1135 (0.002)	0.1092 (0.002)	0.1	0.1077 (0.001)	5	0.1047 (0.002)
						20	*0.0998* (0.001)
						50	0.1086 (0.004)
				0.3	0.1044 (0.001)		
				0.5	0.1042 (0.001)	5	0.1048 (0.002)
						20	0.0999 (0.001)
						50	0.1076 (0.004)
				0.7	0.1051 (0.001)		
				0.9	0.1044 (0.001)	5	0.1058 (0.002)
						20	0.1025 (0.002)
						50	0.1058 (0.003)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	0.1251 (0.001)	0.1175 (0.002)	0.1139 (0.003)	0.1	0.1312 (0.002)	5	0.1079 (0.002)
						20	*0.1028* (0.002)
						50	0.1096 (0.004)
				0.3	0.1263 (0.001)		
				0.5	0.1251 (0.001)	5	0.1080 (0.002)
						20	*0.1028* (0.002)
						50	0.1084 (0.003)
				0.7	0.1244 (0.001)		
				0.9	0.1187 (0.001)	5	0.1092 (0.002)
						20	0.1054 (0.002)
						50	0.1078 (0.002)
300	0.1464 (0.001)	0.1236 (0.002)	0.1204 (0.003)	0.1	0.1553 (0.002)	5	0.1138 (0.002)
						20	*0.1084* (0.002)
						50	0.1125 (0.004)
				0.3	0.1489 (0.001)		
				0.5	0.1464 (0.001)	5	0.1140 (0.002)
						20	0.1085 (0.002)
						50	0.1112 (0.003)
				0.7	0.1437 (0.001)		
				0.9	0.1321 (0.001)	5	0.1152 (0.002)
						20	0.1115 (0.002)
						50	0.1124 (0.002)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0786 (Standardabweichung 0)

Die durchschnittlichen euklidischen Abstände zwischen wahren und mittleren geschätzten (ursprüngliche Methoden) bzw. mittleren prognostizierten (erweiterte Methoden) Erwartungswertvektoren über die Zeit (vgl. Tabelle 9.16) lassen ähnliche Schlussfolgerungen zu wie die durchschnittlichen mittleren Prognosefehler über die Zeit. Für *ILDA* und *OLDC fix* können die wahren Erwartungswertvektoren durch die Erweiterung besser geschätzt werden, sodass verbesserte Schätzer in die Klassifikationsregel der Diskriminanzanalyse zu jedem Zeitpunkt einfließen können. Dies lässt sich daraus schließen, dass bei Betrachtung von Fenstern der Breite $N_{\text{trend}} = 20$ die durchschnittlichen euklidischen Abstände über die Zeit minimal sind (vgl. Spalten „ILDA“ und „OLDC fix“). Für *QDA-AF* und *LDA-AF* können die Schätzungen der Erwartungswertvektoren im Mittel über die Zeit durch Erweiterung der Methode bei diesen sudden drifts nicht verbessert werden.

Auch bezüglich der durchschnittlichen euklidischen Abstände ist bei der ursprünglichen Methode *OLDC* eine große Lernrate λ am besten. Nach Erweiterung der Methode führen kleine Lernraten zu niedrigeren durchschnittlichen euklidischen Abständen, jedoch unterscheiden sich die Ergebnisse für die einzelnen Lernraten nicht mehr so stark wie zuvor.

Bei *OLDC adaptive* führt die Erweiterung der Methode zu keiner Verbesserung der Prognosegüte im Falle dieser sudden drifts. Dies zeigt sich anhand der durchschnittlichen mittleren Prognosefehler über die Zeit sowie auch anhand der durchschnittlichen euklidischen Abstände zwischen wahren und mittleren geschätzten (ursprüngliche Methoden) bzw. mittleren prognostizierten (erweiterte Methoden) Erwartungswertvektoren über die Zeit (vgl. Spalte „OLDC adaptive“). Die durchschnittlichen euklidischen Abstände sind hier bereits für die ursprüngliche Methode unabhängig von den Parametern λ_{start} und L im Vergleich zu den anderen Methoden gering.

Tabelle 9.16: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **Sudden Drift ($p = 2$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	1.6811 (0.004)	**0.1114** (0.127)	**0.1215** (0.141)	0.1	2.2859 (0.134)	5	0.0866 (0.148)
	1.6790 (0.005)	**0.1108** (0.127)	**0.1205** (0.144)		2.2407 (0.104)		0.0900 (0.148)
						20	***0.0782*** (0.084)
							0.0831 (0.082)
						50	0.1139 (0.076)
							0.1132 (0.070)
				0.3	1.9486 (0.010)		
					1.9390 (0.009)		
				0.5	1.6811 (0.004)	5	0.0878 (0.148)
					1.6790 (0.005)		0.0888 (0.149)
						20	0.0790 (0.095)
							0.0802 (0.095)
						50	0.1042 (0.073)
							0.1034 (0.074)
				0.7	1.4092 (0.006)		
					1.4073 (0.007)		
				0.9	0.6413 (0.016)	5	0.0889 (0.155)
					0.6364 (0.017)		0.0905 (0.156)
						20	0.0829 (0.127)
							0.0840 (0.127)
						50	0.0909 (0.118)
							0.0912 (0.118)
10	0.1162 (0.348)	0.1486 (0.620)	0.1487 (0.631)	0.1	*0.1155* (0.345)	5	0.1679 (0.874)
	0.1134 (0.349)	0.1455 (0.632)	0.1454 (0.635)		*0.1126* (0.345)		0.1647 (0.857)
						20	0.1375 (0.541)
							0.1342 (0.533)
						50	0.1244 (0.442)
							0.1219 (0.422)
				0.3	0.1158 (0.346)		
					0.1129 (0.346)		
				0.5	0.1162 (0.348)	5	0.1685 (0.882)
					0.1134 (0.349)		0.1651 (0.871)
						20	0.1427 (0.588)
							0.1390 (0.585)
						50	0.1270 (0.465)
							0.1239 (0.446)
				0.7	0.1173 (0.354)		
					0.1144 (0.354)		
				0.9	0.1219 (0.379)	5	0.1708 (0.911)
					0.1188 (0.380)		0.1671 (0.902)
						20	0.1559 (0.721)
							0.1525 (0.722)
						50	0.1418 (0.660)
							0.1391 (0.607)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
20	**0.0940** (0.130)	0.1249 (0.275)	0.1241 (0.277)	0.1	***0.0923*** (0.128)	5	0.1364 (0.353)
	0.0935 (0.130)	0.1240 (0.275)	0.1227 (0.279)		***0.0917*** (0.128)		0.1371 (0.354)
						20	0.1158 (0.276)
							0.1171 (0.274)
						50	0.1028 (0.218)
							0.1033 (0.221)
				0.3	0.0929 (0.128)		
					0.0923 (0.129)		
				0.5	0.0940 (0.130)	5	0.1357 (0.356)
					0.0935 (0.130)		0.1374 (0.358)
						20	0.1192 (0.292)
							0.1194 (0.295)
						50	0.1047 (0.226)
							0.1048 (0.229)
				0.7	0.0967 (0.133)		
					0.0961 (0.134)		
				0.9	0.1052 (0.148)	5	0.1362 (0.363)
					0.1044 (0.149)		0.1384 (0.366)
						20	0.1267 (0.332)
							0.1271 (0.333)
						50	0.1155 (0.284)
							0.1163 (0.285)
50	0.1237 (0.053)	0.1339 (0.137)	0.1299 (0.137)	0.1	*0.1167* (0.052)	5	0.1307 (0.157)
	0.1251 (0.054)	0.1306 (0.136)	0.1300 (0.138)		*0.1168* (0.052)		0.1289 (0.158)
						20	0.1405 (0.206)
							0.1432 (0.202)
						50	0.1270 (0.338)
							0.1300 (0.345)
				0.3	0.1190 (0.052)		
					0.1200 (0.052)		
				0.5	0.1237 (0.053)	5	0.1302 (0.157)
					0.1251 (0.054)		0.1295 (0.159)
						20	0.1406 (0.191)
							0.1432 (0.191)
						50	0.1272 (0.336)
							0.1291 (0.335)
				0.7	0.1341 (0.056)		
					0.1355 (0.057)		
				0.9	0.1510 (0.068)	5	0.1295 (0.158)
					0.1496 (0.069)		0.1295 (0.160)
						20	0.1351 (0.171)
							0.1374 (0.173)
						50	0.1315 (0.278)
							0.1339 (0.281)
100	0.2185 (0.032)	0.1690 (0.086)	0.1679 (0.087)	0.1	0.2027 (0.032)	5	0.1536 (0.085)
	0.2201 (0.033)	0.1652 (0.083)	0.1672 (0.085)		0.2025 (0.032)		0.1528 (0.083)
						20	0.1990 (0.170)
							0.2050 (0.162)
						50	0.1845 (0.723)
							0.1858 (0.716)
				0.3	0.2077 (0.031)		
					0.2089 (0.032)		
				0.5	0.2185 (0.032)	5	0.1532 (0.085)
					0.2201 (0.033)		0.1521 (0.082)
						20	0.1963 (0.142)
							0.2014 (0.139)
						50	0.1921 (0.691)
							0.1857 (0.686)
				0.7	0.2402 (0.036)		
					0.2406 (0.036)		
				0.9	0.2528 (0.046)	5	*0.1526* (0.085)
					0.2507 (0.046)		*0.1520* (0.083)
						20	0.1791 (0.113)
							0.1791 (0.110)
						50	0.1849 (0.393)
							0.1852 (0.383)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.4211 (0.024)	0.2519 (0.062)	0.2553 (0.064)	0.1	0.3960 (0.026)	5	0.2248 (0.047)
	0.4230 (0.025)	0.2517 (0.061)	0.2545 (0.063)		0.3960 (0.026)		0.2243 (0.046)
						20	0.3058 (0.147)
							0.3157 (0.152)
						50	0.3223 (0.776)
							0.3216 (0.759)
				0.3	0.4041 (0.023)		
					0.4058 (0.024)		
				0.5	0.4211 (0.024)	5	0.2244 (0.046)
					0.4230 (0.025)		0.2232 (0.046)
						20	0.2917 (0.124)
							0.2991 (0.138)
						50	0.3365 (0.694)
							0.3300 (0.698)
				0.7	0.4541 (0.027)		
					0.4543 (0.029)		
				0.9	0.4487 (0.035)	5	*0.2237* (0.047)
					0.4459 (0.036)		*0.2228* (0.046)
						20	0.2593 (0.079)
							0.2601 (0.085)
						50	0.3262 (0.406)
							0.3237 (0.427)
300	0.6230 (0.021)	0.3483 (0.053)	0.3533 (0.058)	0.1	0.5938 (0.026)	5	0.3087 (0.034)
	0.6257 (0.023)	0.3454 (0.055)	0.3498 (0.057)		0.5949 (0.026)		0.3092 (0.035)
						20	0.4066 (0.130)
							0.4176 (0.142)
						50	0.4978 (0.625)
							0.5012 (0.612)
				0.3	0.6032 (0.021)		
					0.6060 (0.022)		
				0.5	0.6230 (0.021)	5	0.3081 (0.034)
					0.6257 (0.023)		0.3082 (0.035)
						20	0.3853 (0.101)
							0.3917 (0.123)
						50	0.5146 (0.595)
							0.5133 (0.593)
				0.7	0.6603 (0.024)		
					0.6616 (0.027)		
				0.9	0.6314 (0.031)	5	*0.3077* (0.034)
					0.6292 (0.033)		*0.3077* (0.035)
						20	0.3461 (0.062)
							0.3474 (0.067)
						50	0.4769 (0.436)
							0.4724 (0.419)

In Abbildung 9.71 ist der Verlauf der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren beider Klassen für die Methode *ILDA* und ihrer Erweiterung veranschaulicht. Während in Tabelle 9.16 die Ergebnisse der „Güte der Schätzer der Erwartungswertvektoren“ auf einen einzelnen Wert über den gesamten Datenstrom zusammengefasst sind, liefert diese Abbildung mehr Erkenntnisse über den Verlauf der verwendeten Schätzer in der Klassifikationsregel der LDA über die Zeit. Anhand der linken oberen Grafik wird deutlich, dass auch hier bei der ursprünglichen Update-Methode die Schätzer für die Erwartungswertvektoren „zeitlich hinterher hängen“, da die Schätzer zeitlich verzögert auf die sudden drifts reagieren. Nach 1000 Beobachtungen erfolgt in der Datensituation der erste sudden drift. Die Mittelwertschätzer reagieren langsam auf diesen Drift, sodass sie sich ab Zeitpunkt $t = 1000$ von $(2, -2)^T$ zu $(-2, 2)^T$ (Klasse 1) bzw. von $(-2, 2)^T$ zu $(2, -2)^T$ (Klasse 2) bewegen. Nach weiteren 1000 Zeitpunkten (Zeitpunkt $t = 2000$) erfolgt jedoch

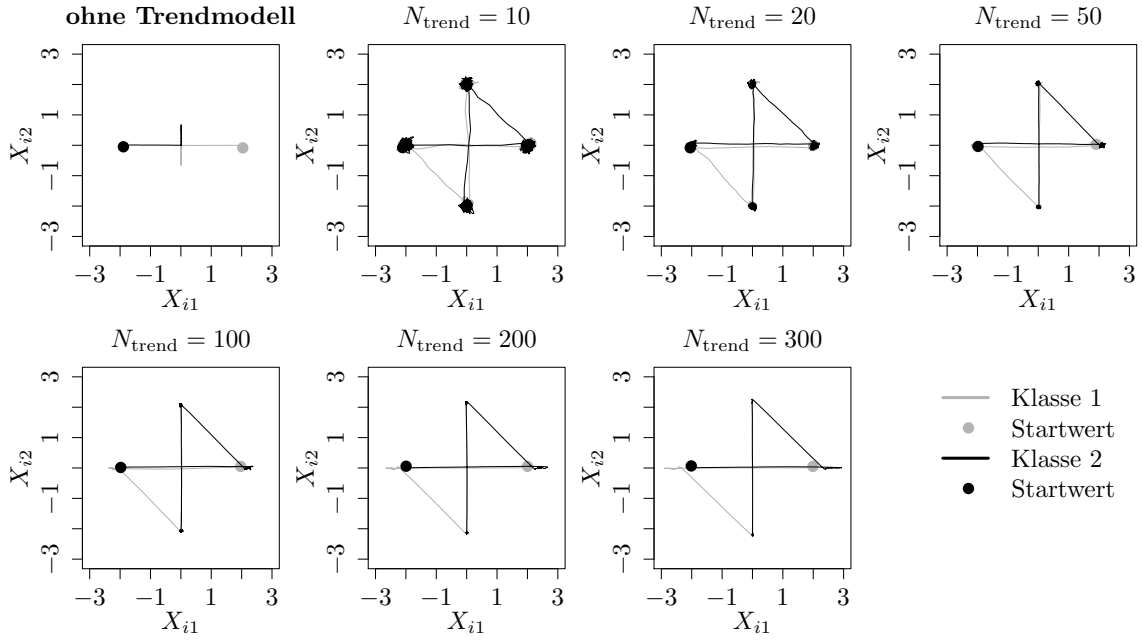


Abbildung 9.71: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der geschätzten bzw. prognostizierten zweidimensionalen Erwartungswertvektoren auf der Datensituation **Sudden Drift** für den gesamten Datenstrom. Ergebnisse aus *ILDA* und Erweiterung mit verschiedenen N_{trend}.

bereits der zweite Strukturbruch. Zu diesem Zeitpunkt sind die beiden Mittelwertvektoren erst bei $(0, 0)^T$ angekommen und repräsentieren beide noch nicht die aktuellen Erwartungswertvektoren des zweiten Konzepts von $(-2, 2)^T$ (Klasse 1) bzw. $(2, -2)^T$ (Klasse 2). Da nun eine Reaktion auf den zweiten Strukturbruch erfolgt, bewegen sich die Mittelwertschätzer in Richtung der Erwartungswertvektoren $(0, -2)^T$ (Klasse 1) bzw. $(2, 0)^T$ (Klasse 2) des dritten Konzepts. Zum Zeitpunkt $t = 3000$ erfolgt der nächste Strukturbruch. Zu diesem Zeitpunkt sind die Schätzer noch nicht bei $(0, -2)^T$ (Klasse 1) bzw. $(2, 0)^T$ (Klasse 2) angekommen, reagieren langsam auf den nächsten Strukturbruch und bewegen sich daher langsam wieder in die genau entgegengesetzte Richtung.

Insgesamt hängen die Mittelwertschätzer zeitlich immer hinterher und repräsentieren nicht die aktuellen Erwartungswertvektoren bzw. auch nicht jene des jeweils folgenden Zeitpunktes $t + 1$. Dies kann auch bei Datensituationen mit sudden drifts durch die Erweiterung der Methoden durch Modellierung des Trends der Erwartungswerkvektoren durch lokale lineare Regressionsmodelle und darauf folgende Prognose der Erwartungswertvektoren verbessert werden. Zwar kann auch bei der Erweiterung keine direkte Reaktion auf den sudden drift erfolgen, da ein linearer Trend und keine plötzliche Änderung unterstellt wird. Jedoch laufen die verbesserten Schätzer relativ schnell in die neue Richtung der Erwartungswertvektoren. Die zeitliche Verzögerung kann zudem ausgeschaltet werden, sodass hier auch die Erwartungswertvektoren aller vier betrachteten Konzepte im Laufe des Datenstroms repräsentativ geschätzt werden können (vgl. Abbildung 9.71). Ab $N_{\text{trend}} = 100$ ist zu erkennen, dass die Erwartungswertvektoren durch die Prognose basierend auf lokalen linearen Regressionsmodellen sogar leicht „überschätzt“ werden, da die Schätzer über die Werte -2

bzw. 2 in erster und zweiter Dimension hinauslaufen. Daher werden die durchschnittlichen euklidischen Abstände über die Zeit mit steigender Fenstergröße N_{trend} auch wieder größer.

Fazit: Diese Datensituation ist ein Beispiel für den Raum der Datensituationen mit sudden drifts, ebenso wie die Datensituation „STAGGER“ (vgl. Seite 246 ff.). Es zeigt sich, dass wie bei „STAGGER“ der Prognosefehler für einige der Methoden (hier *ILDA* und *OLDC* mit fester Lernrate) durch Integration lokaler linearer Regressionsmodelle zur Modellierung des Trends der Erwartungswerte und deren Prognose in der Klassifikationsregel verringert werden kann, obwohl die Annahme eines linearen Trends der Erwartungswertvektoren in dieser Datensituation nicht gerechtfertigt ist. Ebenso kann der durchschnittliche euklidische Abstand zwischen wahren und mittleren geschätzten (ursprüngliche Methoden) bzw. mittleren prognostizierten (erweiterte Methoden) Erwartungswertvektoren über die Zeit für diese Methoden verringert werden. Allgemein greift die lineare Approximation der plötzlichen Strukturbrüche und die Prognosegüte der Klassifikatoren kann für diese Methoden verbessert werden. Dabei hängt die optimale Wahl von N_{trend} wie bei „STAGGER“ von den zeitlichen Abständen zwischen den einzelnen sudden drifts ab und kann nicht pauschal festgelegt werden. Auch hier wäre in der Praxis wieder die Möglichkeit gegeben den Parameter zunächst eine Zeit lang auf dem Datenstrom zu tunen oder Vorkenntnisse über die Datensituation einfließen zu lassen.

Datensituation ohne Drift ($p = 2$) Die Datensituation ohne Drift (vgl. Seite 224 f.) wird betrachtet, um zu untersuchen, ob die Erweiterungen der Methoden den Prognosefehler nicht allzu stark verschlechtern. Da in praktischen Anwendungen nicht immer klar ist, ob ein concept drift vorliegt, und vor allem, ob dieser kontinuierlich die ganze Zeit auftritt, ist es wünschenswert, dass die entwickelten Erweiterungen der Methoden auch im Falle konstanter Verteilungen herangezogen werden können ohne das Ergebnis negativ zu beeinflussen.

In den Abbildungen 9.73 und 9.74 ist der Verlauf der Prognosefehler für einige der betrachteten Methoden und ihrer Erweiterungen zusammen mit dem Bayesfehler, welcher konstant 0.0786 beträgt, vergleichend dargestellt. Es ist zu erkennen, dass sich der Verlauf für alle Methoden und alle betrachteten Werte für N_{trend} relativ ähnlich verhält. Die Kurven für vergleichsweise kleine Fenster durch $N_{\text{trend}} \in \{10, 20, 50\}$ sind leicht nach oben verschoben. Generell liegen jedoch alle Prognosefehler nahe am Bayesfehler. Beispielhafte (äquivalente) Grafiken für *OLDC adaptive* sind in Abbildungen D.1–D.3 in Anhang D zu finden.

Beim Vergleich der durchschnittlichen mittleren Prognosefehler über die Zeit (vgl. Tabelle 9.17) ist zusätzlich zu erkennen, dass *ILDA* bezüglich der Prognosefehler im Vergleich immer am besten ist. Dies ist nicht verwunderlich, da *ILDA* die Methode ohne Gewichte zur Anpassung an einen möglichen concept drift ist. Den geringsten durchschnittlichen mittleren Prognosefehler über die Zeit von 0.0799 weist daher *ILDA* bzw. *OLDC* mit fester Lernrate $\lambda = 0.5$ (und $\lambda = 0.7$) ohne Erweiterung durch lokale lineare Regressionsmodelle auf. Dieser Wert liegt sehr nahe am (durchschnittlichen) Bayesfehler von 0.0786.

Auffällig ist, dass für *QDA-AF* und *LDA-AF* der durchschnittliche mittlere Prognosefehler über die Zeit bei Einbindung lokaler linearer Regressionsmodelle auf recht breiten Fenstern von $N_{\text{trend}} = 300$ am besten ist. Allerdings ist hier zu sagen, dass die Größenordnung dieser durchschnittlichen mittleren Prognosefehler nur gering von jener der Ergebnisse der ursprünglichen Methoden ohne Erweiterung abweicht.

Wird *ILDA* oder *OLDC* mit fester Lernrate betrachtet, so fällt auf, dass zwar der Prognosefehler bei Einbindung der Trendmodelle immer schlechter ist als bei den Methoden ohne Trendmodell, dieser mit wachsendem N_{trend} im Allgemeinen aber wieder sinkt. Dies ist auf mehr Beobachtungen pro Regressionsmodell und demnach eine stabilere Schätzung der Erwartungswertvektoren und folglich stabilere Prognose zurückzuführen.

Bei den Ergebnissen von *OLDC* mit adaptiver Lernrate wird deutlich, dass die durchschnittlichen mittleren Prognosefehler über die Zeit der ursprünglichen Methode für einige Kombinationen aus L und λ_{start} durch die Betrachtung eines breiten Fensters N_{trend} für die Regressionsmodelle verringert werden können. Bei kleinem $L = 5$ sinken die Prognosefehler bereits ab $N_{\text{trend}} = 100$ unterhalb die ursprünglichen Ergebnisse. Bei größerem Fenster L zur Adaption der Lernrate muss teilweise ein breiteres Fenster N_{trend} für die lokalen linearen Regressionsmodelle betrachtet werden. Dann ist es jedoch für einige Kombinationen mit λ_{start} möglich den durchschnittlichen Prognosefehler über die Zeit zu verringern. Dies

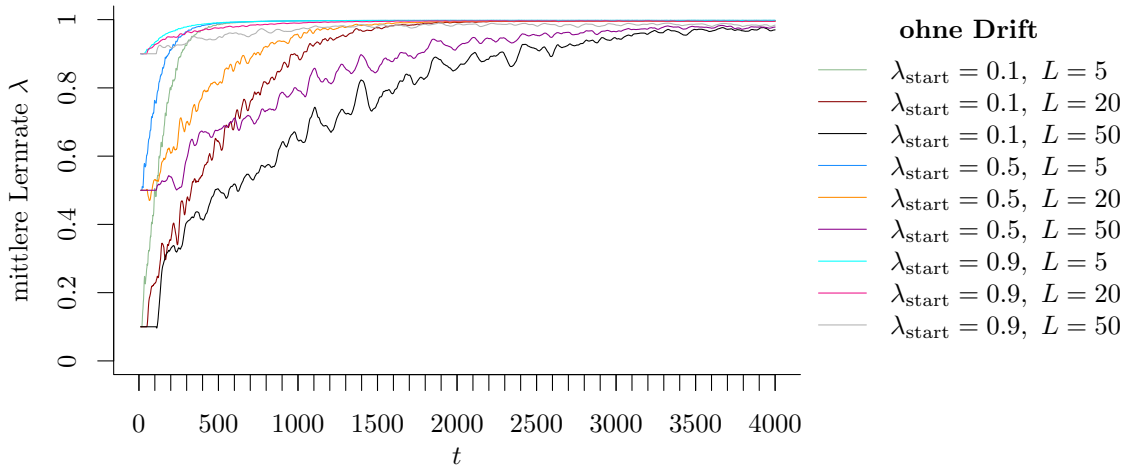
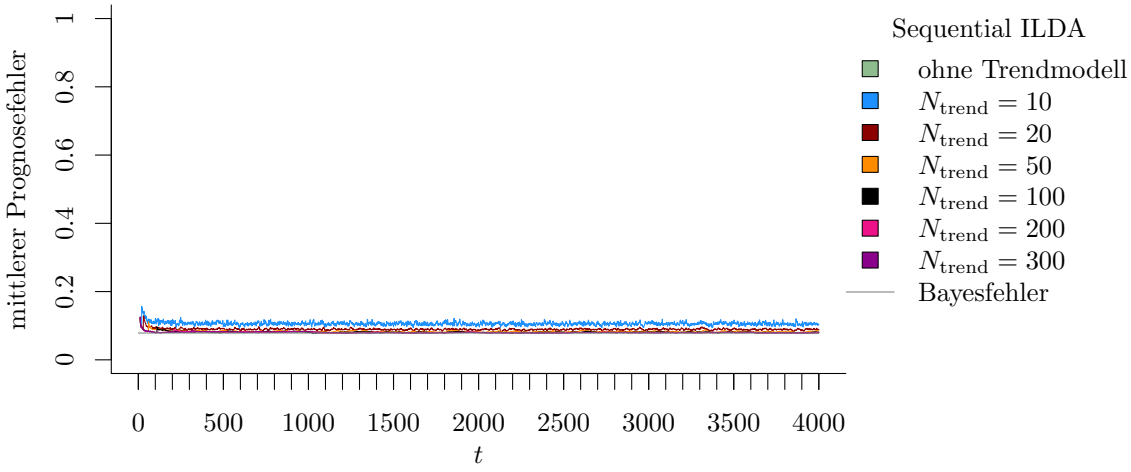


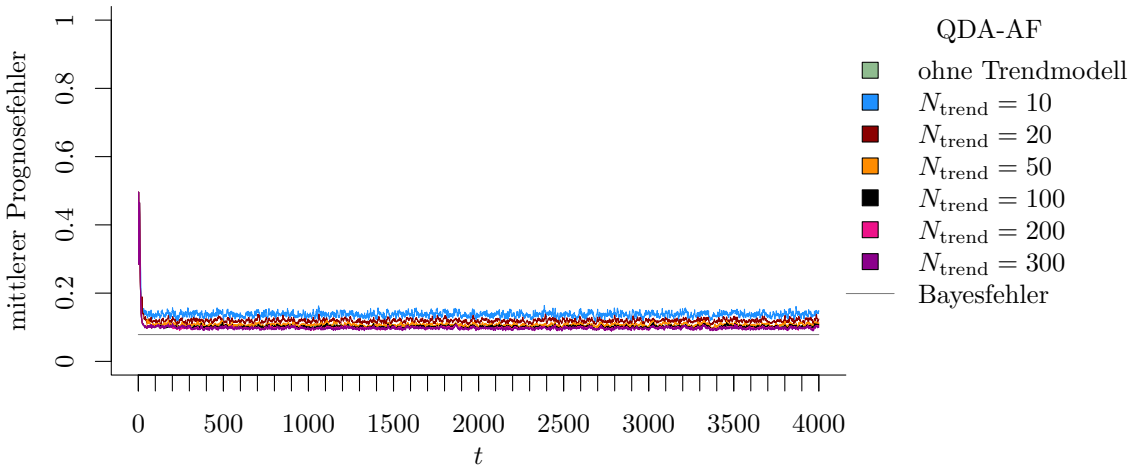
Abbildung 9.72: Mittlerer Verlauf (über 100 Simulationsdurchläufe) der adaptiven Lernrate bei *OLDC* auf der Datensituation **ohne Drift**.

lässt sich dadurch erklären, dass die Prognosefehler bei *OLDC adaptive* verhältnismäßig hoch sind. Sie liegen fast alle über jenen von *OLDC* mit beliebiger fester Lernrate (vgl. erste „Zeile“ in Tabelle 9.17). Mit Blick auf Abbildung 9.72 fällt auf, dass trotz konstanter Verteilung die adaptive Lernrate λ für alle Kombinationen aus Startwert λ_{start} und Fenstergröße L zur Adaption im Laufe des Datenstroms immer weiter ansteigt und sich dem Maximum nähert. Dabei liegt die (schwarze) Kurve für $\lambda_{\text{start}} = 0.1$ und $L = 50$ unterhalb allen anderen. Mit dieser Parameterkombination wird der verhältnismäßig geringste Wert von 0.0846 bei den durchschnittlichen mittleren Prognosefehlern über die Zeit erzielt (vgl. erste „Zeile“). Da eine konstante Verteilung vorliegt, sollte intuitiv $\lambda = 0.5$ am besten sein, da mit dieser Lernrate alle Beobachtungen gleich stark gewichtet werden. Die heuristische Methode zur Adaption der Lernrate funktioniert offenbar bei einer konstanten Verteilung nicht besonders gut. Die Lernrate wird im Mittel immer weiter vergrößert, obwohl dies nicht nötig bzw. für die Prognosegüte eher nachteilig ist. Daher kann bei *OLDC adaptive* die Erweiterung auf dieser Datensituation teilweise niedrigere Prognosefehler generieren als die ursprüngliche Methode.

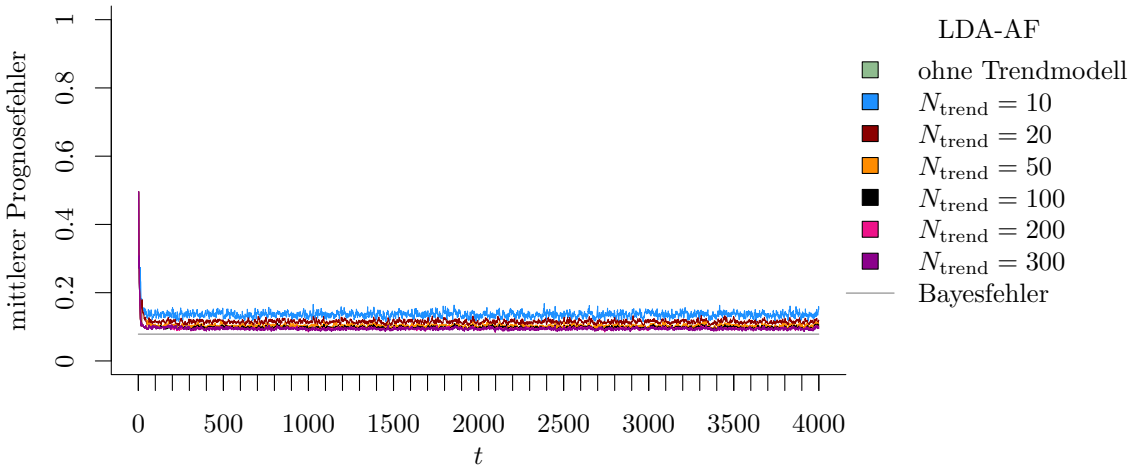
Die Befürchtung, dass die Varianz des Prognosefehlers bei Integration der Trendmodelle auf Datensituationen ohne concept drift im Gegensatz zu den ursprünglichen nicht-erweiterten Methoden systematisch zunimmt, wird anhand der Ergebnisse nicht bestätigt. Lediglich für kleine Fenster N_{trend} zur Anpassung der lokalen linearen Regressionsmodelle steigt die Varianz des Prognosefehlers zunächst leicht gegenüber jener der ursprünglichen Methoden an (vgl. Tabelle 9.17). Für steigendes N_{trend} sinkt die Varianz jedoch für alle Methoden wieder auf das Niveau der Ursprungsmethode. Dies liegt daran, dass mehr aktualisierte (und teilweise gewichtete) Mittelwerte in die einzelnen Regressionsmodelle einfließen und somit (auch bei einem nicht-linearen Trend) eine stabilere Schätzung und demnach „Prognose“ zukünftiger Erwartungswerte möglich ist.



(a) **Sequential ILDA** und Erweiterung durch lokale lineare Regressionsmodelle.

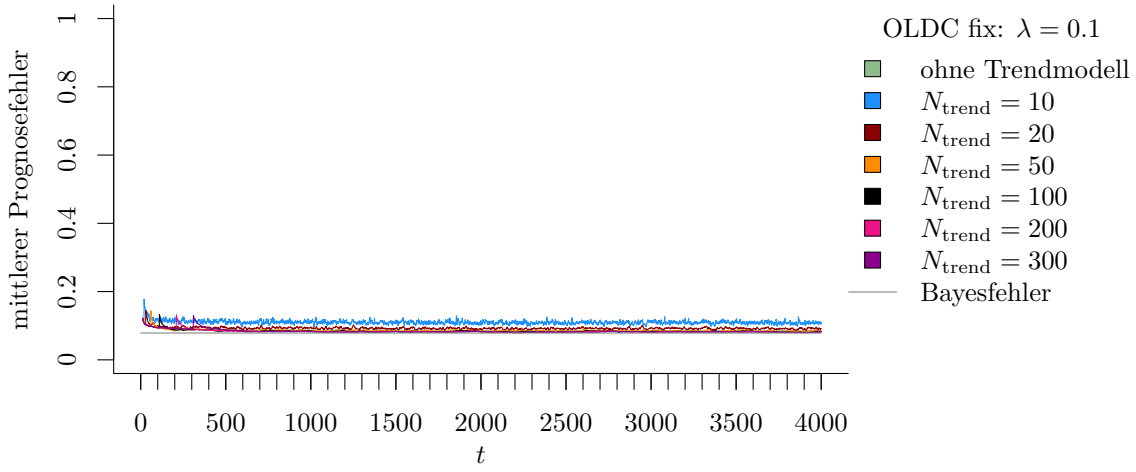


(b) **QDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

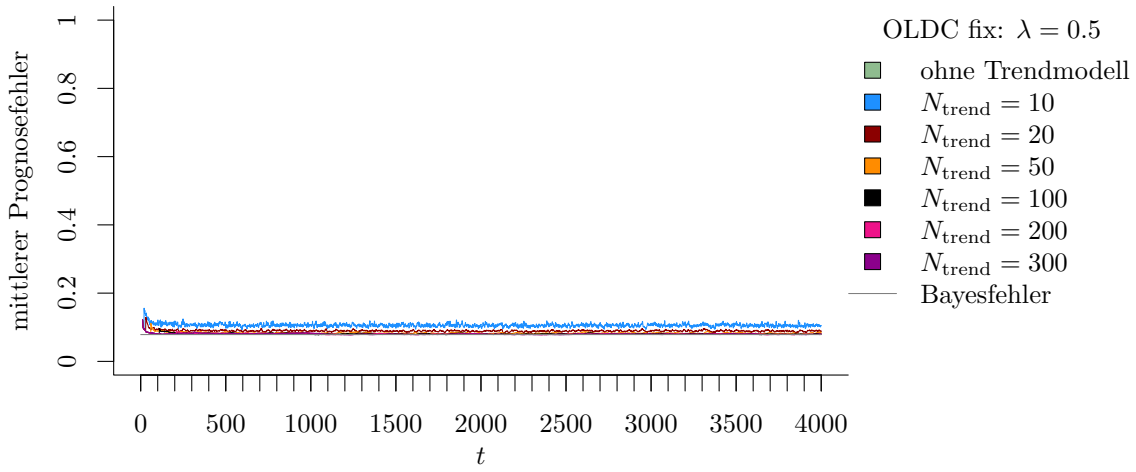


(c) **LDA-AF** und Erweiterung durch lokale lineare Regressionsmodelle.

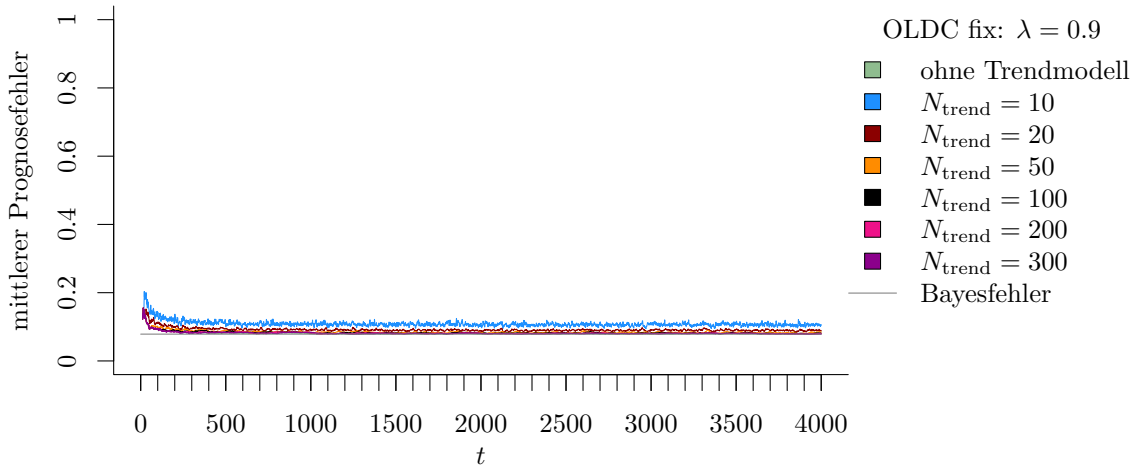
Abbildung 9.73: Mittlerer Prognosefehler über die Zeit für verschiedene Methoden und Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **ohne Drift** im zweidimensionalen Raum.



(a) **OLDC fix** mit $\lambda = 0.1$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC fix** mit $\lambda = 0.5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC fix** mit $\lambda = 0.9$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung 9.74: Mittlerer Prognosefehler über die Zeit für *OLDC* mit verschiedenen festen Lernraten λ und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **ohne Drift** im zweidimensionalen Raum.

Tabelle 9.17: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **ohne Drift ($p = 2$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	***0.0799*** (0.001)	0.1028 (0.002)	0.0990 (0.002)	0.1	0.0871 (0.001)	5	0.1029 (0.002)
						20	0.0900 (0.001)
						50	**0.0846** (0.001)
				0.3	0.0804 (0.001)		
				0.5	***0.0799*** (0.001)	5	0.1033 (0.002)
						20	0.0914 (0.001)
						50	0.0857 (0.001)
				0.7	***0.0799*** (0.001)		
				0.9	0.0806 (0.001)	5	0.1047 (0.002)
						20	0.0970 (0.002)
						50	0.0926 (0.002)
10	*0.1061* (0.002)	0.1385 (0.006)	0.1364 (0.006)	0.1	0.1101 (0.003)	5	0.1464 (0.008)
						20	0.1243 (0.005)
						50	0.1129 (0.003)
				0.3	0.1062 (0.002)		
				0.5	*0.1061* (0.002)	5	0.1475 (0.008)
						20	0.1278 (0.005)
						50	0.1159 (0.004)
				0.7	0.1066 (0.002)		
				0.9	0.1089 (0.003)	5	0.1499 (0.009)
						20	0.1379 (0.007)
						50	0.1290 (0.006)
20	*0.0896* (0.001)	0.1200 (0.003)	0.1151 (0.003)	0.1	0.0928 (0.001)	5	0.1200 (0.004)
						20	0.1050 (0.002)
						50	0.0962 (0.001)
				0.3	0.0897 (0.001)		
				0.5	*0.0896* (0.001)	5	0.1207 (0.004)
						20	0.1074 (0.002)
						50	0.0984 (0.002)
				0.7	0.0900 (0.001)		
				0.9	0.0919 (0.001)	5	0.1223 (0.004)
						20	0.1143 (0.003)
						50	0.1081 (0.003)
50	*0.0832* (0.001)	0.1079 (0.002)	0.1030 (0.002)	0.1	0.0859 (0.001)	5	0.1036 (0.002)
						20	0.0953 (0.001)
						50	0.0910 (0.001)
				0.3	*0.0832* (0.001)		
				0.5	*0.0832* (0.001)	5	0.1039 (0.002)
						20	0.0965 (0.001)
						50	0.0927 (0.001)
				0.7	0.0835 (0.001)		
				0.9	0.0851 (0.001)	5	0.1049 (0.002)
						20	0.1005 (0.002)
						50	0.0987 (0.002)
100	*0.0811* (0.001)	0.1020 (0.002)	0.0980 (0.002)	0.1	0.0840 (0.001)	5	0.0970 (0.001)
						20	0.0911 (0.001)
						50	0.0891 (0.001)
				0.3	0.0813 (0.001)		
				0.5	*0.0811* (0.001)	5	0.0972 (0.001)
						20	0.0916 (0.001)
						50	0.0897 (0.001)
				0.7	0.0814 (0.001)		
				0.9	0.0829 (0.001)	5	0.0982 (0.001)
						20	0.0946 (0.001)
						50	0.0933 (0.001)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	*0.0803* (0.001)	0.0989 (0.002)	0.0953 (0.002)	0.1	0.0835 (0.001)	5	0.0936 (0.001)
						20	0.0882 (0.001)
						50	0.0868 (0.001)
				0.3	0.0804 (0.001)		
				0.5	*0.0803* (0.001)	5	0.0938 (0.001)
						20	0.0884 (0.001)
						50	0.0865 (0.001)
				0.7	0.0805 (0.001)		
				0.9	0.0819 (0.001)	5	0.0949 (0.001)
						20	0.0912 (0.001)
						50	0.0898 (0.001)
300	*0.0801* (0.001)	**0.0979** (0.002)	**0.0945** (0.001)	0.1	0.0835 (0.001)	5	0.0926 (0.001)
						20	0.0869 (0.001)
						50	0.0858 (0.001)
				0.3	0.0802 (0.001)		
				0.5	*0.0801* (0.001)	5	0.0928 (0.001)
						20	0.0872 (0.001)
						50	0.0853 (0.001)
				0.7	0.0802 (0.001)		
				0.9	0.0814 (0.001)	5	0.0940 (0.001)
						20	0.0901 (0.001)
						50	0.0886 (0.001)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0786 (Standardabweichung 0)

Tabelle 9.18 fasst die durchschnittlichen euklidischen Abstände zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren zusammen. Für diese Datensituation ist nicht interessant, ob die Ergebnisse durch Erweiterung der Methoden verbessert werden können. Da kein Drift vorliegt, liefert die einfache Update-Methode *ILDA* bzw. *OLDC* mit fester Lernrate $\lambda = 0.5$ die besten Ergebnisse, da die Erwartungswertvektoren erwartungstreu geschätzt werden können bzw. dies vermutet wird (vgl. Kapitel 6.2.1 und 6.3.1). Da der euklidische Abstand durch 0 beschränkt ist, sind die Werte 0.0056 (Klasse 1) und 0.0070 (Klasse 2) für diese Methoden bereits recht gering. Durch die Erweiterung werden die durchschnittlichen euklidischen Abstände zwischen wahren und mittleren geschätzten (ursprüngliche Methoden) bzw. mittleren prognostizierten (erweiterte Methoden) Erwartungswertvektoren über die Zeit zunächst für alle Methoden größer. Jedoch sinken die Werte mit steigender Fenstergröße N_{trend}, was darauf zurückzuführen ist, dass mehr Beobachtungen in die einzelnen lokalen linearen Regressionsmodelle einfließen, diese stabiler werden und dadurch die folgenden Erwartungswerte nicht „fälschlicherweise“ linear in eine falsche Richtung prognostiziert werden. Für *QDA-AF* und *LDA-AF* sinken die Ergebnisse ab $N_{\text{trend}} = 100$ sogar unter jene der jeweils ursprünglichen Methode.

Auch bei *OLDC adaptive* können die durchschnittlichen euklidischen Abstände über die Zeit der ursprünglichen Methode für verschiedene Parameterkombinationen durch die Erweiterung unterschritten werden. Bei kleinen Fenstern $L = 5$ zur Adaption der Lernrate tritt beispielsweise eine Verbesserung ab etwa $N_{\text{trend}} = 100$ ein und für $L = 20$ ab etwa $N_{\text{trend}} = 200$. Wie bei den Prognosefehlern ist auffällig, dass die Ergebnisse der ursprünglichen Methode verhältnismäßig groß sind. Die euklidischen Abstände liegen deutlich über den entsprechenden Werten bei fester Lernrate λ. Dies lässt sich wieder durch den nichtoptimalen kontinuierlichen Anstieg der adaptiven Lernrate erklären.

Tabelle 9.18: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **ohne Drift** ($p = 2$) getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	***0.0056*** (0.003)	0.0526 (0.099)	0.0530 (0.107)	0.1	0.0384 (0.082)	5	0.0713 (0.145)
	0.0070 (0.003)	0.0522 (0.100)	0.0535 (0.106)		0.0755 (0.069)		0.0735 (0.144)
						20	0.0531 (0.087)
							0.0553 (0.085)
						50	0.0433 (0.056)
							0.0450 (0.055)
				0.3	0.0062 (0.007)		
					0.0183 (0.006)		
				0.5	***0.0056*** (0.003)	5	0.0733 (0.145)
					0.0070 (0.003)		0.0742 (0.146)
						20	0.0598 (0.098)
							0.0586 (0.098)
						50	0.0493 (0.066)
							0.0443 (0.064)
				0.7	0.0084 (0.004)		
					0.0084 (0.004)		
				0.9	0.0157 (0.011)	5	0.0745 (0.152)
					0.0164 (0.011)		0.0756 (0.153)
						20	0.0659 (0.128)
							0.0664 (0.128)
						50	0.0592 (0.117)
							0.0587 (0.116)
10	0.1025 (0.345)	0.1302 (0.617)	0.1305 (0.625)	0.1	*0.1020* (0.342)	5	0.1503 (0.870)
	0.1030 (0.345)	0.1322 (0.642)	0.1331 (0.646)		*0.1025* (0.341)		0.1520 (0.840)
						20	0.1264 (0.566)
							0.1278 (0.563)
						50	0.1121 (0.423)
							0.1124 (0.423)
				0.3	0.1022 (0.343)		
					0.1027 (0.343)		
				0.5	0.1025 (0.345)	5	0.1508 (0.881)
					0.1030 (0.345)		0.1530 (0.856)
						20	0.1310 (0.621)
							0.1325 (0.617)
						50	0.1170 (0.486)
							0.1172 (0.473)
				0.7	0.1031 (0.350)		
					0.1038 (0.352)		
				0.9	0.1061 (0.378)	5	0.1528 (0.915)
					0.1073 (0.391)		0.1551 (0.898)
						20	0.1418 (0.774)
							0.1435 (0.753)
						50	0.1319 (0.661)
							0.1333 (0.657)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
20	0.0679 (0.126)	0.0943 (0.267)	0.0949 (0.273)	0.1	*0.0673* (0.124)	5	0.1096 (0.347)
	0.0676 (0.127)	0.0955 (0.267)	0.0961 (0.272)		*0.0670* (0.125)		0.1130 (0.347)
						20	0.0968 (0.271)
							0.0975 (0.268)
						50	0.0815 (0.194)
							0.0810 (0.194)
				0.3	0.0675 (0.125)		
					0.0672 (0.126)		
				0.5	0.0679 (0.126)	5	0.1092 (0.348)
					0.0676 (0.127)		0.1135 (0.350)
						20	0.0994 (0.293)
							0.1005 (0.290)
						50	0.0852 (0.223)
							0.0849 (0.221)
				0.7	0.0687 (0.129)		
					0.0684 (0.130)		
				0.9	0.0720 (0.144)	5	0.1096 (0.355)
					0.0719 (0.145)		0.1137 (0.357)
						20	0.1056 (0.332)
							0.1075 (0.332)
						50	0.0975 (0.295)
							0.0979 (0.291)
50	0.0430 (0.049)	0.0658 (0.122)	0.0659 (0.124)	0.1	*0.0425* (0.048)	5	0.0760 (0.148)
	0.0417 (0.049)	0.0655 (0.124)	0.0658 (0.124)		*0.0414* (0.048)		0.0764 (0.148)
						20	0.0832 (0.144)
							0.0829 (0.145)
						50	0.0727 (0.164)
							0.0732 (0.163)
				0.3	0.0426 (0.048)		
					0.0414 (0.048)		
				0.5	0.0430 (0.049)	5	0.0750 (0.146)
					0.0417 (0.049)		0.0762 (0.148)
						20	0.0817 (0.149)
							0.0818 (0.150)
						50	0.0751 (0.186)
							0.0750 (0.184)
				0.7	0.0438 (0.051)		
					0.0426 (0.051)		
				0.9	0.0478 (0.061)	5	0.0743 (0.147)
					0.0462 (0.062)		0.0746 (0.149)
						20	0.0775 (0.151)
							0.0784 (0.153)
						50	0.0770 (0.195)
							0.0790 (0.192)
100	0.0321 (0.026)	0.0505 (0.067)	0.0491 (0.067)	0.1	*0.0313* (0.027)	5	0.0564 (0.075)
	0.0297 (0.026)	0.0478 (0.065)	0.0491 (0.064)		0.0299 (0.027)		0.0557 (0.074)
						20	0.0685 (0.092)
							0.0686 (0.089)
						50	0.0785 (0.152)
							0.0746 (0.159)
				0.3	0.0315 (0.026)		
					0.0293 (0.026)		
				0.5	0.0321 (0.026)	5	0.0555 (0.075)
					0.0297 (0.026)		0.0546 (0.074)
						20	0.0668 (0.087)
							0.0655 (0.087)
						50	0.0805 (0.154)
							0.0753 (0.158)
				0.7	0.0331 (0.028)		
					0.0308 (0.028)		
				0.9	0.0373 (0.037)	5	0.0550 (0.074)
					0.0348 (0.036)		0.0533 (0.074)
						20	0.0600 (0.081)
							0.0581 (0.079)
						50	0.0684 (0.118)
							0.0687 (0.118)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	0.0248 (0.015)	0.0366 (0.035)	0.0359 (0.035)	0.1	0.0241 (0.019)	5	0.0401 (0.038)
	0.0238 (0.015)	0.0349 (0.033)	0.0348 (0.033)		0.0253 (0.019)		0.0395 (0.036)
						20	0.0510 (0.055)
							0.0535 (0.052)
						50	0.0634 (0.085)
							0.0597 (0.089)
				0.3	*0.0238* (0.015)		
					0.0237 (0.015)		
				0.5	0.0248 (0.015)	5	0.0395 (0.037)
					0.0238 (0.015)		0.0382 (0.036)
						20	0.0489 (0.047)
							0.0510 (0.046)
						50	0.0648 (0.077)
							0.0575 (0.077)
				0.7	0.0263 (0.017)		
					0.0253 (0.016)		
				0.9	0.0303 (0.023)	5	0.0396 (0.037)
					0.0293 (0.022)		0.0375 (0.036)
						20	0.0424 (0.041)
							0.0425 (0.039)
						50	0.0509 (0.054)
							0.0503 (0.052)
300	0.0215 (0.011)	**0.0318** (0.023)	**0.0326** (0.023)	0.1	0.0218 (0.018)	5	0.0358 (0.024)
	0.0217 (0.011)	**0.0302** (0.022)	**0.0296** (0.022)		0.0239 (0.018)		0.0329 (0.024)
						20	0.0458 (0.038)
							0.0460 (0.036)
						50	0.0558 (0.061)
							0.0523 (0.070)
				0.3	*0.0204* (0.012)		
					0.0215 (0.012)		
				0.5	0.0215 (0.011)	5	0.0349 (0.024)
					0.0217 (0.011)		0.0320 (0.024)
						20	0.0434 (0.030)
							0.0441 (0.030)
						50	0.0567 (0.055)
							0.0488 (0.051)
				0.7	0.0229 (0.013)		
					0.0232 (0.013)		
				0.9	0.0268 (0.017)	5	**0.0348** (0.024)
					0.0276 (0.017)		**0.0316** (0.024)
						20	0.0392 (0.026)
							0.0366 (0.026)
						50	0.0448 (0.035)
							0.0423 (0.034)

Fazit: Generell lässt sich zusammenfassen, dass insgesamt geringe Schwankungen des Prognosefehlers für alle ursprünglichen Methoden sowie erweiterten Methoden vorliegen. Die Prognosefehler weichen nicht auffällig vom Bayesfehler ab. Wie zu erwarten ist in dieser Datensituation die Update-Methode für Diskriminanzanalyse mit identischer Gewichtung aller Beobachtungen (*Sequential ILDA* und *OLDC* mit fester Lernrate $\lambda = 0.5$) am besten. Allerdings wird die Prognosegüte durch Integration lokaler linearer Regressionsmodelle zur Modellierung des (hier nicht vorhandenen bzw. „konstanten“ Trends) auch nicht allzu stark „verschlechtert“.

Für *OLDC* mit adaptiver Lernrate, *QDA-AF* und *LDA-AF* kann der durchschnittliche mittlere Prognosefehler über die Zeit durch die Einbindung von Regressionsmodellen auf entsprechend großen Fenstern N_{trend} sogar leicht verringert werden, obwohl gar kein concept drift vorliegt. Die Varianz des Prognosefehlers steigt auch allenfalls bei kleinem N_{trend}

der erweiterten Methoden zunächst leicht an, fällt jedoch für wachsende Fensterbreite N_{trend} für die lokalen linearen Regressionsmodelle wieder auf das Niveau der Ursprungsmethoden zurück. Die Repräsentation der Prognosegüte durch die durchschnittlichen euklidischen Abstände zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren über die Zeit lässt eine ähnliche Schlussfolgerung zu.

Insgesamt wird die Prognosegüte durch die Erweiterung der Methoden im Falle einer stabilen Verteilung nicht auffällig verschlechtert. Für praktische Anwendungen ist dies von Vorteil, da die Erweiterung somit auch in Fällen Anwendung finden kann, in denen nicht klar ist, ob ein concept drift vorliegt oder sich Phasen von concept drift und Phasen mit stabiler Verteilung im Datenstrom abwechseln.

9.6.3 Ergebnisse der Datensituationen in $p = 3$ und $p = 10$

Klassifikationsprobleme werden mit steigender Anzahl an Dimensionen p schwieriger. Dies wird unter anderem an den durchschnittlichen mittleren Prognosefehlern über die Zeit deutlich. Tabellen 9.21–9.24 (vgl. Seite 370 ff.) enthalten beispielhaft die Ergebnisse für die Datensituationen „Kreisen“ und „Vorbeilaufen“ (gerade) in $p = 3$ und $p = 10$ Dimensionen. Entsprechende Tabellen für die übrigen Datensituationen sind in Anhang C.1 (Tabellen C.1–C.12) zu finden.

Für größeres p kann bei den meisten Datensituationen der Bayesfehler durch die Update-Methoden bzw. ihre Erweiterungen nicht mehr so gut approximiert werden wie im zweidimensionalen Raum. Dies lässt den Schluss zu, dass auch die einzelnen Prognosefehler zu jedem Zeitpunkt etwas größer sind. Hierbei ist zu beachten, dass die Bayesfehler aufgrund der Konstruktion der Daten (vgl. Abschnitt 9.2) nicht bei jeder Datensituation für alle Dimensionen $p \in \{2, 3, 10\}$ identisch sind. Dies ist lediglich bei den Datensituationen „Kreisen“, „Vorbeilaufen“ (gerade), Gradual Drift mit „Austausch“, Sudden Drift und ohne Drift der Fall. Bei diesen sind bei steigender Anzahl der Dimensionen p die zusätzlichen Elemente der Erwartungswertvektoren zu jedem Zeitpunkt in beiden Klassen identisch. Aufgrund symmetrischer Kovarianzmatrizen ändert sich daher bei mehr Dimensionen nichts an der „Überschneidung“ der Verteilungen beider Klassen. Die Datensituation Gradual Drift mit „Kreuzen“ weist für jedes p zu jedem Zeitpunkt einen Bayesfehler nahe Null auf.

Bei den übrigen Datensituationen sind die Bayesfehler bei der Betrachtung verschiedener Anzahlen an Dimensionen p verschieden. Bei steigender Anzahl der Dimensionen unterscheiden sich auch die weiteren Elemente der Erwartungswertvektoren beider Klassen. In den Datensituationen „Kreuzen“ und Gradual Drift mit „Kreuzen“ sind lediglich die Bayesfehler im Fall $p \in \{2, 3\}$ identisch. In der Datensituationen „Vorbeilaufen“ unterscheiden sie sich für alle betrachteten Dimensionen. In Tabelle 9.19 sind alle durchschnittlichen Bayesfehler über den gesamten Datenstrom inklusive Standardabweichungen für alle Datensituationen und Dimensionen vergleichend aufgeführt, für welche der Bayesfehler numerisch in angemessener Zeit ermittelt werden konnte.

Tabelle 9.19: Durchschnittlicher Bayesfehler über den gesamten Datenstrom (sofern er numerisch in angemessener Zeit ermittelt werden konnte). Standardabweichung in Klammern.

Datensituation	$p = 2$	$p = 3$	$p = 10$
„Kreisen“	0.0786 (0.000)	0.0786 (0.000)	0.0786 (0.000)
„Kreuzen“	0.0564 (0.116)	0.0564 (0.116)	
„Vorbeilaufen“	0.0076 (0.017)	0.0039 (0.010)	
„Vorbeilaufen“ (gerade)	0.0222 (0.042)	0.0222 (0.042)	0.0222 (0.042)
Gradual Drift mit „Kreuzen“	0.0000 (0.000)	0.0000 (0.000)	0.0000 (0.000)
Gradual Drift mit „Austausch“	0.2893 (0.122)	0.2893 (0.122)	0.2893 (0.122)
Sudden Drift	0.0786 (0.000)	0.0786 (0.000)	0.0786 (0.000)
ohne Drift	0.0786 (0.000)	0.0786 (0.000)	0.0786 (0.000)

Bei den Datensituationen mit identischem Bayesfehler für jede Anzahl an Dimensionen p kann analysiert werden, wie sich die Approximation dieses Fehlers durch den Prognosefehler mit steigender Anzahl der Dimensionen p verändert. Je mehr Dimensionen betrachtet werden, desto höher liegen die durchschnittlichen mittleren Prognosefehler über die Zeit der ursprünglichen Methoden für die meisten Datensituationen (vgl. beispielhaft jeweils erste „Zeile“ aus Tabellen 9.21/9.22 mit Tabelle 9.5 auf Seite 263 f. bzw. erste „Zeile“ aus Tabellen 9.23/9.24 mit Tabelle 9.11 auf Seite 308 f.). Eine Ausnahme bilden hier *Sequential ILDA* und *OLDC fix*. Bei diesen beiden Methoden verändert sich der Prognosefehler qualitativ bei mehr Dimensionen p kaum. Der Anstieg ist zumindest weniger stark als bei den anderen Methoden. Zusätzlich kann die durchschnittliche Streuung des Prognosefehlers bei *ILDA* mit steigendem p abnehmen.

Auch bei den erweiterten Methoden durch lokale lineare Regressionsmodelle sind die durchschnittlichen Prognosefehler über die Zeit bei größerem $p \in \{3, 10\}$ höher als bei $p = 2$. Bei schmalen Fensterbreiten N_{trend} für die Regressionsmodelle liegt der durchschnittliche Prognosefehler über die Zeit bei mehr Dimensionen bereits vergleichsweise etwas höher. Mit steigendem p erfolgt demnach zunächst ein höherer (absoluter und relativer) Anstieg des Prognosefehlers im Vergleich zu jenem im zweidimensionalen Problem (vgl. Zeile $N_{\text{trend}} = 10$ in den Tabellen pro Datensituation). Mit wachsender Fensterbreite N_{trend} für die Regressionsmodelle nehmen die Unterschiede zu den Ergebnissen mit $p = 2$ wieder stärker ab. Es erfolgt also ein schnellerer Abfall des durchschnittlichen Prognosefehlers über die Zeit mit steigendem N_{trend} bei mehr Dimensionen p.

Die Veränderung des durchschnittlichen Prognosefehlers in Abhängigkeit der Parameter N_{trend} sowie λ bzw. λ_{start} und L bei *OLDC* ist bei $p \in \{3, 10\}$ Dimensionen ähnlich wie im zweidimensionalen Problem. Für *ILDA* und *OLDC fix* mit kleiner Lernrate λ wird der durchschnittliche Prognosefehler durch Einführung lokaler linearer Regressionsmodelle direkt verkleinert. Für *QDA-AF*, *LDA-AF*, *OLDC* mit hoher fester Lernrate sowie *OLDC adaptive* erfolgt für ein schmales Fenster N_{trend} teilweise zunächst eine Erhöhung entgegen der Ergebnisse der ursprünglichen Methoden. Mit wachsender Fensterbreite N_{trend} zur Anpassung der Regressionsmodelle sinkt der durchschnittliche Prognosefehler jedoch bei allen Methoden je nach Datensituation bis zu einem gewissen N_{trend}. Im Falle eines linearen Trends der Erwartungswertvektoren („Kreuzen“, „Vorbeilaufen“, „Vorbeilaufen“ (gerade)) sinkt der Prognosefehler mit steigendem N_{trend} auch bei $p = 3$ und $p = 10$ Dimensionen.

Generell ist es jedoch so, dass bei steigender Anzahl an Dimensionen p der Bayesfehler durch den Prognosefehler der erweiterten Methoden nicht mehr so gut approximiert werden kann, da das Problem schwieriger wird (vgl. Tabelle 9.20). In der Datensituation „Kreisen“ beträgt der Bayesfehler beispielsweise zu jedem Zeitpunkt (und somit auch der durchschnittliche Bayesfehler) 0.0786. Der minimale erzielte durchschnittliche Prognosefehler in der Simulationsstudie steigt mit wachsender Dimension p und beträgt 0.0928 bei $p = 2$, 0.1061 bei $p = 3$ und 0.1220 bei $p = 10$ Dimensionen. Der Bayesfehler kann demnach immer schlechter approximiert werden.

Tabelle 9.20: Minimaler durchschnittlicher Prognosefehler über die Zeit (*min.*) für alle Datensituationen und alle betrachteten Anzahlen an Dimensionen p inklusive der jeweiligen Parameterkombination. Zum Vergleich ist der jeweilige durchschnittliche Bayesfehler über die Zeit (*BF*) aufgeführt.

Datensituation	BF	min.	Parameterkombination
$p = 2$			
„Kreisen“	0.0786	0.0928	*ILDA* ($N_{\text{trend}} = 20$)
„Kreuzen“	0.0564	0.0576	*OLDC* ($N_{\text{trend}} = 300$; $\lambda = 0.1$)
„Vorbeilaufen“	0.0076	0.0082	*OLDC* ($N_{\text{trend}} = 300$; $\lambda_{\text{start}} = 0.9$, $L = 20$)
„Vorbeilaufen“ (gerade)	0.0222	0.0233	*OLDC* ($N_{\text{trend}} = 300$; $\lambda = 0.9$)
Gradual Drift m. „Kreuzen“	0.0000	0.0040	*QDA-AF* ($N_{\text{trend}} = 200$)
Gradual Drift m. „Austausch“	0.2893	0.2958	*OLDC* ($N_{\text{trend}} = 300$; $\lambda = 0.3$)
Sudden Drift	0.0786	0.0938	*OLDC* (ohne; $\lambda_{\text{start}} = 0.1$, $L = 20$)
ohne Drift	0.0786	0.0799	*ILDA* (ohne)
$p = 3$			
„Kreisen“	0.0786	0.1061	*OLDC* ($N_{\text{trend}} = 20$; $\lambda = 0.7$)
„Kreuzen“	0.0564	0.0587	*OLDC* ($N_{\text{trend}} = 300$; $\lambda = 0.9$)
„Vorbeilaufen“	0.0039	0.0045	*OLDC* (ohne; $\lambda_{\text{start}} = 0.9$, $L = 5$)
„Vorbeilaufen“ (gerade)	0.0222	0.0236	*OLDC* ($N_{\text{trend}} = 300$; $\lambda = 0.9$)
Gradual Drift m. „Kreuzen“	0.0000	0.0036	*QDA-AF* ($N_{\text{trend}} = 200$)
Gradual Drift m. „Austausch“	0.2893	0.2982	*OLDC* ($N_{\text{trend}} = 300$; $\lambda = 0.3$)
Sudden Drift	0.0786	0.0976	*OLDC* ($N_{\text{trend}} = 50$; $\lambda = 0.3$)
ohne Drift	0.0786	0.0787	*OLDC* (ohne; $\lambda = 0.7$)
$p = 10$			
„Kreisen“	0.0786	0.1220	*LDA-AF* ($N_{\text{trend}} = 100$)
„Kreuzen“		0.0294	*LDA-AF* ($N_{\text{trend}} = 300$)
„Vorbeilaufen“		0.0000	
„Vorbeilaufen“ (gerade)	0.0222	0.0253	*OLDC* ($N_{\text{trend}} = 300$; $\lambda = 0.9$)
Gradual Drift m. „Kreuzen“	0.0000	0.0018	*QDA-AF* (alle)
Gradual Drift m. „Austausch“	0.2893	0.3130	*OLDC* ($N_{\text{trend}} = 300$; $\lambda = 0.3$)
Sudden Drift	0.0786	0.1193	*OLDC* ($N_{\text{trend}} = 100$; $\lambda = 0.3$)
ohne Drift	0.0786	0.0831	*OLDC* (ohne; $\lambda = 0.5$)

Eine Auffälligkeit besteht darin, dass sich bei steigender Dimension p des Klassifikationsproblems die Abhängigkeit des Prognosefehlers von freien Parametern der Methoden erhöhen kann. Während sich in den Datensituationen mit $p = 2$ nach Erweiterung der Methoden eine Abhängigkeit von λ bei *OLDC* deutlich verringert, ist eine solche deutliche Verringerung bei Problemen in höherer Dimension p nicht erkennbar. In der Datensituation „Kreuzen“ ist beispielsweise auch nach Erweiterung von *OLDC* immer eine Wahl von $\lambda = 0.9$ am besten (vgl. Tabellen C.1/C.7 in Anhang C.1). Bei $p = 2$ sind die Prognosefehler hingegen nahezu unabhängig von der Lernrate (vgl. Tabelle 9.7 auf Seite 279 f.). In der Datensituation Gradual Drift mit „Kreuzen“ ist bei der erweiterten Methode *OLDC* bei $p = 3$ zunächst eine niedrige Lernrate $\lambda = 0.1$ besser (vgl. Tabelle C.3). Erst mit steigender Fensterbreite N_{trend} ist der durchschnittliche Prognosefehler auch bei großer Lernrate ähnlich niedrig. Bei $p = 10$ ist bei den betrachteten Fensterbreiten N_{trend} immer eine kleine Lernrate $\lambda = 0.1$ am besten (vgl. Tabelle C.9). Der Prognosefehler steigt mit wachsender Lernrate λ. In der Datensituation mit $p = 2$ Dimensionen resultieren jedoch durch eine große Lernrate $\lambda = 0.9$ kleinere Prognosefehler (vgl. Tabelle 9.13 auf Seite 325 f.).

Auch die Abhängigkeit von den Parametern λ_{start} und L bei *OLDC* mit adaptiver Lernrate kann sich in höherer Dimension p ändern. Während in der Datensituation „Kreuzen“ bei $p = 2$ Dimensionen hauptsächlich eine Abhängigkeit des Prognosefehlers von der Fensterbreite L zu erkennen ist (vgl. Tabelle 9.7 auf Seite 279 f.), hat bei steigender Anzahl an Dimensionen p auch der Startwert λ_{start} einen stärkeren Einfluss auf die Ergebnisse. So sinken bei $p = 3$ für eine Vielzahl der Parameterkombinationen die durchschnittlichen mittleren Prognosefehler mit wachsendem Startwert (vgl. Tabelle C.1 in Anhang C.1). Bei $p = 10$ ist für $N_{\text{trend}} \leq 100$ tendenziell ein „kurvenförmiger“ Verlauf der Ergebnisse in Abhängigkeit des Startwertes λ_{start} zu erkennen, wobei die geringsten Prognosefehler jeweils für einen hohen Startwert $\lambda_{\text{start}} = 0.9$ erzielt werden (vgl. Tabelle C.7). Bei der Datensituation ohne Drift auf $p = 10$ Dimensionen steigt der durchschnittliche mittlere Prognosefehler bei Betrachtung eines Startwertes von $\lambda_{\text{start}} = 0.9$ für ein breites Fenster $L = 50$ zur Adaption der Lernrate im Vergleich zu $L = 20$ wieder. Bei $p = 2$ und $p = 3$ Dimensionen sinkt der Fehler hingegen mit wachsender Fensterbreite L (vgl. Tabelle 9.17 auf Seite 357 f. sowie Tabellen C.6/C.12 in Anhang C.1). All dies lässt den Schluss zu, dass insbesondere bei höherer Anzahl an Dimensionen p die Wahl von freien Parametern der Methoden ausschlaggebend für die Prognosegüte sein kann. In der Praxis sollten die Parameter daher am besten eine Zeit lang auf dem Datenstrom optimiert werden, sofern dies möglich ist.

Mit der Methode *QDA-AF* und ihrer Erweiterung resultieren in höher-dimensionalen Klassifikationsproblemen teilweise deutlich höhere Prognosefehler. Hervorzuheben sind dabei insbesondere die Ergebnisse auf den Datensituationen „Kreisen“ (vgl. Tabellen 9.21/9.22 mit Tabelle 9.5 auf Seite 263 f.) oder auch Sudden Drift. Als Fazit lässt sich sagen, dass *QDA-AF* insbesondere bei großem p nur angewandt werden sollte, wenn vermutet wird, dass nicht-lineare Trennungen nötig sind. In diesem Falle ist die Methode als Online Variante der Quadratischen Diskriminanzanalyse auch in hohen Dimensionen den anderen Methoden sehr stark überlegen (vgl. Tabellen C.3/C.9 in Anhang C.1 und Tabelle 9.13 auf Seite 325 f. zur Datensituation Gradual Drift mit „Kreuzen“ sowie Tabelle 9.20).

Bei höherer Dimension p wird häufig ein breiteres Fenster N_{trend} für die einzelnen Regressionsmodelle benötigt, damit der Prognosefehler möglichst minimiert bzw. der Bayesfehler approximiert werden kann. In der Datensituation „Kreisen“ mit $p = 10$ liegt das Minimum für *ILDA* beispielsweise bei $N_{\text{trend}} = 50$ (vgl. Tabelle 9.22), während es bei $p \in \{2, 3\}$ durch $N_{\text{trend}} = 20$ erzielt wird (vgl. Tabelle 9.5 auf Seite 263 f. und Tabelle 9.21). Für *LDA-AF* ist der durchschnittliche Prognosefehler bei $p = 10$ minimal bei Fenstern der Breite $N_{\text{trend}} = 100$ anstelle $N_{\text{trend}} = 50$ im zweidimensionalen Problem oder bei $p = 3$.

Weitere Unterschiede bei verschiedenen Dimensionen resultieren für die Datensituation Gradual Drift mit „Kreuzen“. Bei der erweiterten Methode *LDA-AF* erfolgt bei $p = 3$ und $p = 10$ Dimensionen bei schmalen Fenstern N_{trend} für die Regressionsmodelle zunächst eine Vergrößerung des durchschnittlichen Prognosefehlers über die Zeit. Mit wachsender Fensterbreite N_{trend} für die Regressionsmodelle sinkt der durchschnittliche Prognosefehler

bei den betrachteten Werten jedoch und liegt bei $N_{\text{trend}} = 200$ ($p = 10$) erstmals unterhalb des Ergebnisses der ursprünglichen Methode (vgl. Tabellen C.3/C.9 in Anhang C.1). Im zweidimensionalen Klassifikationsproblem steigt der durchschnittliche mittlere Prognosefehler hingegen ab einer Fensterbreite von $N_{\text{trend}} = 20$ für wachsendes N_{trend} wieder an (vgl. Tabelle 9.13 auf Seite 325 f.). Bei *ILDA* sinkt der durchschnittliche Prognosefehler mit steigender Dimension für immer breitere Fenster N_{trend} für die Regressionsmodelle, bevor er wieder steigt. So wird bei $p = 2$ das Minimum bei $N_{\text{trend}} = 10$ erreicht (0.2360), bei $p = 3$ durch $N_{\text{trend}} = 50$ (0.2454) und bei $p = 10$ durch $N_{\text{trend}} = 300$ (0.2542). Der durchschnittliche Bayesfehler wird durch *QDA-AF* bei $p \in \{3, 10\}$ sogar etwas besser approximiert als bei $p = 2$ (vgl. Tabelle 9.20).

Die euklidischen Abstände zwischen wahren und geschätzten bzw. prognostizierten Erwartungswertvektoren der ursprünglichen Update-Methoden sind in den meisten Datensituationen bei mehr Dimensionen p deutlich höher. Beispielhafte entsprechende Tabellen 9.25–9.28 für die Datensituationen „Kreisen“ und „Vorbeilaufen“ (gerade) auf $p = 3$ und $p = 10$ Dimensionen sind ebenfalls am Ende des Abschnittes zu finden. Die Tabellen der übrigen Datensituationen sind in Anhang C.2 zusammengefasst (Tabellen C.13–C.20). Bei den Datensituationen „Kreisen“, Sudden Drift und ohne Drift ist nur ein leichter Anstieg der euklidischen Abstände bei höheren Dimensionen p festzustellen. (Bei der Datensituation „Kreisen“ sind die Werte in höherer Dimension für *OLDC fix* teilweise sogar niedriger.) In den Datensituationen mit linearem Trend der Erwartungswertvektoren („Kreuzen“, „Vorbeilaufen“ und „Vorbeilaufen“ (gerade)) ist diese Zunahme für steigendes p jedoch teilweise sehr stark. Insbesondere bei *QDA-AF* (ohne Erweiterung) sind die euklidischen Abstände im Falle von $p = 10$ Dimensionen im Vergleich zu $p = 2$ Dimensionen teilweise zehnfach größer (vgl. Tabellen 9.8 auf Seite 282 ff. und C.17 zu „Kreuzen“ sowie Tabellen 9.10 auf Seite 298 ff. und C.18 zu „Vorbeilaufen“). Auch in der Datensituation „Vorbeilaufen“ (gerade) sind die euklidischen Abstände der ursprünglichen Methode *QDA-AF* bei $p = 10$ deutlich erhöht (vgl. Tabellen 9.12 auf Seite 312 ff. und 9.28 auf Seite 387 ff.).

Auf der anderen Seite resultieren ohne Drift bei $p = 10$ Dimensionen mit der ursprünglichen Methode *QDA-AF* sogar geringere euklidische Abstände als bei $p = 2$ Dimensionen (vgl. Tabellen 9.18 auf Seite 359 ff. und C.20). Auch *OLDC* mit adaptiver Lernrate liefert in einigen Datensituationen für bestimmte Parameterkombinationen geringere euklidische Abstände bei höherer Dimension p. So zum Beispiel teilweise bei einem Fenster der Breite $L = 50$ in den Datensituationen „Kreisen“ und „Vorbeilaufen“ (gerade) (vgl. Tabellen 9.25/9.26 mit Tabelle 9.6 auf Seite 267 ff. und Tabellen 9.27/9.28 mit Tabelle 9.12 auf Seite 312 ff.).

In der Datensituation „Vorbeilaufen“ ist für die Ergebnisse von *OLDC adaptive* (ohne Erweiterung) zudem auffällig, dass die euklidischen Abstände in höherer Dimension $p = 10$ viel weniger vom Startwert λ_{start} abhängen. Da bei $p = 2$ die Abstände mit wachsender Fensterbreite L zur Adaption zunehmen, erfolgt demnach für schmalere Fenster L in

höherer Dimension eine stärkere Zunahme im Vergleich zu den Ergebnissen bei $p = 2$ (vgl. Tabellen 9.10 auf Seite 298 ff. und C.18).

Für die erweiterten Methoden ist die Abhängigkeit der euklidischen Abstände von der Fensterbreite N_{trend} für die Regressionsmodelle sowie die teilweise (schwache) Abhängigkeit von den Parametern λ, λ_{start} und L bei höheren Dimensionen vergleichbar mit den Ergebnissen in $p = 2$. Der Verlauf ist demnach ähnlich. Durch die Integration lokaler linearer Regressionsmodelle zur Modellierung und Prognose des Trends der Erwartungswertvektoren können die euklidischen Abstände in allen Datensituationen stark reduziert werden. Tendenziell sind jedoch alle euklidischen Abstände bei höherer Dimension p vergleichsweise größer.

Für Situationen mit einem linearen Trend der Erwartungswertvektoren bedeutet dies, dass mit wachsendem p ein breiteres Fenster N_{trend} für die Regressionsmodelle benötigt wird, um entsprechend kleine euklidische Abstände zu erzielen und demnach die wahren Erwartungswertvektoren möglichst gut schätzen zu können (vgl. dazu z. B. Tabelle 9.12 auf Seite 312 ff. und Tabellen 9.27/9.28 der Datensituation „Vorbeilaufen“ (gerade)). Es kann also davon ausgegangen werden, dass die Erwartungstreue der erweiterten Schätzfunktionen für alle Methoden auch bei mehr Dimensionen p gilt. Für gute Schätzungen der Erwartungswertvektoren müssen lediglich mehr Beobachtungen in die einzelnen Regressionsmodelle einfließen.

Für Situationen ohne einen strikten linearen Trend der Erwartungswertvekoren bedeutet dies, dass die lineare Approximation bei mehr Dimensionen p nicht mehr so gut greifen kann. Ein Beispiel ist die Datensituation „Kreisen“ (vgl. dazu Tabelle 9.6 auf Seite 267 ff. und Tabellen 9.25/9.26). Bei Betrachtung von $p \in \{2, 3, 10\}$ wird je nach Update-Methode für die Diskriminanzanalyse der minimale Wert für beide Klassen durch Fenster der Breite $N_{\text{trend}} = 10$ oder $N_{\text{trend}} = 20$ für die lokalen linearen Regressionsmodelle erreicht. Bei der Datensituation mit Sudden Drift wird der minimale Wert bei großem p teilweise erst für ein breiteres Fenster N_{trend} erreicht. Trotzdem liegen auch hier alle durchschnittlichen euklidischen Abstände bei größerem p vergleichsweise höher (vgl. Tabellen 9.16 auf Seite 348 ff. und C.15/C.19). Für breitere Fenster N_{trend} steigen die euklidischen Abstände wieder, weshalb sie für höhere p vergleichsweise immer etwas größer sind. In diesem Fall ist es nicht möglich die euklidischen Abstände durch mehr Beobachtungen für die Regressionsmodelle bzw. ein breiteres Fenster N_{trend} zu verbessern. Es liegt ein Trade-off vor, da bei zu breiten Fenstern N_{trend} die lineare Approximation eines nicht-linearen Trends der Erwartungswertvektoren nicht mehr greift. Demnach können durch die Erweiterung der Methoden auch bei großem p verbesserte Schätzer in die Klassifikatoren der Diskriminanzanalyse einfließen, das Klassifikationsproblem wird jedoch mit steigender Anzahl an Dimensionen generell komplexer.

Die durchschnittliche Streuung über die Zeit ist in den meisten Fällen für hohes p nicht deutlich größer, sondern im Verhältnis zu der absoluten Größe des durchschnittlichen euklidischen Abstandes vergleichbar mit den Ergebnissen bei $p = 2$ Dimensionen. In manchen

Fällen ist sogar trotz Zunahme des durchschnittlichen euklidischen Abstandes über die Zeit mit steigendem p ein Abfall der durchschnittlichen Varianz erkennbar (u. a. bei der erweiterten Methode *QDA-AF* in der Datensituation ohne Drift, vgl. Tabelle 9.18 auf Seite 359 ff. und Tabelle C.20).

Die wichtigsten Ergebnisse zusammengefasst sind die Folgenden:

- Je höher die Dimension p ist, desto höher liegen die durchschnittlichen Prognosefehler über die Zeit der ursprünglichen Methoden für die meisten Datensituationen. Am geringsten sind die Unterschiede über verschiedene Dimensionen p bei den Methoden *Sequential ILDA* und *OLDC fix*.
- Nach Erweiterung der Methoden liegen die durchschnittlichen Prognosefehler über die Zeit bei größerem p ebenfalls höher. Mit wachsender Fensterbreite N_{trend} für die Regressionsmodelle nehmen die Unterschiede zu den Ergebnissen mit $p = 2$ bei Datensituationen mit linearem Trend der Erwartungswertvektoren wieder stärker ab. Es erfolgt also ein schnellerer Abfall des durchschnittlichen Prognosefehlers mit steigendem N_{trend} bei mehr Dimensionen.
- Der Bayesfehler kann durch den Prognosefehler bei steigender Dimension p tendenziell immer schlechter approximiert werden. Bei linearem Trend der Erwartungswertvektoren werden mit wachsendem p daher breitere Fenster N_{trend} benötigt, um den Bayesfehler durch den Prognosefehler gleich gut zu approximieren. Bei nicht-linearem Trend ist dies nicht möglich, da die lineare Approximation nicht mehr greift und breitere Fenster die Ergebnisse wieder verschlechtern.
- Erwartungstreue (für die Erwartungswertvektoren der Klassen des kommenden Zeitpunktes) der prognostizierten Erwartungswertvektoren durch lineare Regressionsmodelle unter linearem Trend bleibt vermutlich auch bei mehr Dimensionen p erhalten. Allerdings werden verhältnismäßig breitere Fenster N_{trend} für die Regressionsmodelle benötigt, um den Trend der Erwartungswertvektoren gleich gut zu modellieren und die Erwartungswertvektoren daraufhin gleich gut prognostizieren zu können.
- Bei nicht-linearem Trend werden die Erwartungswertvektoren bei höherem p durch die lokalen linearen Regressionsmodelle unsicherer prognostiziert. Es entsteht ein Trade-off bei der Wahl von N_{trend} zwischen möglichst vielen Beobachtungen für jedes Regressionsmodell (N_{trend} möglichst groß) und linearer Approximation des nicht-linearen Trends (N_{trend} klein).
- Bei steigender Dimension p des Klassifikationsproblems kann sich die Abhängigkeit des Prognosefehlers von freien Parametern der Methoden erhöhen. Die Wahl dieser Parameter kann ausschlaggebend für die Prognosegüte sein und wird daher wichtiger.
- Mit der Methode *QDA-AF* und ihrer Erweiterung resultieren in höher-dimensionalen Klassifikationsproblemen teilweise deutlich höhere Prognosefehler. Daher sollte *QDA-AF* insbesondere bei großem p nur angewandt werden, wenn vermutet wird, dass nicht-lineare Trennungen nötig sind.

Tabelle 9.21: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Kreisen“ ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4986 (0.016)	0.1588 (0.005)	0.1476 (0.007)	0.1	0.4979 (0.008)	5	0.1299 (0.003)
						20	0.1257 (0.003)
						50	0.1396 (0.005)
				0.3	0.4975 (0.004)		
				0.5	0.4985 (0.016)	5	0.1284 (0.003)
						20	0.1196 (0.003)
						50	0.1320 (0.004)
				0.7	0.4713 (0.016)		
				0.9	0.3714 (0.007)	5	0.1290 (0.003)
						20	*0.1174* (0.003)
						50	0.1209 (0.003)
10	*0.1375* (0.004)	0.1710 (0.007)	0.1626 (0.008)	0.1	0.1376 (0.004)	5	0.1638 (0.008)
						20	0.1495 (0.007)
						50	0.1596 (0.008)
				0.3	0.1378 (0.004)		
				0.5	*0.1375* (0.004)	5	0.1638 (0.008)
						20	0.1492 (0.007)
						50	0.1571 (0.008)
				0.7	0.1376 (0.004)		
				0.9	0.1389 (0.004)	5	0.1658 (0.009)
						20	0.1536 (0.007)
						50	0.1538 (0.008)
20	**0.1063** (0.002)	0.1479 (0.004)	0.1358 (0.004)	0.1	0.1099 (0.002)	5	0.1354 (0.004)
						20	0.1246 (0.003)
						50	0.1412 (0.006)
				0.3	0.1070 (0.002)		
				0.5	0.1063 (0.002)	5	0.1353 (0.004)
						20	0.1239 (0.003)
						50	0.1374 (0.005)
				0.7	***0.1061*** (0.002)		
				0.9	0.1068 (0.002)	5	0.1368 (0.004)
						20	0.1271 (0.003)
						50	0.1298 (0.004)
50	0.1089 (0.001)	**0.1297** (0.003)	**0.1226** (0.003)	0.1	0.1164 (0.002)	5	0.1148 (0.002)
						20	0.1108 (0.002)
						50	0.1517 (0.010)
				0.3	0.1105 (0.001)		
				0.5	0.1089 (0.001)	5	0.1142 (0.002)
						20	0.1088 (0.002)
						50	0.1447 (0.009)
				0.7	0.1078 (0.001)		
				0.9	*0.1064* (0.001)	5	0.1151 (0.002)
						20	0.1098 (0.002)
						50	0.1232 (0.004)
100	0.1846 (0.002)	0.1320 (0.002)	0.1314 (0.004)	0.1	0.1987 (0.004)	5	0.1109 (0.002)
						20	0.1130 (0.002)
						50	0.1554 (0.008)
				0.3	0.1883 (0.002)		
				0.5	0.1846 (0.002)	5	0.1096 (0.002)
						20	0.1089 (0.002)
						50	0.1459 (0.006)
				0.7	0.1813 (0.002)		
				0.9	0.1709 (0.002)	5	0.1099 (0.002)
						20	***0.1073*** (0.002)
						50	0.1237 (0.004)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.5840 (0.004)	0.2188 (0.007)	0.2137 (0.012)	0.1	0.5997 (0.007)	5	0.1634 (0.005)
						20	0.1801 (0.006)
						50	0.2586 (0.016)
				0.3	0.5954 (0.004)		
				0.5	0.5840 (0.004)	5	0.1614 (0.005)
						20	0.1711 (0.005)
						50	0.2448 (0.015)
				0.7	0.5634 (0.004)		
				0.9	0.4911 (0.003)	5	*0.1608* (0.005)
						20	0.1632 (0.005)
						50	0.2022 (0.010)
300	0.8541 (0.001)	0.4230 (0.013)	0.4129 (0.022)	0.1	0.8577 (0.002)	5	0.3530 (0.018)
						20	0.3876 (0.015)
						50	0.5090 (0.022)
				0.3	0.8631 (0.001)		
				0.5	0.8541 (0.001)	5	0.3505 (0.018)
						20	0.3759 (0.015)
						50	0.4894 (0.021)
				0.7	0.8287 (0.001)		
				0.9	0.7562 (0.002)	5	*0.3489* (0.018)
						20	0.3616 (0.015)
						50	0.4252 (0.019)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0786 (Standardabweichung 0)

Tabelle 9.22: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Kreisen“ ($p = 10$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4985 (0.009)	0.3669 (0.005)	0.1528 (0.005)	0.1	0.4986 (0.005)	5	0.1536 (0.003)
						20	0.1613 (0.003)
						50	0.2215 (0.008)
				0.3	0.4985 (0.004)		
				0.5	0.4985 (0.009)	5	0.1522 (0.003)
						20	0.1540 (0.003)
						50	0.2051 (0.007)
				0.7	0.4766 (0.008)		
				0.9	0.3894 (0.005)	5	0.1520 (0.003)
						20	*0.1495* (0.003)
						50	0.1640 (0.004)
10	0.2546 (0.008)	0.2817 (0.008)	0.2411 (0.011)	0.1	0.2428 (0.008)	5	0.2007 (0.007)
						20	0.2046 (0.007)
						50	0.2397 (0.009)
				0.3	0.2527 (0.008)		
				0.5	0.2546 (0.008)	5	0.1998 (0.007)
						20	0.2008 (0.007)
						50	0.2320 (0.009)
				0.7	0.2540 (0.008)		
				0.9	0.2501 (0.008)	5	0.2000 (0.007)
						20	*0.1984* (0.007)
						50	0.2081 (0.007)
20	0.1759 (0.003)	0.2250 (0.005)	0.1753 (0.005)	0.1	0.1742 (0.003)	5	0.1749 (0.004)
						20	0.1695 (0.004)
						50	0.2137 (0.008)
				0.3	0.1754 (0.003)		
				0.5	0.1759 (0.003)	5	0.1741 (0.004)
						20	*0.1660* (0.004)
						50	0.2038 (0.008)
				0.7	0.1751 (0.003)		
				0.9	0.1735 (0.003)	5	0.1755 (0.004)
						20	0.1663 (0.004)
						50	0.1768 (0.005)
50	**0.1427** (0.002)	**0.1969** (0.004)	*0.1303* (0.002)	0.1	0.1531 (0.002)	5	0.1541 (0.003)
						20	0.1494 (0.003)
						50	0.2105 (0.010)
				0.3	0.1446 (0.002)		
				0.5	0.1427 (0.002)	5	0.1533 (0.003)
						20	0.1459 (0.003)
						50	0.1965 (0.008)
				0.7	0.1411 (0.002)		
				0.9	**0.1389** (0.002)	5	0.1545 (0.003)
						20	0.1460 (0.003)
						50	0.1578 (0.004)
100	0.2047 (0.002)	0.2276 (0.004)	***0.1220*** (0.002)	0.1	0.2235 (0.003)	5	0.1452 (0.002)
						20	0.1483 (0.002)
						50	0.1992 (0.007)
				0.3	0.2092 (0.002)		
				0.5	0.2047 (0.002)	5	0.1438 (0.002)
						20	0.1431 (0.002)
						50	0.1851 (0.006)
				0.7	0.2006 (0.002)		
				0.9	0.1901 (0.002)	5	0.1442 (0.002)
						20	**0.1409** (0.002)
						50	0.1514 (0.003)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	0.5797 (0.004)	0.4262 (0.005)	0.2093 (0.008)	0.1	0.5899 (0.006)	5	0.2015 (0.004)
						20	0.2200 (0.005)
						50	0.2783 (0.011)
				0.3	0.5892 (0.004)		
				0.5	0.5797 (0.004)	5	0.1994 (0.004)
						20	0.2109 (0.004)
						50	0.2612 (0.010)
				0.7	0.5607 (0.004)		
				0.9	0.4941 (0.003)	5	*0.1983* (0.004)
						20	0.2033 (0.004)
						50	0.2178 (0.006)
300	0.8323 (0.002)	0.6411 (0.008)	0.4484 (0.013)	0.1	0.8281 (0.002)	5	0.3894 (0.011)
						20	0.4198 (0.010)
						50	0.4800 (0.016)
				0.3	0.8382 (0.002)		
				0.5	0.8323 (0.002)	5	0.3870 (0.011)
						20	0.4099 (0.010)
						50	0.4629 (0.015)
				0.7	0.8106 (0.002)		
				0.9	0.7459 (0.002)	5	*0.3853* (0.011)
						20	0.3998 (0.010)
						50	0.4143 (0.011)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0786 (Standardabweichung 0)

Tabelle 9.23: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ (gerade)** ($p = 3$) getrennt nach Methoden für Online DA und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.1443 (0.001)	0.0392 (0.001)	0.0312 (0.001)	0.1	0.4328 (0.002)	5	0.0315 (0.001)
						20	***0.0252*** (0.000)
						50	0.0271 (0.000)
				0.3	0.2535 (0.002)		
				0.5	0.1443 (0.001)	5	0.0316 (0.001)
						20	0.0255 (0.000)
						50	0.0275 (0.001)
				0.7	0.0705 (0.001)		
				0.9	0.0290 (0.000)	5	0.0324 (0.001)
						20	0.0272 (0.000)
						50	0.0290 (0.001)
10	0.0518 (0.001)	0.0492 (0.002)	0.0446 (0.002)	0.1	0.0451 (0.001)	5	0.0463 (0.002)
						20	0.0389 (0.001)
						50	0.0411 (0.001)
				0.3	0.0529 (0.001)		
				0.5	0.0518 (0.001)	5	0.0463 (0.002)
						20	0.0399 (0.001)
						50	0.0411 (0.001)
				0.7	0.0433 (0.001)		
				0.9	*0.0341* (0.001)	5	0.0438 (0.002)
						20	0.0386 (0.001)
						50	0.0396 (0.002)
20	0.0419 (0.000)	0.0433 (0.001)	0.0374 (0.001)	0.1	0.0362 (0.000)	5	0.0385 (0.001)
						20	0.0324 (0.000)
						50	0.0343 (0.001)
				0.3	0.0427 (0.001)		
				0.5	0.0419 (0.000)	5	0.0383 (0.001)
						20	0.0331 (0.001)
						50	0.0342 (0.001)
				0.7	0.0349 (0.000)		
				0.9	*0.0274* (0.000)	5	0.0364 (0.001)
						20	0.0322 (0.001)
						50	0.0331 (0.001)
50	0.0383 (0.000)	0.0390 (0.001)	0.0329 (0.001)	0.1	0.0330 (0.000)	5	0.0336 (0.001)
						20	0.0299 (0.000)
						50	0.0323 (0.001)
				0.3	0.0390 (0.000)		
				0.5	0.0383 (0.000)	5	0.0333 (0.001)
						20	0.0302 (0.000)
						50	0.0318 (0.001)
				0.7	0.0319 (0.000)		
				0.9	*0.0249* (0.000)	5	0.0312 (0.001)
						20	0.0287 (0.000)
						50	0.0300 (0.001)
100	0.0372 (0.000)	0.0364 (0.001)	0.0308 (0.001)	0.1	0.0320 (0.000)	5	0.0316 (0.000)
						20	0.0290 (0.000)
						50	0.0305 (0.001)
				0.3	0.0379 (0.000)		
				0.5	0.0372 (0.000)	5	0.0312 (0.000)
						20	0.0290 (0.000)
						50	0.0298 (0.000)
				0.7	0.0310 (0.000)		
				0.9	*0.0242* (0.000)	5	0.0290 (0.000)
						20	0.0270 (0.000)
						50	0.0279 (0.000)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	0.0367 (0.000)	0.0348 (0.001)	0.0295 (0.000)	0.1	0.0315 (0.000)	5	0.0304 (0.000)
						20	0.0282 (0.000)
						50	0.0292 (0.000)
				0.3	0.0374 (0.000)		
				0.5	0.0367 (0.000)	5	0.0300 (0.000)
						20	0.0280 (0.000)
						50	0.0284 (0.000)
				0.7	0.0306 (0.000)		
				0.9	*0.0238* (0.000)	5	0.0277 (0.000)
						20	0.0258 (0.000)
						50	0.0266 (0.000)
300	**0.0365** (0.000)	**0.0342** (0.001)	**0.0291** (0.000)	0.1	0.0314 (0.000)	5	0.0301 (0.000)
						20	0.0279 (0.000)
						50	0.0287 (0.000)
				0.3	0.0372 (0.000)		
				0.5	0.0365 (0.000)	5	0.0296 (0.000)
						20	0.0275 (0.000)
						50	0.0279 (0.000)
				0.7	0.0305 (0.000)		
				0.9	***0.0236*** (0.000)	5	0.0273 (0.000)
						20	0.0254 (0.000)
						50	0.0261 (0.000)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0222 (Standardabweichung 0.042)

Tabelle 9.24: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ (gerade)** ($p = 10$) getrennt nach Methoden für Online DA und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.1455 (0.001)	0.0912 (0.004)	0.0313 (0.000)	0.1	0.4280 (0.002)	5	0.0356 (0.001)
						20	0.0296 (0.000)
						50	0.0335 (0.001)
				0.3	0.2546 (0.002)		
				0.5	0.1455 (0.001)	5	0.0348 (0.000)
						20	***0.0294*** (0.000)
						50	0.0329 (0.001)
				0.7	0.0714 (0.001)		
				0.9	0.0301 (0.000)	5	0.0357 (0.001)
						20	0.0314 (0.000)
						50	0.0338 (0.001)
10	0.0969 (0.004)	0.1010 (0.005)	0.0653 (0.003)	0.1	0.0858 (0.003)	5	0.0642 (0.002)
						20	0.0616 (0.002)
						50	0.0647 (0.002)
				0.3	0.0990 (0.004)		
				0.5	0.0969 (0.004)	5	0.0631 (0.002)
						20	0.0608 (0.002)
						50	0.0627 (0.002)
				0.7	0.0787 (0.003)		
				0.9	0.0541 (0.002)	5	0.0544 (0.002)
						20	*0.0520* (0.002)
						50	0.0539 (0.002)
20	0.0600 (0.001)	0.0830 (0.004)	0.0460 (0.001)	0.1	0.0556 (0.001)	5	0.0518 (0.001)
						20	0.0452 (0.001)
						50	0.0494 (0.001)
				0.3	0.0616 (0.001)		
				0.5	0.0600 (0.001)	5	0.0502 (0.001)
						20	0.0445 (0.001)
						50	0.0478 (0.001)
				0.7	0.0491 (0.001)		
				0.9	*0.0364* (0.001)	5	0.0459 (0.001)
						20	0.0411 (0.001)
						50	0.0439 (0.001)
50	0.0452 (0.000)	0.0756 (0.004)	0.0354 (0.001)	0.1	0.0433 (0.000)	5	0.0451 (0.001)
						20	0.0388 (0.001)
						50	0.0439 (0.001)
				0.3	0.0465 (0.000)		
				0.5	0.0452 (0.000)	5	0.0436 (0.001)
						20	0.0381 (0.001)
						50	0.0425 (0.001)
				0.7	0.0375 (0.000)		
				0.9	*0.0291* (0.000)	5	0.0410 (0.001)
						20	0.0367 (0.001)
						50	0.0400 (0.001)
100	0.0408 (0.000)	0.0714 (0.004)	0.0307 (0.000)	0.1	0.0395 (0.000)	5	0.0408 (0.001)
						20	0.0363 (0.000)
						50	0.0406 (0.001)
				0.3	0.0420 (0.000)		
				0.5	0.0408 (0.000)	5	0.0393 (0.001)
						20	0.0355 (0.000)
						50	0.0389 (0.001)
				0.7	0.0340 (0.000)		
				0.9	*0.0269* (0.000)	5	0.0372 (0.001)
						20	0.0341 (0.000)
						50	0.0366 (0.001)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	0.0387 (0.000)	0.0680 (0.004)	0.0278 (0.000)	0.1	0.0376 (0.000)	5	0.0378 (0.000)
						20	0.0342 (0.000)
						50	0.0379 (0.001)
				0.3	0.0398 (0.000)		
				0.5	0.0387 (0.000)	5	0.0363 (0.000)
						20	0.0331 (0.000)
						50	0.0361 (0.001)
				0.7	0.0323 (0.000)		
				0.9	*0.0257* (0.000)	5	0.0342 (0.000)
						20	0.0316 (0.000)
						50	0.0338 (0.001)
300	**0.0380** (0.000)	**0.0667** (0.003)	**0.0268** (0.000)	0.1	0.0371 (0.000)	5	0.0368 (0.000)
						20	0.0332 (0.000)
						50	0.0366 (0.001)
				0.3	0.0391 (0.000)		
				0.5	0.0380 (0.000)	5	0.0352 (0.000)
						20	0.0321 (0.000)
						50	0.0348 (0.001)
				0.7	0.0318 (0.000)		
				0.9	***0.0253*** (0.000)	5	0.0331 (0.000)
						20	0.0304 (0.000)
						50	0.0325 (0.001)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0222 (Standardabweichung 0.042)

Tabelle 9.25: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Kreisen“ ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	1.9936 (0.007)	0.6427 (0.189)	0.6145 (0.212)	0.1	2.2177 (0.144)	5	0.2741 (0.201)
	1.9960 (0.007)	0.6518 (0.189)	0.6178 (0.208)		2.2533 (0.116)		0.2806 (0.201)
						20	0.4768 (0.181)
							0.4849 (0.180)
						50	0.5969 (0.263)
							0.6038 (0.262)
				0.3	2.0348 (0.014)		
					2.0399 (0.013)		
				0.5	1.9936 (0.007)	5	0.2597 (0.197)
					1.9960 (0.007)		0.2649 (0.198)
						20	0.4247 (0.166)
							0.4328 (0.166)
						50	0.5651 (0.244)
							0.5675 (0.242)
				0.7	1.9536 (0.010)		
					1.9554 (0.010)		
				0.9	1.7830 (0.026)	5	*0.2376* (0.201)
					1.7842 (0.026)		*0.2416* (0.202)
						20	0.3287 (0.169)
							0.3337 (0.168)
						50	0.4230 (0.209)
							0.4266 (0.207)
10	**0.2005** (0.403)	*0.1728* (0.716)	0.1736 (0.732)	0.1	0.2018 (0.399)	5	0.1943 (0.981)
	0.2035 (0.404)	*0.1728* (0.740)	0.1744 (0.753)		0.2045 (0.400)		0.1944 (0.995)
						20	0.1787 (0.791)
							0.1786 (0.795)
						50	0.1875 (0.927)
							0.1860 (0.963)
				0.3	0.2013 (0.401)		
					0.2041 (0.401)		
				0.5	0.2005 (0.403)	5	0.1933 (0.985)
					0.2035 (0.404)		0.1934 (1.001)
						20	0.1787 (0.806)
							0.1785 (0.808)
						50	0.1867 (0.929)
							0.1831 (0.891)
				0.7	0.1991 (0.410)		
					0.2024 (0.411)		
				0.9	**0.1930** (0.441)	5	0.1946 (1.012)
					0.1977 (0.444)		0.1943 (1.029)
						20	0.1824 (0.872)
							0.1825 (0.885)
						50	0.1821 (0.898)
							0.1817 (0.876)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
20	0.3392 (0.154)	**0.1630** (0.302)	**0.1610** (0.311)	0.1	0.3430 (0.152)	5	0.1600 (0.385)
	0.3439 (0.152)	**0.1657** (0.298)	**0.1628** (0.305)		0.3477 (0.150)		0.1600 (0.383)
						20	0.1703 (0.394)
							0.1727 (0.384)
						50	0.2015 (0.655)
							0.2025 (0.660)
				0.3	0.3415 (0.153)		
					0.3462 (0.150)		
				0.5	0.3392 (0.154)	5	0.1578 (0.385)
					0.3439 (0.152)		0.1578 (0.382)
						20	0.1634 (0.372)
							0.1659 (0.364)
						50	0.1952 (0.591)
							0.1959 (0.589)
				0.7	0.3342 (0.158)		
					0.3391 (0.155)		
				0.9	0.3113 (0.173)	5	0.1553 (0.390)
					0.3170 (0.170)		***0.1545*** (0.387)
						20	***0.1544*** (0.379)
							0.1558 (0.367)
						50	0.1693 (0.446)
							0.1726 (0.438)
50	0.8529 (0.098)	0.3694 (0.147)	0.3555 (0.147)	0.1	0.8602 (0.097)	5	0.2628 (0.167)
	0.8580 (0.098)	0.3767 (0.148)	0.3613 (0.148)		0.8652 (0.097)		0.2657 (0.169)
						20	0.3679 (0.240)
							0.3739 (0.238)
						50	0.4915 (2.222)
							0.4952 (2.251)
				0.3	0.8574 (0.096)		
					0.8624 (0.097)		
				0.5	0.8529 (0.098)	5	0.2589 (0.166)
					0.8580 (0.098)		0.2611 (0.168)
						20	0.3496 (0.200)
							0.3545 (0.199)
						50	0.4776 (1.817)
							0.4738 (1.820)
				0.7	0.8429 (0.101)		
					0.8483 (0.102)		
				0.9	0.7918 (0.112)	5	*0.2499* (0.165)
					0.7980 (0.113)		*0.2516* (0.168)
						20	0.3077 (0.184)
							0.3114 (0.184)
						50	0.3838 (0.771)
							0.3910 (0.769)
100	1.6803 (0.072)	0.9074 (0.189)	0.8763 (0.197)	0.1	1.6834 (0.072)	5	0.6921 (0.130)
	1.6855 (0.072)	0.9186 (0.195)	0.8857 (0.198)		1.6890 (0.072)		0.6990 (0.130)
						20	0.8872 (0.216)
							0.8960 (0.212)
						50	1.2373 (2.230)
							1.2378 (2.306)
				0.3	1.6821 (0.070)		
					1.6874 (0.071)		
				0.5	1.6803 (0.072)	5	0.6866 (0.130)
					1.6855 (0.072)		0.6922 (0.129)
						20	0.8530 (0.186)
							0.8594 (0.185)
						50	1.2002 (1.825)
							1.1899 (1.786)
				0.7	1.6745 (0.077)		
					1.6801 (0.077)		
				0.9	1.6207 (0.093)	5	*0.6736* (0.128)
					1.6275 (0.094)		*0.6784* (0.127)
						20	0.7801 (0.157)
							0.7855 (0.158)
						50	0.9987 (0.860)
							1.0029 (0.824)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	2.8784 (0.052)	2.1866 (0.256)	2.1305 (0.260)	0.1	2.8424 (0.056)	5	1.8847 (0.105)
	2.8830 (0.052)	2.1998 (0.271)	2.1426 (0.263)		2.8488 (0.055)		1.8919 (0.108)
						20	2.1562 (0.178)
							2.1665 (0.186)
						50	2.7628 (1.026)
							2.7515 (1.000)
				0.3	2.8559 (0.051)		
					2.8610 (0.051)		
				0.5	2.8784 (0.052)	5	1.8792 (0.104)
					2.8830 (0.052)		1.8860 (0.107)
						20	2.1118 (0.161)
							2.1203 (0.161)
						50	2.6924 (0.909)
							2.6791 (0.928)
				0.7	2.9173 (0.058)		
					2.9227 (0.058)		
				0.9	2.9844 (0.078)	5	*1.8674* (0.103)
					2.9916 (0.080)		*1.8737* (0.106)
						20	2.0209 (0.134)
							2.0280 (0.136)
						50	2.3783 (0.526)
							2.3791 (0.546)
300	3.1237 (0.039)	3.0538 (0.175)	2.9934 (0.170)	0.1	3.0393 (0.049)	5	2.8111 (0.089)
	3.1258 (0.038)	3.0634 (0.183)	3.0042 (0.171)		3.0439 (0.046)		2.8169 (0.091)
						20	3.0284 (0.134)
							3.0343 (0.140)
						50	3.4637 (0.475)
							3.4568 (0.465)
				0.3	3.0707 (0.039)		
					3.0732 (0.037)		
				0.5	3.1237 (0.039)	5	2.8112 (0.089)
					3.1258 (0.038)		2.8167 (0.090)
						20	3.0045 (0.126)
							3.0099 (0.126)
						50	3.4298 (0.435)
							3.4177 (0.443)
				0.7	3.2206 (0.046)		
					3.2236 (0.044)		
				0.9	3.4632 (0.067)	5	*2.8079* (0.088)
					3.4682 (0.067)		*2.8132* (0.090)
						20	2.9459 (0.110)
							2.9518 (0.112)
						50	3.2269 (0.288)
							3.2262 (0.298)

Tabelle 9.26: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Kreisen“ ($p = 10$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	1.9974 (0.007)	1.8942 (0.054)	1.0225 (0.144)	0.1	2.2476 (0.077)	5	0.5265 (0.099)
	1.9981 (0.007)	1.8779 (0.060)	1.0284 (0.149)		2.2346 (0.067)		0.5302 (0.098)
						20	0.7337 (0.112)
							0.7395 (0.112)
						50	1.0025 (0.351)
							1.0031 (0.345)
				0.3	2.0414 (0.013)		
					2.0398 (0.010)		
				0.5	1.9974 (0.007)	5	0.5152 (0.096)
					1.9981 (0.007)		0.5195 (0.096)
						20	0.6960 (0.102)
							0.7003 (0.103)
						50	0.9310 (0.327)
							0.9347 (0.320)
				0.7	1.9575 (0.009)		
					1.9587 (0.010)		
				0.9	1.7869 (0.024)	5	*0.4908* (0.098)
					1.7887 (0.024)		*0.4958* (0.098)
						20	0.6272 (0.099)
							0.6323 (0.101)
						50	0.7039 (0.178)
							0.7114 (0.174)
10	**0.2910** (0.691)	**0.7024** (0.784)	**0.6517** (1.318)	0.1	***0.2908*** (0.684)	5	0.3200 (1.218)
	0.2961 (0.703)	**0.7070** (0.792)	0.6596 (1.414)		***0.2960*** (0.697)		0.3225 (1.256)
						20	0.3078 (1.108)
							0.3110 (1.144)
						50	0.3356 (1.900)
							0.3385 (2.035)
				0.3	***0.2908*** (0.686)		
					0.2960 (0.699)		
				0.5	0.2910 (0.691)	5	0.3191 (1.219)
					0.2961 (0.703)		0.3222 (1.261)
						20	0.3067 (1.103)
							0.3101 (1.132)
						50	0.3294 (1.779)
							0.3312 (1.779)
				0.7	0.2914 (0.701)		
					0.2963 (0.713)		
				0.9	0.2935 (0.764)	5	0.3208 (1.255)
					0.2970 (0.767)		0.3236 (1.284)
						20	0.3090 (1.139)
							0.3121 (1.166)
						50	0.3159 (1.334)
							0.3181 (1.359)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
20	0.3636 (0.184)	0.8128 (0.237)	0.6579 (0.477)	0.1	0.3656 (0.181)	5	0.2549 (0.402)
	0.3663 (0.185)	0.8134 (0.238)	**0.6587** (0.485)		0.3685 (0.182)		0.2544 (0.404)
						20	0.2641 (0.473)
							0.2624 (0.481)
						50	0.3011 (1.253)
							0.3007 (1.236)
				0.3	0.3648 (0.182)		
					0.3677 (0.183)		
				0.5	0.3636 (0.184)	5	0.2526 (0.398)
					0.3663 (0.185)		0.2531 (0.401)
						20	0.2581 (0.410)
							0.2574 (0.412)
						50	0.2876 (1.008)
							0.2873 (1.003)
				0.7	0.3607 (0.188)		
					0.3631 (0.189)		
				0.9	0.3460 (0.210)	5	***0.2513*** (0.402)
					0.3473 (0.211)		***0.2512*** (0.406)
						20	0.2522 (0.392)
							0.2523 (0.399)
						50	0.2580 (0.511)
							0.2552 (0.511)
50	0.8564 (0.082)	1.2527 (0.122)	0.8758 (0.175)	0.1	0.8614 (0.080)	5	0.3686 (0.176)
	0.8586 (0.082)	1.2506 (0.123)	0.8745 (0.176)		0.8638 (0.080)		0.3659 (0.178)
						20	0.4530 (0.259)
							0.4522 (0.264)
						50	0.5579 (5.014)
							0.5573 (4.875)
				0.3	0.8595 (0.081)		
					0.8619 (0.080)		
				0.5	0.8564 (0.082)	5	0.3649 (0.174)
					0.8586 (0.082)		0.3637 (0.176)
						20	0.4387 (0.206)
							0.4388 (0.207)
						50	0.5164 (3.346)
							0.5152 (3.328)
				0.7	0.8481 (0.086)		
					0.8499 (0.085)		
				0.9	0.8013 (0.099)	5	*0.3566* (0.175)
					0.8018 (0.098)		*0.3549* (0.176)
						20	0.4125 (0.186)
							0.4124 (0.188)
						50	0.4205 (0.476)
							0.4217 (0.486)
100	1.6826 (0.067)	2.0657 (0.116)	1.4263 (0.160)	0.1	1.6828 (0.064)	5	0.8529 (0.117)
	1.6836 (0.066)	2.0719 (0.116)	1.4498 (0.166)		1.6839 (0.064)		0.8504 (0.118)
						20	1.0194 (0.176)
							1.0181 (0.178)
						50	1.1963 (2.408)
							1.1971 (2.293)
				0.3	1.6827 (0.065)		
					1.6839 (0.065)		
				0.5	1.6826 (0.067)	5	0.8487 (0.116)
					1.6836 (0.066)		0.8477 (0.118)
						20	0.9969 (0.158)
							0.9963 (0.155)
						50	1.1542 (1.752)
							1.1575 (1.774)
				0.7	1.6789 (0.071)		
					1.6793 (0.071)		
				0.9	1.6289 (0.087)	5	*0.8363* (0.115)
					1.6274 (0.086)		*0.8347* (0.117)
						20	0.9516 (0.137)
							0.9511 (0.138)
						50	0.9912 (0.305)
							0.9927 (0.316)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	2.8800 (0.051)	3.4026 (0.122)	2.6858 (0.218)	0.1	2.8425 (0.049)	5	2.1141 (0.109)
	2.8790 (0.050)	3.3947 (0.120)	2.6956 (0.212)		2.8415 (0.051)		2.1124 (0.107)
						20	2.3453 (0.172)
							2.3461 (0.168)
						50	2.6406 (0.683)
							2.6450 (0.679)
				0.3	2.8562 (0.049)		
					2.8552 (0.049)		
				0.5	2.8800 (0.051)	5	2.1111 (0.109)
					2.8790 (0.050)		2.1096 (0.106)
						20	2.3198 (0.156)
							2.3194 (0.147)
						50	2.5778 (0.580)
							2.5768 (0.572)
				0.7	2.9219 (0.058)		
					2.9206 (0.056)		
				0.9	2.9942 (0.079)	5	*2.1024* (0.109)
					2.9916 (0.075)		*2.1008* (0.107)
						20	2.2687 (0.136)
							2.2687 (0.131)
						50	2.3435 (0.264)
							2.3453 (0.262)
300	3.1234 (0.038)	3.7701 (0.113)	3.3605 (0.147)	0.1	3.0404 (0.039)	5	*3.0104* (0.098)
	3.1198 (0.038)	3.7603 (0.107)	3.3726 (0.148)		3.0371 (0.041)		*3.0076* (0.096)
						20	3.1900 (0.140)
							3.1869 (0.136)
						50	3.3951 (0.410)
							3.3976 (0.407)
				0.3	3.0701 (0.036)		
					3.0666 (0.037)		
				0.5	3.1234 (0.038)	5	3.0114 (0.098)
					3.1198 (0.038)		3.0088 (0.095)
						20	3.1833 (0.133)
							3.1814 (0.125)
						50	3.3661 (0.367)
							3.3625 (0.357)
				0.7	3.2233 (0.044)		
					3.2198 (0.043)		
				0.9	3.4720 (0.067)	5	3.0124 (0.098)
					3.4681 (0.064)		3.0099 (0.096)
						20	3.1582 (0.120)
							3.1570 (0.114)
						50	3.2173 (0.197)
							3.2164 (0.192)

Tabelle 9.27: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ (gerade)** ($p = 3$) getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	5.0161 (0.010)	0.3396 (0.795)	*0.1538* (0.122)	0.1	9.0016 (0.142)	5	1.7545 (0.296)
	5.0269 (0.010)	0.3668 (0.857)	*0.1581* (0.121)		9.0675 (0.195)		1.7883 (0.325)
						20	2.1459 (0.276)
							2.1821 (0.309)
						50	2.6220 (0.554)
							2.6507 (0.577)
				0.3	7.0171 (0.015)		
					7.0407 (0.018)		
				0.5	5.0161 (0.010)	5	0.9454 (0.128)
					5.0269 (0.010)		0.9536 (0.130)
						20	1.0743 (0.067)
							1.0860 (0.067)
						50	1.3839 (0.241)
							1.3925 (0.238)
				0.7	3.0153 (0.011)		
					3.0192 (0.011)		
				0.9	1.0152 (0.020)	5	0.2255 (0.111)
					1.0172 (0.019)		0.2281 (0.110)
						20	0.2388 (0.058)
							0.2464 (0.057)
						50	0.3313 (0.072)
							0.3356 (0.071)
10	0.1287 (0.400)	0.1631 (0.710)	0.1650 (0.740)	0.1	*0.1281* (0.396)	5	0.1554 (0.641)
	0.1291 (0.397)	0.1664 (0.715)	0.1674 (0.735)		*0.1284* (0.393)		0.1559 (0.652)
						20	0.1325 (0.425)
							0.1330 (0.426)
						50	0.1326 (0.438)
							0.1333 (0.433)
				0.3	0.1283 (0.397)		
					0.1286 (0.394)		
				0.5	0.1287 (0.400)	5	0.1563 (0.650)
					0.1291 (0.397)		0.1567 (0.662)
						20	0.1358 (0.448)
							0.1362 (0.452)
						50	0.1341 (0.455)
							0.1352 (0.451)
				0.7	0.1295 (0.406)		
					0.1301 (0.403)		
				0.9	0.1332 (0.438)	5	0.1612 (0.699)
					0.1342 (0.434)		0.1626 (0.707)
						20	0.1450 (0.532)
							0.1463 (0.537)
						50	0.1408 (0.517)
							0.1424 (0.516)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
20	0.0839 (0.139)	0.1177 (0.295)	0.1187 (0.307)	0.1	*0.0832* (0.136)	5	0.1101 (0.254)
	0.0844 (0.138)	0.1195 (0.302)	0.1204 (0.303)		*0.0836* (0.136)		0.1101 (0.252)
						20	0.0908 (0.164)
							0.0906 (0.165)
						50	0.0892 (0.169)
							0.0901 (0.169)
				0.3	0.0835 (0.137)		
					0.0839 (0.136)		
				0.5	0.0839 (0.139)	5	0.1112 (0.257)
					0.0844 (0.138)		0.1107 (0.255)
						20	0.0944 (0.180)
							0.0938 (0.178)
						50	0.0907 (0.177)
							0.0919 (0.177)
				0.7	0.0850 (0.142)		
					0.0855 (0.142)		
				0.9	0.0892 (0.158)	5	0.1159 (0.278)
					0.0899 (0.158)		0.1159 (0.277)
						20	0.1039 (0.222)
							0.1041 (0.220)
						50	0.0972 (0.204)
							0.0995 (0.204)
50	0.0534 (0.052)	0.0813 (0.135)	0.0825 (0.136)	0.1	*0.0528* (0.051)	5	0.0753 (0.109)
	0.0530 (0.053)	0.0806 (0.139)	0.0809 (0.137)		*0.0524* (0.051)		0.0738 (0.109)
						20	0.0664 (0.078)
							0.0626 (0.080)
						50	0.0654 (0.114)
							0.0626 (0.114)
				0.3	0.0530 (0.051)		
					0.0525 (0.051)		
				0.5	0.0534 (0.052)	5	0.0760 (0.109)
					0.0530 (0.053)		0.0738 (0.110)
						20	0.0697 (0.086)
							0.0653 (0.086)
						50	0.0659 (0.115)
							0.0644 (0.114)
				0.7	0.0544 (0.055)		
					0.0544 (0.055)		
				0.9	0.0593 (0.066)	5	0.0805 (0.121)
					0.0590 (0.066)		0.0786 (0.122)
						20	0.0764 (0.107)
							0.0744 (0.108)
						50	0.0696 (0.117)
							0.0709 (0.116)
100	0.0393 (0.028)	0.0607 (0.073)	0.0632 (0.072)	0.1	0.0393 (0.030)	5	0.0557 (0.060)
	0.0401 (0.029)	0.0592 (0.074)	0.0567 (0.073)		0.0405 (0.029)		0.0541 (0.060)
						20	0.0543 (0.051)
							0.0509 (0.052)
						50	0.0533 (0.087)
							0.0515 (0.079)
				0.3	*0.0390* (0.028)		
					0.0399 (0.028)		
				0.5	0.0393 (0.028)	5	0.0557 (0.058)
					0.0401 (0.029)		0.0529 (0.059)
						20	0.0554 (0.053)
							0.0521 (0.054)
						50	0.0522 (0.079)
							0.0522 (0.080)
				0.7	0.0405 (0.030)		
					0.0412 (0.031)		
				0.9	0.0459 (0.039)	5	0.0602 (0.066)
					0.0451 (0.039)		0.0558 (0.066)
						20	0.0596 (0.064)
							0.0574 (0.066)
						50	0.0547 (0.076)
							0.0560 (0.074)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	*0.0293* (0.016)	0.0449 (0.038)	0.0452 (0.037)	0.1	0.0317 (0.021)	5	0.0425 (0.036)
	0.0331 (0.016)	0.0457 (0.037)	0.0440 (0.037)		0.0363 (0.020)		0.0478 (0.035)
						20	0.0460 (0.036)
							0.0474 (0.037)
						50	0.0447 (0.062)
							0.0490 (0.053)
				0.3	0.0297 (0.017)		
					0.0334 (0.017)		
				0.5	*0.0293* (0.016)	5	0.0405 (0.031)
					0.0331 (0.016)		0.0446 (0.030)
						20	0.0441 (0.032)
							0.0464 (0.032)
						50	0.0425 (0.053)
							0.0480 (0.047)
				0.7	0.0304 (0.018)		
					0.0344 (0.018)		
				0.9	0.0355 (0.024)	5	0.0436 (0.035)
					0.0379 (0.024)		0.0453 (0.035)
						20	0.0458 (0.037)
							0.0487 (0.037)
						50	0.0437 (0.056)
							0.0480 (0.052)
300	***0.0254*** (0.012)	**0.0375** (0.026)	**0.0365** (0.025)	0.1	0.0301 (0.019)	5	0.0384 (0.030)
	0.0302 (0.012)	**0.0408** (0.025)	**0.0394** (0.024)		0.0365 (0.018)		0.0450 (0.028)
						20	0.0436 (0.032)
							0.0465 (0.031)
						50	0.0430 (0.052)
							0.0489 (0.047)
				0.3	0.0264 (0.013)		
					0.0309 (0.013)		
				0.5	***0.0254*** (0.012)	5	**0.0344** (0.022)
					0.0302 (0.012)		**0.0397** (0.021)
						20	0.0389 (0.024)
							0.0423 (0.023)
						50	0.0397 (0.042)
							0.0448 (0.041)
				0.7	0.0266 (0.014)		
					0.0321 (0.013)		
				0.9	0.0314 (0.018)	5	0.0370 (0.025)
					0.0361 (0.018)		0.0409 (0.024)
						20	0.0403 (0.027)
							0.0447 (0.025)
						50	0.0388 (0.036)
							0.0444 (0.035)

Tabelle 9.28: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ (gerade)** ($p = 10$) getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	5.0302 (0.008)	2.3053 (2.828)	0.2779 (0.216)	0.1	8.9685 (0.076)	5	1.7448 (0.211)
	5.0320 (0.008)	1.9861 (2.827)	0.2747 (0.208)		8.9882 (0.068)		1.7503 (0.197)
						20	2.0669 (0.221)
							2.0759 (0.200)
						50	2.8405 (0.519)
							2.8561 (0.487)
				0.3	7.0232 (0.012)		
					7.0301 (0.011)		
				0.5	5.0302 (0.008)	5	0.9843 (0.077)
					5.0320 (0.008)		0.9824 (0.077)
						20	1.1347 (0.060)
							1.1385 (0.056)
						50	1.7728 (0.434)
							1.7744 (0.433)
				0.7	3.0235 (0.010)		
					3.0239 (0.010)		
				0.9	1.0199 (0.018)	5	*0.2619* (0.059)
					1.0206 (0.018)		*0.2592* (0.058)
						20	0.3028 (0.046)
							0.3046 (0.046)
						50	0.5957 (0.248)
							0.5965 (0.242)
10	0.2490 (0.700)	0.2948 (0.867)	0.3412 (1.390)	0.1	*0.2480* (0.694)	5	0.2835 (0.961)
	0.2477 (0.696)	0.3019 (0.943)	0.3382 (1.356)		*0.2466* (0.690)		0.2805 (0.946)
						20	0.2570 (0.751)
							0.2555 (0.744)
						50	0.2570 (0.794)
							0.2552 (0.774)
				0.3	0.2484 (0.696)		
					0.2470 (0.692)		
				0.5	0.2490 (0.700)	5	0.2844 (0.969)
					0.2477 (0.696)		0.2816 (0.951)
						20	0.2609 (0.778)
							0.2593 (0.769)
						50	0.2600 (0.833)
							0.2579 (0.789)
				0.7	0.2506 (0.709)		
					0.2493 (0.705)		
				0.9	0.2579 (0.760)	5	0.2929 (1.035)
					0.2567 (0.758)		0.2906 (1.018)
						20	0.2751 (0.878)
							0.2732 (0.872)
						50	0.2707 (0.892)
							0.2691 (0.866)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
20	0.1607 (0.182)	0.2096 (0.272)	0.2572 (0.473)	0.1	*0.1595* (0.179)	5	0.1980 (0.290)
	0.1587 (0.179)	0.2168 (0.320)	0.2539 (0.463)		*0.1574* (0.176)		0.1931 (0.290)
						20	0.1724 (0.224)
							0.1706 (0.224)
						50	0.1710 (0.259)
							0.1696 (0.248)
				0.3	0.1599 (0.180)		
					0.1578 (0.177)		
				0.5	0.1607 (0.182)	5	0.1984 (0.291)
					0.1587 (0.179)		0.1938 (0.290)
						20	0.1766 (0.235)
							0.1744 (0.235)
						50	0.1740 (0.262)
							0.1718 (0.252)
				0.7	0.1626 (0.186)		
					0.1606 (0.184)		
				0.9	0.1704 (0.207)	5	0.2073 (0.315)
					0.1685 (0.206)		0.2031 (0.315)
						20	0.1906 (0.275)
							0.1880 (0.274)
						50	0.1846 (0.279)
							0.1829 (0.274)
50	0.0996 (0.062)	0.1557 (0.151)	0.1883 (0.188)	0.1	*0.0983* (0.059)	5	0.1369 (0.122)
	0.1013 (0.059)	0.1653 (0.167)	0.1896 (0.184)		*0.0997* (0.057)		0.1350 (0.119)
						20	0.1193 (0.101)
							0.1201 (0.096)
						50	0.1223 (0.291)
							0.1238 (0.265)
				0.3	0.0987 (0.060)		
					0.1002 (0.057)		
				0.5	0.0996 (0.062)	5	0.1370 (0.122)
					0.1013 (0.059)		0.1356 (0.118)
						20	0.1220 (0.104)
							0.1239 (0.100)
						50	0.1223 (0.238)
							0.1230 (0.219)
				0.7	0.1020 (0.065)		
					0.1037 (0.062)		
				0.9	0.1110 (0.079)	5	0.1460 (0.136)
					0.1128 (0.075)		0.1450 (0.132)
						20	0.1347 (0.124)
							0.1346 (0.121)
						50	0.1298 (0.173)
							0.1319 (0.162)
100	0.0719 (0.032)	0.1288 (0.090)	0.1513 (0.091)	0.1	0.0716 (0.031)	5	0.1017 (0.066)
	0.0748 (0.031)	0.1358 (0.095)	0.1562 (0.090)		*0.0733* (0.030)		0.1027 (0.066)
						20	0.0947 (0.065)
							0.0984 (0.064)
						50	0.0946 (0.148)
							0.0970 (0.151)
				0.3	*0.0712* (0.031)		
					0.0736 (0.030)		
				0.5	0.0719 (0.032)	5	0.1007 (0.065)
					0.0748 (0.031)		0.1024 (0.064)
						20	0.0949 (0.063)
							0.1003 (0.062)
						50	0.0925 (0.111)
							0.0950 (0.110)
				0.7	0.0746 (0.035)		
					0.0777 (0.034)		
				0.9	0.0839 (0.045)	5	0.1086 (0.075)
					0.0880 (0.043)		0.1122 (0.073)
						20	0.1048 (0.074)
							0.1090 (0.073)
						50	0.0989 (0.088)
							0.1059 (0.084)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	*0.0550* (0.018)	0.1097 (0.044)	0.1209 (0.044)	0.1	0.0578 (0.018)	5	0.0807 (0.037)
	0.0556 (0.018)	0.1151 (0.054)	0.1263 (0.043)		0.0553 (0.018)		0.0786 (0.036)
						20	0.0816 (0.042)
							0.0804 (0.043)
						50	0.0809 (0.073)
							0.0793 (0.066)
				0.3	0.0555 (0.017)		
					0.0545 (0.017)		
				0.5	*0.0550* (0.018)	5	0.0762 (0.034)
					0.0556 (0.018)		0.0766 (0.034)
						20	0.0772 (0.037)
							0.0798 (0.037)
						50	0.0768 (0.062)
							0.0738 (0.060)
				0.7	0.0574 (0.020)		
					0.0589 (0.020)		
				0.9	0.0662 (0.027)	5	0.0810 (0.040)
					0.0693 (0.027)		0.0852 (0.039)
						20	0.0817 (0.042)
							0.0864 (0.041)
						50	0.0783 (0.049)
							0.0829 (0.044)
300	***0.0486*** (0.012)	**0.1042** (0.031)	**0.1117** (0.028)	0.1	0.0544 (0.015)	5	0.0732 (0.027)
	0.0474 (0.013)	**0.1061** (0.034)	**0.1152** (0.028)		0.0486 (0.015)		0.0662 (0.026)
						20	0.0776 (0.033)
							0.0711 (0.033)
						50	0.0793 (0.071)
							0.0739 (0.063)
				0.3	0.0501 (0.013)		
					0.0466 (0.013)		
				0.5	***0.0486*** (0.012)	5	**0.0656** (0.023)
					0.0474 (0.013)		**0.0630** (0.023)
						20	0.0693 (0.026)
							0.0682 (0.026)
						50	0.0703 (0.055)
							0.0658 (0.050)
				0.7	0.0507 (0.014)		
					0.0503 (0.014)		
				0.9	0.0588 (0.020)	5	0.0684 (0.026)
					0.0597 (0.019)		0.0696 (0.026)
						20	0.0708 (0.028)
							0.0725 (0.028)
						50	0.0693 (0.035)
							0.0701 (0.032)

9.6.4 Zusammenfassung der Ergebnisse und Fazit

Es wurden verschiedene Datensituationen (vgl. Abschnitte 9.1 und 9.2) betrachtet, um die „Prognosegüte“ der einzelnen Update-Methoden für Diskriminanzanalyse sowie der jeweiligen Erweiterungen durch die Integration lokaler linearer Regressionsmodelle bei verschiedenen Arten und Ausprägungen von concept drift zu untersuchen. Die Prognosegüte wird dabei zum einen durch den Prognosefehler charakterisiert, zum anderen durch den euklidischen Abstand zwischen wahren und geschätzten bzw. prognostizierten Erwartungswertvektoren zur Untersuchung der Erwartungstreue der verwendeten Schätzer in den kontinuierlich aktualisierten Klassifikatoren der Methoden für Online Diskriminanzanalyse (vgl. Seite 232 ff.).

In der Simulationsstudie wird deutlich, dass die Prognosegüte der Klassifikatoren durch Erweiterung der Methoden stark verbessert werden kann, wenn die Annahme eines linearen Trends der Erwartungswertvektoren erfüllt ist. Dies ist bei den betrachteten Datensituationen „Kreuzen“ (vgl. Seite 271 ff.), „Vorbeilaufen“ (vgl. Seite 286 ff.) und „Vorbeilaufen“ (gerade) (vgl. Seite 302 ff.) der Fall.

Für alle Update-Methoden für Diskriminanzanalyse sinkt bei Integration der vorgestellten Erweiterung sowohl der durchschnittliche Prognosefehler als auch der durchschnittliche euklidische Abstand über die Zeit mit wachsender Fenstergröße N_{trend} für die lokalen linearen Regressionsmodelle zur Modellierung des Trends der Erwartungswerte und Prognose der Erwartungswertvektoren. Dies untermauert die Ergebnisse aus Kapitel 8. Dort wurde die Erwartungstreue der Schätzer $T_2^{(c),P}$, $T_2^{(c),K}$ und $T_2^{(c),A}$ für die Erwartungswertvektoren, die in die Klassifikatoren der erweiterten Methoden einfließen, für Spezialfälle bewiesen (Zusammenfassung auf Seite 209 f.). Nun wird deutlich, dass die Erwartungstreue vermutlich auch im allgemeinen Fall und unabhängig von der Dimension p gilt, sofern die Annahme eines linearen Trends der Erwartungswertvektoren in den Klassen gerechtfertigt ist. Bei höherer Dimension p werden nur breitere Fenster für die lokalen linearen Regressionsmodelle benötigt, damit die Prognose der Erwartungswertvektoren gleich gut ist. Der Bayesfehler der Datensituationen mit allen Dimensionen p kann zu vielen Zeitpunkten durch den Prognosefehler der erweiterten Methoden approximiert werden.

Da in der Praxis im Allgemeinen jedoch nicht bekannt ist, ob ein betrachteter Datenstrom einem concept drift unterliegt bzw. wenn ja, wie genau dieser concept drift aussieht, wurden auch Datensituationen mit anderen Arten von Drift betrachtet, um zu untersuchen, ob die Erweiterung der Methoden auch in solchen Fällen die Prognosegüte der Klassifikatoren verbessern kann. Untersucht wurden im Speziellen nicht-linearer Trend der Erwartungswertvektoren („Kreisen“, vgl. Seite 255 ff.) und zusätzlicher Verletzung der Normalverteilungsannahme (moving hyperplane, vgl. Seite 235 ff.), verschiedene sudden drifts (Sudden Drift, vgl. Seite 338 ff.), u. a. auch mit Verletzung der Normalverteilungsannahme (STAGGER, vgl. Seite 246 ff.) und gradual drifts statt eines incremental drifts (Gradual Drift

mit „Kreuzen“ (vgl. Seite 316 ff.), Gradual Drift mit „Austausch“ (vgl. Seite 327 ff.)) sowie eine Datensituation ohne Drift (vgl. Seite 353 ff.).

Es wird deutlich, dass in vielen Situationen eine Verbesserung der Prognosegüte der Update-Methoden durch die eingeführte Erweiterung erreicht werden kann. Die lineare Approximation greift in vielen verschiedenen Datensituationen und nicht-linearer Trend kann bis zu einem gewissen Grad lokal linear approximiert werden. Es zeigt sich außerdem, dass sich die Prognosefehler der verschiedenen erweiterten Methoden annähern, während jene der ursprünglichen Methoden stark variieren können. Bei geeigneter Wahl der Fensterbreite N_{trend} für die lokalen linearen Regressionsmodelle kann der Bayesfehler zu vielen Zeitpunkten im Datenstrom durch den Prognosefehler approximiert werden, sofern bei Betrachtung einer Update-Methode für LDA eine lineare Trennung der Daten möglich ist. Auch reduziert sich insbesondere bei Datensituationen mit wenigen Variablen (kleines p) häufig der Einfluss von den Parametern λ, λ_{start} und L bei *OLDC* nach Integration lokaler linearer Regressionsmodelle. Wenn irgendeine Art von Drift vermutet wird, kann $\lambda = 0.9$ bzw. $\lambda_{\text{start}} = 0.9$ als Startwert bei *OLDC* zur Reduzierung des Prognosefehlers empfohlen werden. Die Breite des Fensters L zur Adaption lässt sich jedoch nicht pauschal festlegen. Der optimale Wert hängt sehr von der Stärke des Drifts ab, teilweise auch noch bei den erweiterten Methoden. Dieser sollte daher mit Vorkenntnissen gewählt oder zunächst optimiert werden. In höher-dimensionalen Klassifikationsproblemen zeigt sich teilweise eine stärkere Abhängigkeit der Ergebnisse von den freien Parametern der einzelnen Update-Methoden. Die Prognosegüte kann daher von der Wahl der Parameter abhängen. Bei großem p sollten diese Parameter daher nach Möglichkeit mit Vorkenntnissen gewählt oder zunächst auf dem Datenstrom optimiert werden.

Letztendlich kann nicht pauschal eine Methode inklusive aller Parameter und die optimale Fensterbreite N_{trend} für alle möglichen Arten von Drifts und (den unendlichen Raum der) Datensituationen empfohlen werden. Die Wahl hängt auch von Vorkenntnissen, Kapazitäten und dem Analyseziel ab:

- Gibt es bereits Vorkenntnisse darüber, ob ein Drift vorliegt und wenn ja, wie dieser aussieht?
  - Bei linearem Trend der Erwartungswertvektoren sollte N_{trend} möglichst groß sein. Zudem ist meistens die Erweiterung von *OLDC* mit fester Lernrate eine gute Wahl, wobei $\lambda = 0.9$ gewählt werden kann (vgl. auch Tabelle 9.20).
  - Bei nicht-linearem Trend der Erwartungswertvektoren ist N_{trend} so zu wählen, dass die lineare Approximation noch greift.
  - Bei sudden drifts kann der Parameter N_{trend} anhand des Musters zeitlich wiederkehrender Strukturbrüche angepasst werden.
  - Wenn bekannt ist, dass die Klassen nicht-linear trennbar sind, ist *QDA-AF* unschlagbar.

- Welche Zeit- und Rechenkapazitäten stehen zur Verfügung? Sind die entsprechenden Update-Methoden für Diskriminanzanalyse auf dem jeweiligen System bereits implementiert?
  - Häufig liefert auch die Erweiterung der standardmäßigen Fisher LDA (*Sequential ILDA*) sehr gute Ergebnisse. Diese basiert lediglich auf Update-Formeln für die nötigen Größen der Fisher LDA sowie linearen Regressionsmodellen.
  - Zu beachten sind jedoch eventuelle Probleme beim Eigenwertproblem bei Vorliegen von diskreten Variablen (vgl. Seite 252). In diesem Fall sind die anderen Methoden stabiler.
  - Wenn lediglich die einfache Update-Methode der Kanonischen LDA ohne Gewichtungen wie bei *OLDC fix* implementiert ist und die Gewichtung nicht einfach hinzugefügt werden kann, entspricht dies $\lambda = 0.5$. Da in vielen Fällen nach Erweiterung von *OLDC fix* die Lernrate λ weniger wichtig wird, kann auch eine Erweiterung durch lokale lineare Regressionsmodelle der einfachen Kanonischen LDA betrachtet werden.
  - Insbesondere die Erweiterung von *OLDC adaptive* ist aufgrund der Adaption der Lernrate sehr rechenintensiv. Auch *QDA-AF* und *LDA-AF* benötigen mehr Rechenzeit als die anderen Methoden.
- Für welche Zielgruppe wird die Analyse durchgeführt? Liegt der Fokus auf einer möglichst anschaulichen Darstellung und Erklärbarkeit der Methode und Ergebnisse für den Anwender oder steht die Optimierung der Prognosegüte im Vordergrund?
  - *LDA-AF* und *QDA-AF* sind sehr komplex und nicht anschaulich zu beschreiben.
  - Auch die Adaption der Lernrate bei *OLDC adaptive* ist weniger anschaulich.
  - Anschaulich darstellbar ist die Methode *OLDC* mit fester Lernrate λ, die auf einfachen Update-Formeln für die nötigen Größen der Kanonischen LDA basiert.

Vor dem Hintergrund der jeweiligen Antworten ergeben sich unterschiedliche Präferenzen. Insgesamt kann jedoch festgehalten werden, dass es in den meisten Fällen möglich ist die Prognosegüte noch zu verbessern, wenn zuvor eine konkrete Methode für Online Diskriminanzanalyse festgelegt wurde bzw. fokussiert wird.

Wenn die Wahl der Update-Methode offen ist, kann besonders herausgestellt werden, dass sich die Unterschiede der durchschnittlichen Prognosefehler über die Zeit für die unterschiedlichen Update-Methoden nach der Erweiterung verringern. Bei vielen Datensituationen sind ohne die Erweiterung *LDA-AF* und *OLDC adaptive* mit geeigneter Wahl der beiden Parameter λ_{start} und L den anderen Methoden deutlich überlegen. Diese beiden Parameter sind im besten Fall zu optimieren. Die heuristische Vorgehensweise zur Adaption der Lernrate (vgl. Seite 75) führt dazu, dass im Datenstrom mehr Daten gespeichert werden müssen und der Rechenaufwand größer wird. Auch *LDA-AF* ist eine recht aufwändige und theoretisch komplexe Methode. Zudem kann auch die Prognosegüte dieser beiden Methoden durch die vorgestellte Erweiterung vielfach noch verbessert werden.

Die Erweiterung der Methoden durch Integration lokaler linearer Regressionsmodelle kann daher Vorteile bieten. Bei *OLDC* wird die Wahl der festen Lernrate vielfach weniger wichtig als vor der Erweiterung und kann auf einen hohen Wert $\lambda = 0.9$ festgesetzt werden, wenn ein Drift vermutet wird. Damit kann vielfach sogar eine bessere Prognosegüte erzielt werden als mit dem Aufwand einer adaptiven Lernrate. Generell ist die Wahl der entsprechenden Methode für Online Diskriminanzanalyse sowie auch die Wahl eventueller Parameter bei den erweiterten Methoden nicht mehr ganz so wichtig wie ohne Erweiterung. Die Methode kann daher nach Präferenzen und Möglichkeiten ausgewählt werden. Im Falle eines linearen Trends der Erwartungswertvektoren ist ein möglichst großes Fenster N_{trend} für die Anpassung der lokalen linearen Regressionsmodelle am besten. Es fließen somit mehr Beobachtungen in die einzelnen Modelle ein. Die Modellierung und Prognose der Erwartungswertvektoren der Klassen wird stabiler und die Varianz reduziert.

Wichtig zu beachten ist jedoch, dass in Situationen, in denen die Erwartungswertvektoren keinem linearen Trend unterliegen, ein möglichst großes Fenster der Breite N_{trend} für die Anpassung der lokalen linearen Regressionsmodelle nicht die besten Ergebnisse liefert. Vielmehr gilt es daher diesen Parameter geeignet zu wählen, da bei zu geringem N_{trend} unsichere Schätzungen aufgrund weniger Beobachtungen für die einzelnen linearen Modelle resultieren. Bei zu breitem Fenster N_{trend} fließen hingegen viele Beobachtungen in die Modelle ein, aber eventuell greift die lineare Approximation des Trends nicht mehr und die Erwartungswertvektoren werden falsch prognostiziert. Die Wahl des optimalen N_{trend} hängt bei sudden drifts insbesondere von den Abständen zwischen den einzelnen Strukturbrüchen ab, bei nicht-linearem Trend insbesondere von der Struktur des Drifts. In der Praxis könnten zum einen Vorinformationen in die Wahl von N_{trend} einfließen, zum anderen könnte der Prognosefehler eine Zeit lang im Datenstrom erfasst und der Parameter getuned werden, bevor er festgesetzt wird.

Die Befürchtung, dass im Falle keines Drifts bzw. einer stabilen Verteilung die Varianz des Prognosefehlers für die erweiterten Methoden steigt, wird nicht bestätigt. Bei allen Datensituationen steigt die Varianz des Prognosefehlers gegenüber jener der ursprünglichen Methode zunächst leicht an, wenn schmale Fenster N_{trend} für die einzelnen lokalen linearen Regressionsmodelle betrachtet werden. Mit steigendem N_{trend} sinkt die Varianz jedoch bei allen Methoden. In den Datensituationen mit Drift ist die Varianz des Prognosefehlers spätestens bei dem N_{trend}, für welches der durchschnittliche mittlere Prognosefehler der ursprünglichen Methode verringert werden kann, auch wieder auf dem ursprünglichen Niveau. In der Datensituation ohne Drift wird die Prognosegüte in Form des durchschnittlichen Prognosefehlers bei allen Methoden nicht auffallend verschlechtert, wenn N_{trend} groß genug gewählt wird. Bei *QDA-AF*, *LDA-AF* und *OLDC adaptive* wurden in der Simulationsstudie die Fehler sogar noch etwas verringert.

10 Zusammenfassung und Ausblick

Vor dem Hintergrund eines immer weiter wachsenden Datenbestandes und insbesondere der Existenz von immer mehr Datenströmen anstelle von Batch-Daten gewinnen Online-Algorithmen zur Analyse von Daten immer mehr an Bedeutung. Spezielle Charakteristiken von Datenströmen sind die beliebige (unbeschränkte) Größe sowie die nicht-kontrollierbare Reihenfolge des Auftretens der Beobachtungen (Abschnitt 2.1). Zudem sind die einzelnen Beobachtungen im Datenstrom häufig nur eine Zeit lang verfügbar und werden nicht dauerhaft gespeichert, sodass sie sequentiell herangezogen werden müssen. Auch für gängige Klassifikationsverfahren wie die Diskriminanzanalyse wird somit eine Adaption zur Online-Fähigkeit erforderlich. Aufgrund der zeitlichen Komponente besteht ein wesentliches Problem von Datenströmen zudem darin, dass sich die zugrunde liegende Verteilung der Beobachtungen im Laufe der Zeit ändern kann, sodass die Annahme einer allen Beobachtungen zugrunde liegenden identischen Verteilung anders als bei Batch-Daten verletzt ist. Für solche Situationen hat sich der Begriff *concept drift* etabliert. Die Problematik erfordert eine weitere Adaption von Online-Algorithmen an veränderliche Verteilungen. Es wurden bereits einige solcher Algorithmen für Online Diskriminanzanalyse entwickelt. Viele dieser Algorithmen haben gemein, dass zwar eine Adaption an einen concept drift ermöglicht wird, eine kontinuierlich fortschreitende Veränderung der Verteilung allerdings nicht beachtet wird, sondern lediglich die Verteilung der bis dato realisierten Beobachtungen einbezogen wird. Eine kontinuierliche Veränderung der Verteilung kann sich somit negativ auf die Prognosegüte der entsprechenden Online-Klassifikatoren auswirken. In dieser Dissertation wird daher eine Methodik zur Verbesserung der Prognosegüte von Methoden für Online Diskriminanzanalyse für Datenströme mit concept drift entwickelt.

Der Begriff *concept drift* wurde als Erstes von Schlimmer und Granger (1986) eingeführt, in der gängigen Literatur gibt es bisher jedoch keine einheitliche Definition. Daher werden die verschiedenen Formalisierungen und von verschiedenen AutorInnen eingeführten Arten und Formen von concept drift in Abschnitt 2.2 gegenübergestellt, um eine Übersicht über Gemeinsamkeiten und Unterschiede zu erhalten. Die meisten dieser Formalisierungen sind dabei von qualitativer Natur, bevor Webb et al. (2016) formale Definitionen zur Differenzierung zwischen verschiedenen Arten von concept drift vorschlugen.

In der Arbeit wird *concept* (oder dt. *Konzept*) äquivalent für den *datengenerierenden Prozess*/ die *datengenerierende Funktion*, die *zugrunde liegende Verteilung* oder formaler die Menge $\{P(Y = y_c), f_{\boldsymbol{X}}(\boldsymbol{x}), f_{\boldsymbol{X}|Y=y_c}(\boldsymbol{x}|y_c), P(Y = y_c|\boldsymbol{X} = \boldsymbol{x})\}$, $c = 1, \ldots, M$ (Klassen),

verwendet, wobei Y die Zielvariable und $\boldsymbol{X}$ den Zufallsvektor der Einflussvariablen sowie y_c und $\boldsymbol{x}$ die entsprechenden Ausprägungen bezeichnen. Jegliche Veränderung einer der Größen der oben genannten Menge wird als *concept drift* bezeichnet. Zudem erfolgt eine Differenzierung insbesondere nur zwischen den Arten *incremental drift*, *sudden drift* und *gradual drift*. Bei einem incremental drift ersetzen sich viele nur leicht unterschiedliche zugrunde liegende Verteilungen im Laufe der Zeit. Es erfolgt somit eine kontinuierliche mäßige Veränderung. Entgegen dazu zeichnet sich ein sudden drift durch einen plötzlichen Wechsel zwischen zwei sich deutlich unterscheidenden Verteilungen aus. Der gradual drift wird dadurch charakterisiert, dass zwei datengenerierende Funktionen existieren, wobei erstere nach und nach durch zweitere ersetzt wird. Mit der Zeit nimmt demnach die Wahrscheinlichkeit zu, mit der Beobachtungen der zweiten Verteilung folgen, bis die erste Verteilung vollständig ersetzt wurde.

Mit dem Fokus auf Online Diskriminanzanalyse für Datenströme mit concept drift werden in Kapitel 3 zunächst die verschiedenen Diskriminanzanalyseverfahren vorgestellt. Die Lineare Diskriminanzanalyse nach Fisher (*Fisher LDA*, Abschnitt 3.3) leitet sich aus einer geometrischen Idee zur Trennung von Daten mit kategorialer Zielvariable ab. Die Kanonische Diskriminanzanalyse resultiert hingegen als Folgerung der Bayes-Regel und basiert auf der Annahme multivariat normalverteilter Beobachtungen in den einzelnen Klassen. Werden identische Kovarianzmatrizen $\boldsymbol{\Sigma}$ für alle Klassen angenommen, so entspricht dies der Kanonischen Linearen Diskriminanzanalyse (*Kanonische LDA*, Abschnitt 3.2.1), bei welcher nur lineare Trennungen möglich sind. Durch Lockerung der Annahme und Unterstellung von unterschiedlichen Kovarianzmatrizen in den Klassen sind auch nicht-lineare Klassifikationsgrenzen durch die Kanonische Quadratische Diskriminanzanalyse (*Kanonische QDA*, Abschnitt 3.2.2) möglich. Anlehnend an die Bachelorarbeit von van Meegen (2015) sowie die Veröffentlichung von van Meegen et al. (2019) wird gezeigt, dass die Fisher LDA sowie die Kanonische LDA trotz unterschiedlicher Annahmen und Herleitungen unabhängig von der Anzahl an Klassen M, Anzahl an Variablen p und der Form der a-priori Wahrscheinlichkeiten in der Theorie dieselben Ergebnisse liefern (Abschnitt 3.4).

Als Online Varianten der Diskriminanzanalyse werden in Kapitel 4 drei Methoden vorgestellt. Pang et al. (2005a,b) schlagen eine Adaption der Fisher LDA für Datenströme vor. Die *Sequential Incremental LDA* (*Sequential ILDA*) bzw. *Chunk Incremental LDA* (*Chunk ILDA*) ermöglicht die Aktualisierung der Klassifikationsregel der Fisher LDA durch Standard Update-Formeln für Mittelwertvektoren und Kovarianzmatrizen anhand von einer einzelnen bzw. mehreren neuen Beobachtungen im Datenstrom. Es erfolgt dabei jedoch keine Adaption an einen möglichen concept drift. Alle Beobachtungen fließen mit identischem (bzw. ohne) Gewicht ein, sodass die aktualisierte Klassifikationsregel zu jedem Zeitpunkt jener entspricht, welche durch Anpassung auf allen bis zu diesem Zeitpunkt realisierten Beobachtungen gleichzeitig resultiert (Batch Variante). Eine Möglichkeit zur Anpassung an concept drift schlagen Kuncheva und Plumpton (2008) durch die Methode *Online Linear Discriminant Classifier* (*OLDC*) als Online Variante der Kanonischen LDA

vor. Bei dieser können neue Beobachtungen mit einem Gewicht in Form einer (festen oder adaptiven) Lernrate λ in die Aktualisierungsformeln der nötigen Größen zur Bildung der Klassifikationsregel einfließen. Anagnostopoulos et al. (2012) ziehen hingegen die Idee des adaptiven exponentiellen Vergessens heran, wodurch eine Adaption an eine Veränderung der zugrunde liegenden Verteilung der Beobachtungen im Datenstrom erzielt werden soll. Für alle drei Methoden wurden verschiedenste Korrekturen vorgenommen (in Kapitel 4 gekennzeichnet). Die Methoden werden zudem an einigen Stellen entgegen der ursprünglichen Beschreibung ausführlicher erläutert und für die jeweiligen Update-Formeln werden teilweise neue Fallunterscheidungen eingeführt.

Die beiden zuletzt genannten Methoden werden in Kapitel 5 wie die Methode *Chunk ILDA* auf Chunk Varianten erweitert, sodass auch mehr Beobachtungen als eine einzelne in die einzelnen Aktualisierungsschritte einfließen können.

Die Anpassungsgüte der so resultierenden Klassifikatoren der Online Varianten der Diskriminanzanalyse mit möglicher Adaption an einen concept drift kann entgegen jener der standardmäßigen Diskriminanzanalyse stark verbessert werden. Allerdings wird auch bei diesen Online Varianten im Datenstrom zu jedem Zeitpunkt nur Information der bis dato aufgetretenen Beobachtungen einbezogen. Ein kontinuierlicher Trend des concept drifts, der auch in der Zukunft anhält, wird nicht beachtet. Dies hat zur Folge, dass in die Klassifikationsregel, mithilfe derer zukünftige Beobachtungen prognostiziert werden sollen, „veraltete“ Schätzer für die Erwartungswertvektoren einfließen. Wird ein (linearer) Trend der Erwartungswertvektoren der Klassen unterstellt, so verlieren die bisherigen Schätzer für die Erwartungswertvektoren aller Methoden an Güte hinsichtlich der Prognose zukünftiger Beobachtungen durch die Klassifikationsregel der Diskriminanzanalyse.

In Kapitel 6 wird gezeigt, dass die Schätzfunktionen $T_1^{(c)}$ für den Erwartungswertvektor aus Klasse c aller betrachteten Methoden für Online Diskriminanzanalyse zum Zeitpunkt t im Datenstrom bei Annahme einer stabilen Verteilung über die Zeit erwartungstreu für den folgenden Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ sind, falls nur Beobachtungen in Klasse c realisiert werden. Für *Sequential ILDA* wird dies für den allgemeinen Fall bewiesen, dass die Beobachtungen zu jedem Zeitpunkt zufällig in den einzelnen Klassen realisiert werden. Auch bei den anderen Methoden ist von Erwartungstreue der Schätzfunktionen im allgemeinen Fall in Situationen ohne concept drift auszugehen. Bei Annahme eines linearen Trends der Erwartungswertvektoren der Klassen sind die Schätzfunktionen $T_1^{(c)}$ der Methoden für Online Diskriminanzanalyse zum Zeitpunkt t jedoch nicht mehr erwartungstreu für die Erwartungswertvektoren $\boldsymbol{\mu}_{t+1}^{(c)}$ der Prognose (Zusammenfassung in Abschnitt 8.5).

Die Verbesserung der Methoden für Online Diskriminanzanalyse in Kapitel 7 basiert auf der Idee, dass der concept drift geeignet modelliert und prognostiziert wird. Da ein linearer Trend $\boldsymbol{\mu}_i^{(c)} = \boldsymbol{\beta}_0^{(c)} + \boldsymbol{\beta}_1^{(c)} i$ des Erwartungswertvektors jeder Klasse c über die Zeit ($i = 1, \ldots$) unterstellt wird (Abschnitt 7.2), werden im Laufe des Datenstroms für jede

Klasse jeweils lokale lineare Regressionsmodelle an die letzten $n_{\text{trend}}^{(c)}$ durch die Online Diskriminanzanalyse aktualisierten Mittelwertvektoren als Schätzer der Erwartungswertvektoren angepasst. Die Mittelwertvektoren zu jedem festen Zeitpunkt t werden mit jeder neuen (je nach Methode teilweise gewichteten) Beobachtung im Datenstrom aktualisiert. Würde der Mittelwertvektor hingegen analog basierend auf allen Beobachtungen gleichzeitig berechnet werden, so wäre recht intuitiv, dass dieser Mittelwertvektor im Falle eines linearen Trends der Erwartungswertvektoren ein repräsentativer Schätzer für einen zeitlich verzögerten Erwartungswertvektor ist. Im Falle identisch gewichteter Beobachtungen ist dieser zum Beispiel ein repräsentativer Schätzer für den Erwartungswertvektor des Zeitpunktes $\frac{t+1}{2}$. Als Einflussvariablen in den lokalen linearen Regressionsmodellen zu jedem Zeitpunkt t werden daher verschobene Zeitpunkte $z_i^{(c)}$ betrachtet, da die Mittelwertvektoren aufgrund der kontinuierlichen Aktualisierungen die zeitverzögerten Erwartungswertvektoren $\boldsymbol{\mu}_{z_i^{(c)}}^{(c)}$ repräsentieren (Abschnitt 7.3). Die verschobenen Zeitpunkte werden dabei im Datenstrom analog zu den entsprechenden Mittelwertvektoren (z. B. mit identischen Gewichtungen) aktualisiert. Nach Bestimmung der Kleinste-Quadrate-Schätzer $\hat{\boldsymbol{\beta}}_{0t}^{(c)}$ und $\hat{\boldsymbol{\beta}}_{1t}^{(c)}$ im linearen Modell zum Zeitpunkt t kann der Erwartungswertvektor für Klasse c des kommenden Zeitpunktes prognostiziert werden durch $\hat{\boldsymbol{\mu}}_{t+1}^{(c)} = \hat{\boldsymbol{\beta}}_{0t}^{(c)} + \hat{\boldsymbol{\beta}}_{1t}^{(c)}(t+1)$ (Abschnitt 7.4). Diese Schätzer für alle Klassen c ersetzen die aktualisierten Mittelwertvektoren in der jeweiligen Klassifikationsregel der Online Diskriminanzanalyse zum Zeitpunkt t, um eine bessere Prognose für Beobachtungen des folgenden Zeitpunktes $t+1$ gewährleisten zu können. Die Lokalität durch Beschränkung auf die letzten $n_{\text{trend}}^{(c)}$ aktualisierten Mittelwertvektoren durch die Online Diskriminanzanalyse im Datenstrom für die Anpassung lokaler linearer Regressionsmodelle hat für die praktische Anwendung den Vorteil, dass auch nicht-lineare Trends geeignet linear approximiert werden können.

In Kapitel 8 wird gezeigt, dass die Schätzfunktionen $T_2^{(c)}$ für die Erwartungswertvektoren der erweiterten Methoden in dem Spezialfall, dass bis zum Zeitpunkt t nur Beobachtungen in Klasse c realisiert werden, nun jeweils erwartungstreu für den Erwartungswertvektor $\boldsymbol{\mu}_{t+1}^{(c)}$ der Verteilung der Prognose sind. Ebenso bleibt die Erwartungstreue auch im Falle stabiler Verteilungen $\boldsymbol{\mu}^{(c)} := \boldsymbol{\mu}_i^{(c)}$ erhalten. Für die Schätzfunktion der Erweiterung von *Sequential ILDA* wird die Erwartungstreue für beide Situationen (linearer Trend und stabile Verteilung) im allgemeinen Fall der zufälligen Realisation der Beobachtungen in einer der Klassen gezeigt (Zusammenfassung in Abschnitt 8.5).

Die Durchführung einer umfangreichen Simulationsstudie soll diese theoretischen Ergebnisse untermauern und auf den allgemeinen Fall der zufälligen Realisation der Beobachtungen in einer der M Klassen zu jedem Zeitpunkt im Datenstrom erweitern. Neben zwei in der gängigen Literatur zu concept drift Methodik etablierten Datensituationen (*STAGGER* und *moving hyperplane*, Abschnitt 9.1) werden weitere Datensituationen mit verschiedenen Arten und Stärken von concept drift als Ausprägungen des unendlichen Raumes aller möglichen Datensituationen mit vorliegendem concept drift erzeugt (Abschnitt 9.2). Bei diesen ist es möglich den Bayesfehler zu bestimmen und demnach die Größenordnung re-

sultierender Prognosefehler der Methoden zu interpretieren. Es werden Situationen mit verschiedener Ausprägung eines *incremental drifts* und linearem Trend der Erwartungswertvektoren erzeugt, da die Erweiterung der Methoden für Online Diskriminanzanalyse für diese Art von concept drift entwickelt wurde. Weiterhin werden auch Datensituationen mit *incremental drift* jedoch nicht-linearem Trend sowie Datensituationen mit *sudden* und *gradual drifts* und eine Datensituation mit stabiler Verteilung über die Zeit betrachtet, um die Auswirkung der Erweiterung auf die Prognosegüte und insbesondere die Approximation durch einen linearen Trend analysieren zu können. Jede Datensituation wird mit $p \in \{2, 3, 10\}$ Einflussvariablen betrachtet.

Anhand der Ergebnisse (Abschnitt 9.6) der in Abschnitt 9.3–9.5 beschriebenen Simulationsstudie wird deutlich, dass die Prognosegüte der Klassifikatoren durch Erweiterung der Methoden im Falle eines linearen Trends der Erwartungswertvektoren deutlich verbessert werden kann. Als Maß zur Analyse der Prognosegüte wird für alle Update-Methoden und Erweiterungen auf allen Datensituationen jeweils der Verlauf des mittleren Prognosefehlers betrachtet. Als einzelne Maßzahl werden zudem jeweils der durchschnittliche mittlere Prognosefehler über die Zeit sowie der durchschnittliche euklidische Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren herangezogen (vgl. Seite 232 ff.). Der jeweilige Bayesfehler der Datensituationen (mit jeder Dimension p) kann zu vielen Zeitpunkten durch den Prognosefehler der erweiterten Methoden approximiert werden. Bei allen erweiterten Methoden sinkt sowohl der durchschnittliche mittlere Prognosefehler als auch der durchschnittliche euklidische Abstand zwischen wahren und mittleren prognostizierten Erwartungswertvektoren über die Zeit mit wachsender Fenstergröße N_{trend} (vgl. Seite 162) für die lokalen linearen Regressionsmodelle entgegen den Werten der ursprünglichen Methoden. Dies untermauert die Ergebnisse zur Erwartungstreue der Schätzfunktionen (Kapitel 8). Es ist zu vermuten, dass die Erwartungstreue der Schätzfunktionen der erweiterten Methoden auch im allgemeinen Fall und unabhängig von der Dimension p gilt, sofern die Annahme eines linearen Trends der Erwartungswertvektoren in den Klassen gerechtfertigt ist. Bei höherer Dimension p werden breitere Fenster N_{trend} für die lokalen linearen Regressionsmodelle für eine vergleichbar gute Prognose der Erwartungswertvektoren und folglich vergleichbar geringe Prognosefehler durch die Diskriminanzanalyse benötigt (Abschnitt 9.6.3). Die Schätzung und Prognose der Erwartungswertvektoren der Klassen wird stabiler und die Varianz reduziert.

Bei den Datensituationen ohne linearen Trend der Erwartungswertvektoren (auch *sudden* und *gradual drift*) wird deutlich, dass die lineare Approximation durch die lokalen linearen Regressionsmodelle häufig greift. Die neuen Schätzfunktionen für die Erwartungswertvektoren der Prognose sind in diesem Fall zwar nicht erwartungstreu, der Prognosefehler und euklidische Abstand zwischen prognostizierten und wahren Erwartungswertvektoren kann durch die Erweiterung bei geeigneter Wahl der Fensterbreite N_{trend} und demnach Anpassung lokaler linearer Regressionsmodelle auf $n_{\text{trend}}^{(c)}$ aktualisierten Mittelwertvektoren (vgl. Seite 162) im Vergleich zu den Ergebnissen der ursprünglichen Methoden trotzdem teilweise

deutlich verringert werden. Eine Approximation des Bayesfehlers durch den Prognosefehler wird im Gegensatz zu den ursprünglichen Methoden für Online Diskriminanzanalyse zu vielen Zeitpunkten ermöglicht, sofern eine lineare Trennung der Daten möglich ist. Der Prognosefehler wird also minimiert.

Während die Prognosefehler der verschiedenen ursprünglichen Methoden für Online Diskriminanzanalyse stark variieren können, erfolgt eine Annäherung der Fehler nach Erweiterung. Bei geringer Dimension p sinkt zudem häufig der Einfluss der einstellbaren Parameter (Lernrate λ und Parameter für adaptive Lernrate bei *OLDC*) nach Integration lokaler linearer Regressionsmodelle. Wird ein concept drift vermutet, so kann die Lernrate bei *OLDC* auf einen hohen Wert von $\lambda = 0.9$ festgesetzt werden. Da in Klassifikationsproblemen mit mehr Einflussvariablen p die Wahl der freien Parameter der Methoden in Hinblick auf die Prognosegüte jedoch nicht irrelevant ist (Abschnitt 9.6.3), ist zu empfehlen diese nach Möglichkeit mit Vorkenntnissen zu wählen oder im Datenstrom zunächst zu optimieren.

Die optimale Wahl der (ursprünglichen) Methode für Online Diskriminanzanalyse und aller Parameter hängt in praktischen Anwendungen von der speziellen Datensituation, dem Analyseziel, der individuellen Zielgruppe sowie den (Rechen-)Kapazitäten ab. Die Prognosegüte jeder einzelnen dieser Methoden kann durch die eingeführte Erweiterung durch Integration lokaler linearer Regressionsmodelle jedoch häufig verbessert werden. Zudem ist herauszustellen, dass sich die Unterschiede der durchschnittlichen Prognosefehler über die Zeit für die unterschiedlichen Update-Methoden für Diskriminanzanalyse nach der Erweiterung verringern, sodass die Wahl der Update-Methode nach Präferenzen erfolgen kann.

In praktischen Anwendungen sollte insbesondere die Fensterbreite N_{trend} für die Regressionsmodelle geeignet gewählt oder im Datenstrom optimiert werden. Bei einem strikt linearen Trend der Erwartungswertvektoren ist ein möglichst breites Fenster ideal und die prognostizierten Erwartungswertvektoren konvergieren aufgrund der Erwartungstreue gegen ihren wahren Wert sowie die Prognosefehler gegen den Bayesfehler. In der Praxis ist ein strikt kontinuierlicher linearer Trend jedoch unwahrscheinlich und der concept drift wird lediglich lokal linear approximiert. Ein zu geringes N_{trend} führt dann zu unsicheren Schätzungen aufgrund weniger Beobachtungen für die einzelnen linearen Regressionsmodelle, bei zu breitem Fenster N_{trend} greift hingegen die lineare Approximation eines nicht-linearen Trends nicht mehr. Bei geeigneter Wahl kann die Erweiterung der Methoden für Online Diskriminanzanalyse die Prognosegüte jedoch in vielen Datensituationen mit concept drift verbessern, selbst wenn kein linearer Trend der Erwartungswertvektoren vorliegt.

In dieser Arbeit liegt der Fokus auf Erwartungstreue der Schätzfunktionen. In Kapitel 7 wird deutlich, dass die herangezogenen KQ-Schätzer der betrachteten linearen Regressionsmodelle bei Unterstellung eines linearen Trends der Erwartungswertvektoren zwar erwartungstreu, allerdings aufgrund der autokorrelierten Fehler nicht BLUE sind, d. h. nicht die geringste Varianz unter allen linearen erwartungstreuen Schätzern aufweisen. Die Autokorrelation resultiert dadurch, dass die herangezogenen Mittelwertvektoren im linearen

Modell aufgrund der kontinuierlichen Aktualisierung in der Online Diskriminanzanalyse teilweise auf denselben Beobachtungen basieren und damit selbst korreliert sind. In einem weiteren Schritt könnte daher analysiert werden, wie diese Korrelationsstruktur aussieht. Diese könnte herangezogen werden, um verallgemeinerte KQ-Schätzer (Abschnitt 7.4) heranzuziehen und somit die Varianz der Schätzer noch weiter zu verringern.

Des Weiteren ist eine weitere Verbesserung in Hinblick auf die Kovarianzmatrizen denkbar. In Abschnitt 7.6 wird erläutert, dass die Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}$ in der Fisher LDA im Zuge der Modellierung des Trends der Erwartungswertvektoren für die verbesserte Klassifikationsregel ebenfalls auf Basis der prognostizierten Erwartungswertvektoren neu angepasst wird. Für die Kovarianzmatrix $\boldsymbol{B}$ ist dies möglich, da diese streng genommen in der Online LDA nicht aktualisiert, sondern vielmehr neu berechnet wird, da diese nur auf den Mittelwertvektoren, nicht jedoch auf den einzelnen Beobachtungen basiert. Dies ist ein wesentlicher Unterschied zu der gepoolten Kovarianzmatrix $\boldsymbol{S}$, die sowohl bei der Fisher LDA als auch der Kanonischen LDA betrachtet wird. Da die einzelnen Beobachtungen im Datenstrom nicht gespeichert werden, wird die aktuelle Kovarianzmatrix $\boldsymbol{S}$ in den einzelnen Update-Methoden für Diskriminanzanalyse mithilfe von Aktualisierungsformeln neu bestimmt. Die vorgestellte Erweiterung der Methoden für Online Diskriminanzanalyse greift bisher lediglich in die Konstruktion der Klassifikationsregel ein, allerdings nicht in die einzelnen Aktualisierungsschritte der nötigen Größen. Vor diesem Hintergrund ist es nicht möglich für die aktuelle Kovarianzmatrix $\boldsymbol{S}$ die prognostizierten Mittelwertvektoren heranzuziehen. Da allerdings auch die gepoolte Kovarianzmatrix in Datenströmen mit concept drift degenerieren kann und sich insbesondere bei linearem Trend der Erwartungswertvektoren im Laufe der Zeit „streckt“, könnte sich eine Annahme dieses Problems weiter positiv auf die Prognosegüte der Klassifikatoren auswirken.

Im thematischen Kontext mit concept drift ist zudem eine Lockerung der Annahme eines strikt linearen Trends der Erwartungswertvektoren interessant. Möglich wäre die Unterstellung nicht-linearer Trends und eine Analyse der Problemstellung mithilfe von weiteren statistischen Modellen anstelle der linearen Regression. In Hinblick auf eine Veränderung der Erwartungswertvektoren über die Zeit kann darüber hinaus die Problemstellung einer zusätzlichen Veränderung der Kovarianzmatrix in Betracht gezogen werden. Besonders in praktischen Datensituationen geht eine Veränderung des Erwartungswertes sehr häufig mit einer Veränderung der Varianz einher. Die Annahme einer konstanten Kovarianzmatrix über die Zeit könnte aufgehoben werden. Stattdessen könnten zum Beispiel zeitinvariante Korrelationsstrukturen unterstellt werden und mithilfe des Trends der Erwartungswertvektoren könnte die Veränderung der Kovarianzmatrizen modelliert werden.

Zudem kann die vorgestellte Idee zur Verbesserung der Prognosegüte unabhängig von den speziell betrachteten Update-Methoden bzw. der Diskriminanzanalyse betrachtet werden, sondern als grundlegende Idee angesehen werden. Eine Übertragung und Adaption auf beliebige weitere Klassifikationsmodelle und entsprechende Online Varianten ist denkbar und interessant für weitergehende Forschungsarbeit.

Anhang

A Beispiele für die Methoden aus Kapitel 4

Zur Veranschaulichung werden alle Methoden im Folgenden auf einen einfachen eindimensionalen Beispiel-Datenstrom angewandt:

Tabelle A.1: 13 Beobachtungen eines Beispiel-Datenstroms.

t	1	2	3	4	5	6	7	8	9	10	11	12	13
x_t	1	1.5	11	2.5	9	3.5	4	6	5	5.5	4	6.5	1
c_t $(g(x_t))$	1	1	2	1	2	1	1	2	2	1	3	1	2

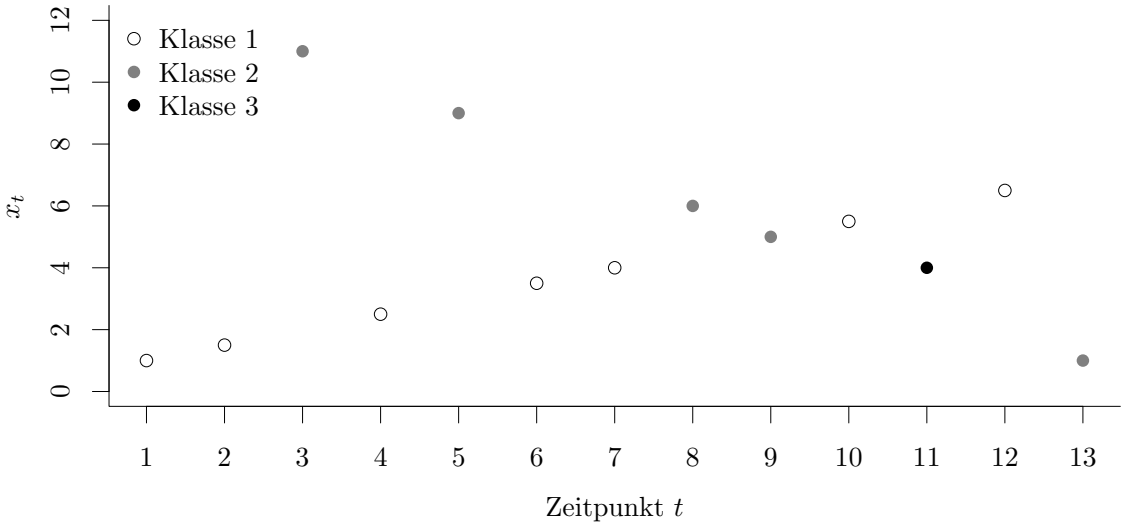


Abbildung A.1: Beispiel-Datenstrom.

In Abbildung A.1 sind die Daten veranschaulicht. Der Datensatz besteht aus 13 Beobachtungen in drei Klassen, wobei die dritte Klasse erst mit der elften Beobachtung auftritt. Die Beobachtungen einer jeden Klasse unterliegen einem linearen Trend.

A.1 Beispiel für Sequential Incremental LDA (Abschnitt 4.2)

Zunächst wird die Methode *Sequential Incremental LDA* aus Abschnitt 4.2 betrachtet.

Initialisierung Die Initialisierung erfolgt auf den ersten $n_{10} = 10$ Beobachtungen.

Die Anzahl der Beobachtungen in jeder Klasse beträgt:

$$n_{10}^{(1)} = 6,\ n_{10}^{(2)} = 4 \quad \text{nach (4.1)}.$$

Der Mittelwert berechnet sich folgendermaßen:

$$\boldsymbol{m}_{n_{10}} = \frac{1}{n_{10}} \sum_{i=1}^{10} \boldsymbol{x}_i = \frac{1}{10} \cdot (1 + 1.5 + \ldots + 5 + 5.5) = 4.9 \quad \text{nach (4.2)}.$$

Die Mittelwerte der Klassen $c \in \{1, 2\}$ können mithilfe von (4.3) bestimmt werden:

$$\boldsymbol{m}_{n_{10}^{(1)}}^{(1)} = \frac{1}{n_{10}^{(1)}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=1 \\ i \leq 10}} \boldsymbol{x}_i = \frac{1}{6} \cdot (1 + 1.5 + 2.5 + 3.5 + 4 + 5.5) = \frac{18}{6} = 3,$$

$$\boldsymbol{m}_{n_{10}^{(2)}}^{(2)} = \frac{1}{n_{10}^{(2)}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=2 \\ i \leq 10}} \boldsymbol{x}_i = \frac{1}{4} \cdot (11 + 9 + 6 + 5) = \frac{31}{4} = 7.75.$$

Die relativen Häufigkeiten als initiale Schätzer für die a-priori Wahrscheinlichkeiten der einzelnen Klassen ergeben sich durch

$$P_{10}^{(1)} = \frac{n_{10}^{(1)}}{n_{10}} = \frac{6}{10} = 0.6, \quad P_{10}^{(2)} = \frac{n_{10}^{(2)}}{n_{10}} = \frac{4}{10} = 0.4.$$

Die Zwischen-den-Klassen Kovarianzmatrix und gepoolte Kovarianzmatrix innerhalb der Klassen (jeweils ohne Vorfaktor und hier Varianzen im eindimensionalen Raum) ergeben sich durch (4.6) und (4.7) und sehen für die ersten zehn Beobachtungen folgendermaßen aus:

$$\tilde{\boldsymbol{B}}_{10} = \sum_{c=1}^{2} n_{10}^{(c)} \left(\boldsymbol{m}_{n_{10}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{10}} \right)^2 = 6 \cdot (3 - 4.9)^2 + 4 \cdot (7.75 - 4.9)^2 = 54.15,$$

$$\tilde{\boldsymbol{S}}_{10} = \sum_{c=1}^{2} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq 10}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{10}^{(c)}}^{(c)} \right)^2 \overset{(4.4)}{=} \sum_{c=1}^{2} \left((n_{10}^{(c)} - 1) \boldsymbol{S}_{10}^{(c)} \right) = 5 \cdot \frac{14}{5} + 3 \cdot \frac{22.75}{3} = 36.75.$$

1. Prognose Mit diesen Größen kann die aktuelle Klassifikationsregel (3.47) zum Zeitpunkt $t = 10$ aufgestellt werden. Dazu wird zunächst der Schätzer für die Inverse der Quadratwurzel der Kovarianzmatrix

$$\boldsymbol{S}_{10} = \frac{1}{n_{10} - M} \cdot \tilde{\boldsymbol{S}}_{10}$$

erstellt, wobei nun der Vorfaktor mit betrachtet wird (vgl. Seite 61):

$$\boldsymbol{S}_{10}^{-1/2} = \sqrt{(n_{10} - M)\tilde{\boldsymbol{S}}_{10}^{-1}} = \sqrt{(10-2) \cdot \frac{1}{36.75}} \approx 0.4666.$$

Aufgrund der Eindimensionalität ist die Berechnung hier vereinfacht. Bei mehr Dimensionen würde mithilfe einer Spektralzerlegung (3.39) der Schätzer $\boldsymbol{S}_{10}^{-1/2}$ für die Inverse der Quadratwurzel der Kovarianzmatrix $\boldsymbol{\Sigma}^{-1/2}$ bestimmt werden.

Eine Eigenwertzerlegung von (vgl. Seite 48 f.)

$$\boldsymbol{S}_{10}^{-1/2}\boldsymbol{B}_{10}\boldsymbol{S}_{10}^{-1/2} \approx 0.4666 \cdot \frac{1}{10} \cdot 54.15 \cdot 0.4666 \approx 1.1788$$

liefert den Eigenvektor $\boldsymbol{\nu}_1 = 1$ zum einzigen (positiven) Eigenwert 1.1788, wobei auch der Schätzer für die Zwischen-den-Klassen Kovarianzmatrix $\boldsymbol{B}_{10} = \frac{1}{n_{10}}\tilde{\boldsymbol{B}}_{10}$ hier mit Vorfaktor betrachtet wird (vgl. Seite 61). Aufgrund der Eindimensionalität ist jedoch auch hier streng genommen keine Eigenwertzerlegung nötig.

Die 1. Diskriminanzkomponente ergibt sich durch Transformation mithilfe von (3.41) unter Verwendung der Schätzer durch

$$\boldsymbol{\alpha}_1 \approx \boldsymbol{S}_{10}^{-1/2}\boldsymbol{\nu}_1 = 0.4666.$$

Mithilfe der Diskriminanzkomponente liefert die folgende Minimierung (vgl. (3.47)) die prognostizierte Klasse für die neue Beobachtung $\boldsymbol{x}_{11} = 4$:

$$\begin{aligned}
\hat{c}_{11} &= \underset{c=1,2}{\arg\min} \left(\left(\boldsymbol{\alpha}_1 \left(\boldsymbol{x}_{11} - \boldsymbol{m}^{(1)}_{n^{(1)}_{10}} \right) \right)^2 - 2\log P^{(1)}_{10}, \left(\boldsymbol{\alpha}_1 \left(\boldsymbol{x}_{11} - \boldsymbol{m}^{(2)}_{n^{(2)}_{10}} \right) \right)^2 - 2\log P^{(2)}_{10} \right) \\
&\approx \underset{c=1,2}{\arg\min} \left((0.4666 \cdot (4-3))^2 - 2\log(0.6), (0.4666 \cdot (4-7.75))^2 - 2\log(0.4) \right) \\
&\approx \underset{c=1,2}{\arg\min} \left(1.2394, 4.8942\right) = 1.
\end{aligned}$$

Natürlich kann an dieser Stelle die wahre Klasse $c_{11} = 3$ nicht prognostiziert werden, da diese erst zum Zeitpunkt 11 zum ersten Mal neu auftritt.

1. Aktualisierung Zunächst erfolgt eine Aktualisierung aller Größen durch die Beobachtung $\boldsymbol{x}_{11} = 4$ mit neuer Klasse $c_{11} = 3$, also $g(\boldsymbol{x}_{11}) = 3$.

Damit ergeben sich die aktualisierten Mittelwerte der Klassen durch (4.8):

$$\boldsymbol{m}^{(1)}_{n^{(1)}_{11}} = \boldsymbol{m}^{(1)}_{n^{(1)}_{10}} = 3, \quad \boldsymbol{m}^{(2)}_{n^{(2)}_{11}} = \boldsymbol{m}^{(2)}_{n^{(2)}_{10}} = 7.75, \quad \boldsymbol{m}^{(3)}_{n^{(3)}_{11}} = \boldsymbol{x}_{11} = 4.$$

Der gesamte Mittelwert verändert sich zu

$$\boldsymbol{m}_{n_{11}} = \frac{n_{10}\boldsymbol{m}_{n_{10}} + \boldsymbol{x}_{11}}{n_{10}+1} = \frac{1}{10+1} \cdot (10 \cdot 4.9 + 4) = \frac{53}{11} = 4.\overline{81} \approx 4.8181 \quad \text{nach (4.9)}$$

und die neuen Anzahlen an Beobachtungen in den einzelnen Klassen betragen nach (4.10):

$$n_{11}^{(1)} = n_{10}^{(1)} = 6, \quad n_{11}^{(2)} = n_{10}^{(2)} = 4, \quad n_{11}^{(3)} = 1.$$

Die aktuellen Schätzer für die a-priori Wahrscheinlichkeiten lauten nach (4.11):

$$P_{11}^{(1)} = \frac{n_{10}^{(1)}}{n_{10}+1} = \frac{6}{11} = 0.\overline{54}, \quad P_{11}^{(2)} = \frac{n_{10}^{(2)}}{n_{10}+1} = \frac{4}{11} = 0.\overline{36},$$
$$P_{11}^{(3)} = \frac{1}{n_{10}+1} = \frac{1}{11} = 0.\overline{09}.$$

$\tilde{\boldsymbol{B}}$ und $\tilde{\boldsymbol{S}}$ werden mittels (4.12) und (4.13) aktualisiert:

$$\begin{aligned}\tilde{\boldsymbol{B}}_{11} &= \sum_{c=1}^{3} n_{11}^{(c)} \left(\boldsymbol{m}_{n_{11}^{(c)}}^{(c)} - \boldsymbol{m}_{n_{11}}\right)^2 = 6 \cdot \left(3 - \frac{53}{11}\right)^2 + 4 \cdot \left(7.75 - \frac{53}{11}\right)^2 + \left(4 - \frac{53}{11}\right)^2 \\ &= 6 \cdot \left(\frac{-20}{11}\right)^2 + 4 \cdot \left(\frac{32.25}{11}\right)^2 + \left(\frac{-9}{11}\right)^2 = \frac{6641.25}{121} \approx 54.8864, \\ \tilde{\boldsymbol{S}}_{11} &= \tilde{\boldsymbol{S}}_{10} = 36.75.\end{aligned}$$

2. Prognose Der Schätzer für die Inverse der Quadratwurzel der Kovarianzmatrix ist

$$\boldsymbol{S}_{11}^{-1/2} = \sqrt{(n_{11} - M)\tilde{\boldsymbol{S}}_{11}^{-1}} = \sqrt{(11-3) \cdot \frac{1}{36.75}} \approx 0.4666.$$

Da es sich um ein eindimensionales Problem handelt und nur Varianzen betrachtet werden, muss keine Eigenwertzerlegung des Produktes $\boldsymbol{S}_{11}^{-1/2}\boldsymbol{B}_{11}\boldsymbol{S}_{11}^{-1/2} = \frac{1}{n_{11}}\boldsymbol{S}_{11}^{-1/2}\tilde{\boldsymbol{B}}_{11}\boldsymbol{S}_{11}^{-1/2} \approx 1.0863$ durchgeführt werden, um die Eigenvektoren zu ermitteln. Der einzige Eigenvektor ist in jedem Fall $\boldsymbol{\nu}_1 = 1$.

Die 1. Diskriminanzkomponente beträgt dann

$$\boldsymbol{\alpha}_1 \approx \boldsymbol{S}_{11}^{-1/2}\boldsymbol{\nu}_1 = 0.4666.$$

Unter Betrachtung der Schätzer liefert (3.47) die prognostizierte Klasse für $\boldsymbol{x}_{12} = 6.5$:

$$\begin{aligned}\hat{c}_{12} = \underset{c=1,2,3}{\arg\min} \Bigg(&\left(\boldsymbol{\alpha}_1 \left(\boldsymbol{x}_{12} - \boldsymbol{m}_{n_{11}^{(1)}}^{(1)}\right)\right)^2 - 2\log P_{11}^{(1)}, \left(\boldsymbol{\alpha}_1 \left(\boldsymbol{x}_{12} - \boldsymbol{m}_{n_{11}^{(2)}}^{(2)}\right)\right)^2 - 2\log P_{11}^{(2)}, \\ &\left(\boldsymbol{\alpha}_1 \left(\boldsymbol{x}_{12} - \boldsymbol{m}_{n_{11}^{(3)}}^{(3)}\right)\right)^2 - 2\log P_{11}^{(3)}\Bigg) \\ \approx \underset{c=1,2,3}{\arg\min} \Bigg(&(0.4666 \cdot (6.5-3))^2 - 2\log\left(\frac{6}{11}\right), (0.4666 \cdot (6.5-7.75))^2 - 2\log\left(\frac{4}{11}\right), \\ &(0.4666 \cdot (6.5-4))^2 - 2\log\left(\frac{1}{11}\right)\Bigg) \\ \approx \underset{c=1,2,3}{\arg\min}\, &(3.8793, 2.3634, 6.1565) = 2.\end{aligned}$$

2. Aktualisierung Die nächste Beobachtung im Datenstrom $\boldsymbol{x}_{12} = 6.5$ weist die Klasse $c_{12} = 1$ auf, es gilt also $g(\boldsymbol{x}_{12}) = 1$. Mithilfe derselben Update-Formeln (4.8)–(4.13) ergeben sich die aktualisierten Größen folgendermaßen:

$$
\begin{aligned}
\boldsymbol{m}^{(1)}_{n^{(1)}_{12}} &= \frac{n^{(1)}_{11}\boldsymbol{m}^{(1)}_{n^{(1)}_{11}} + \boldsymbol{x}_{12}}{n^{(1)}_{11}+1} = \frac{1}{6+1}\cdot(6\cdot 3 + 6.5) = \frac{24.5}{7} = 3.5,\\
\boldsymbol{m}^{(2)}_{n^{(2)}_{12}} &= \boldsymbol{m}^{(2)}_{n^{(2)}_{11}} = 7.75, \quad \boldsymbol{m}^{(3)}_{n^{(3)}_{12}} = \boldsymbol{m}^{(3)}_{n^{(3)}_{11}} = 4,\\
\boldsymbol{m}_{n_{12}} &= \frac{n_{11}\boldsymbol{m}_{n_{11}} + \boldsymbol{x}_{12}}{n_{11}+1} = \frac{1}{11+1}\cdot\left(11\cdot\frac{53}{11} + 6.5\right) = \frac{59.5}{12} = \frac{119}{24} = 4.958\overline{3} \approx 4.9583,\\
n^{(1)}_{12} &= n^{(1)}_{11} + 1 = 6 + 1 = 7, \quad n^{(2)}_{12} = n^{(2)}_{11} = 4, \quad n^{(3)}_{12} = n^{(3)}_{11} = 1,\\
P^{(1)}_{12} &= \frac{n^{(1)}_{11}+1}{n_{11}+1} = \frac{6+1}{11+1} = \frac{7}{12} \approx 0.5833, \quad P^{(2)}_{12} = \frac{n^{(2)}_{11}}{n_{11}+1} = \frac{4}{11+1} = \frac{4}{12} \approx 0.3333,\\
P^{(3)}_{12} &= \frac{n^{(3)}_{11}}{n_{11}+1} = \frac{1}{11+1} = \frac{1}{12} \approx 0.0833,\\
\tilde{\boldsymbol{B}}_{12} &\overset{(4.12)}{=} \sum_{c=1}^{3} n^{(c)}_{12}\left(\boldsymbol{m}^{(c)}_{n^{(c)}_{12}} - \boldsymbol{m}_{n_{12}}\right)^2\\
&= 7\cdot\left(3.5 - \frac{119}{24}\right)^2 + 4\cdot\left(7.75 - \frac{119}{24}\right)^2 + 1\cdot\left(4 - \frac{119}{24}\right)^2\\
&= 7\cdot\left(\frac{-35}{24}\right)^2 + 4\cdot\left(\frac{67}{24}\right)^2 + 1\cdot\left(\frac{-23}{24}\right)^2 = \frac{27060}{576} \approx 46.9792,\\
\tilde{\boldsymbol{S}}_{12} &\overset{(4.13)}{=} \tilde{\boldsymbol{S}}_{11} + \frac{n^{(1)}_{11}}{n^{(1)}_{11}+1}\left(\boldsymbol{x}_{12} - \boldsymbol{m}^{(1)}_{n^{(1)}_{11}}\right)^2 = 36.75 + \frac{6}{6+1}\cdot(6.5-3)^2\\
&= 36.75 + \frac{6}{7}\cdot 12.25 = 36.75 + 10.5 = 47.25.
\end{aligned}
$$

3. Prognose Aufgrund der Eindimensionalität ergibt sich die 1. Diskriminanzkomponente wie in den vorherigen Update-Schritten durch

$$
\boldsymbol{\alpha}_1 = \boldsymbol{S}_{12}^{-1/2}\boldsymbol{\nu}_1 = \sqrt{(n_{12} - M)\tilde{\boldsymbol{S}}_{12}^{-1}}\,\boldsymbol{\nu}_1 = \sqrt{(12-3)\cdot\frac{1}{47.25}}\cdot 1 \approx 0.4364.
$$

Die prognostizierte Klasse für Beobachtung $\boldsymbol{x}_{13} = 1$ ist dann

$$
\begin{aligned}
\hat{c}_{13} &= \operatorname*{arg\,min}_{c=1,2,3}\left(\left(\boldsymbol{\alpha}_1\left(\boldsymbol{x}_{13} - \boldsymbol{m}^{(1)}_{n^{(1)}_{12}}\right)\right)^2 - 2\log P^{(1)}_{12}, \left(\boldsymbol{\alpha}_1\left(\boldsymbol{x}_{13} - \boldsymbol{m}^{(2)}_{n^{(2)}_{12}}\right)\right)^2 - 2\log P^{(2)}_{12},\right.\\
&\qquad\left.\left(\boldsymbol{\alpha}_1\left(\boldsymbol{x}_{13} - \boldsymbol{m}^{(3)}_{n^{(3)}_{12}}\right)\right)^2 - 2\log P^{(3)}_{12}\right)\\
&\approx \operatorname*{arg\,min}_{c=1,2,3}\left((0.4364\cdot(1-3.5))^2 - 2\log\left(\frac{7}{12}\right), (0.4364\cdot(1-7.75))^2 - 2\log\left(\frac{4}{12}\right),\right.\\
&\qquad\left.(0.4364\cdot(1-4))^2 - 2\log\left(\frac{1}{12}\right)\right)\\
&\approx \operatorname*{arg\,min}_{c=1,2,3}(2.2683, 10.8744, 6.6838) = 1.
\end{aligned}
$$

3. Aktualisierung Aktualisierung der Größen mittels der Beobachtung $\boldsymbol{x}_{13} = 1$ aus Klasse 2, also $g(\boldsymbol{x}_{13}) = 2$ führt entsprechend zu folgenden Ergebnissen:

$$\boldsymbol{m}^{(1)}_{n^{(1)}_{13}} = \boldsymbol{m}^{(1)}_{n^{(1)}_{12}} = 3.5, \quad \boldsymbol{m}^{(2)}_{n^{(2)}_{13}} = \frac{n^{(2)}_{12}\boldsymbol{m}^{(2)}_{n^{(2)}_{12}} + \boldsymbol{x}_{13}}{n^{(2)}_{12}+1} = \frac{1}{4+1}\cdot(4\cdot 7.75+1) = \frac{32}{5} = 6.4,$$

$$\boldsymbol{m}^{(3)}_{n^{(3)}_{13}} = \boldsymbol{m}^{(3)}_{n^{(3)}_{12}} = 4,$$

$$\boldsymbol{m}_{n_{13}} = \frac{n_{12}\boldsymbol{m}_{n_{12}} + \boldsymbol{x}_{13}}{n_{12}+1} = \frac{1}{12+1}\cdot\left(12\cdot\frac{119}{24}+1\right) = \frac{60.5}{13} \approx 4.6538,$$

$$n^{(1)}_{13} = n^{(1)}_{12} = 7, \quad n^{(2)}_{13} = n^{(2)}_{12}+1 = 4+1 = 5, \quad n^{(3)}_{13} = n^{(3)}_{12} = 1,$$

$$P^{(1)}_{13} = \frac{n^{(1)}_{12}}{n_{12}+1} = \frac{7}{12+1} = \frac{7}{13} \approx 0.5385, \quad P^{(2)}_{13} = \frac{n^{(2)}_{12}+1}{n_{12}+1} = \frac{4+1}{12+1} = \frac{5}{13} \approx 0.3846,$$

$$P^{(3)}_{13} = \frac{n^{(3)}_{12}}{n_{12}+1} = \frac{1}{12+1} = \frac{1}{13} \approx 0.0769,$$

$$\tilde{\boldsymbol{B}}_{13} \overset{(4.12)}{=} \sum_{c=1}^{3} n^{(c)}_{13}\left(\boldsymbol{m}^{(c)}_{n^{(c)}_{13}} - \boldsymbol{m}_{n_{13}}\right)^2$$

$$\approx 7\cdot(3.5-4.6538)^2 + 5\cdot(6.4-4.6538)^2 + 1\cdot(4-4.6538)^2 \approx 24.9923,$$

$$\tilde{\boldsymbol{S}}_{13} \overset{(4.13)}{=} \tilde{\boldsymbol{S}}_{12} + \frac{n^{(2)}_{12}}{n^{(2)}_{12}+1}\left(\boldsymbol{x}_{13} - \boldsymbol{m}^{(2)}_{n^{(2)}_{12}}\right)^2 = 47.25 + \frac{4}{4+1}\cdot(1-7.75)^2$$

$$= 47.25 + \frac{4}{5}\cdot 45.5625 = 47.25 + 36.45 = 83.7.$$

A.2 OLDC mit fester Lernrate (Abschnitt 4.3)

Für das Rechenbeispiel zur Veranschaulichung der Methode *OLDC (Online Linear Discriminant Classifier)* mit fester Lernrate $\lambda = 1/2$ wird wieder der eindimensionale Datenstrom aus Tabelle A.1 (s. Seite 403) betrachtet.

Initialisierung Die Initialisierung erfolgt erneut auf Basis der ersten zehn Beobachtungen. Die Anzahl der Beobachtungen zum Zeitpunkt $t = 10$ in jeder Klasse beträgt:

$$n_{10}^{(1)} = 6,\ n_{10}^{(2)} = 4 \quad \text{nach (4.1)}.$$

Die Mittelwertvektoren der Klassen $c \in \{1, 2\}$ können auch hier mithilfe von (4.3) bestimmt werden:

$$\boldsymbol{m}_{n_{10}^{(1)}}^{(1)} = \frac{1}{n_{10}^{(1)}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=1 \\ i \le 10}} \boldsymbol{x}_i = \frac{1}{6} \cdot (1 + 1.5 + 2.5 + 3.5 + 4 + 5.5) = \frac{18}{6} = 3,$$

$$\boldsymbol{m}_{n_{10}^{(2)}}^{(2)} = \frac{1}{n_{10}^{(2)}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=2 \\ i \le 10}} \boldsymbol{x}_i = \frac{1}{4} \cdot (11 + 9 + 6 + 5) = \frac{31}{4} = 7.75.$$

Die a-priori Wahrscheinlichkeiten der Klassen werden durch die relativen Häufigkeiten geschätzt:

$$P_{10}^{(1)} = \frac{n_{10}^{(1)}}{n_{10}} = \frac{6}{10} = 0.6, \quad P_{10}^{(2)} = \frac{n_{10}^{(2)}}{n_{10}} = \frac{4}{10} = 0.4.$$

Die inverse gepoolte Kovarianzmatrix innerhalb der Klassen (hier Varianz aufgrund nur einer Dimension) zum Zeitpunkt $t = 10$ ergibt sich durch Invertierung von (4.5) mit Vorfaktor $\frac{1}{n_{10}}$ anstelle von $\frac{1}{n_{10}-M}$ (vgl. Seite 70):

$$\begin{aligned}
\boldsymbol{S}_{10}^{-1} &= \left(\frac{1}{n_{10}} \sum_{c=1}^{2} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le 10}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{10}^{(c)}}^{(c)} \right)^2 \right)^{-1} \\
&= \Big(\frac{1}{10} \cdot \Big((1-3)^2 + (1.5-3)^2 + (2.5-3)^2 + (3.5-3)^2 + (4-3)^2 + (5.5-3)^2 \\
&\qquad + (11-7.75)^2 + (9-7.75)^2 + (6-7.75)^2 + (5-7.75)^2 \Big) \Big)^{-1} \\
&= \left(\frac{1}{10} \cdot 36.75 \right)^{-1} = \frac{1}{3.675} = \frac{40}{147} \approx 0.2721.
\end{aligned}$$

Die Kovarianzmatrix zum Zeitpunkt $t = 10$ kann analog auch auf Basis der empirischen Kovarianzmatrizen der Klassen $c \in \{1, 2\}$ berechnet werden:

$$\boldsymbol{S}_{10} = \frac{1}{n_{10}} \sum_{c=1}^{2} \left((n_{10}^{(c)} - 1) \boldsymbol{S}_{10}^{(c)} \right) = \frac{1}{10} \cdot \left(5 \cdot \frac{14}{5} + 3 \cdot \frac{22.75}{3} \right) = \frac{36.75}{10} = 3.675,$$

wobei die empirischen Varianzen $\boldsymbol{S}_{10}^{(c)}$ der Klassen $c \in \{1,2\}$ berechnet werden durch (4.4):

$$\begin{aligned}
\boldsymbol{S}_{10}^{(1)} &= \frac{1}{n_{10}^{(1)} - 1} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=1 \\ i \le 10}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{10}^{(1)}}^{(1)} \right)^2 \\
&= \frac{1}{5} \cdot \Big((1-3)^2 + (1.5-3)^2 + (2.5-3)^2 + (3.5-3)^2 + (4-3)^2 + (5.5-3)^2 \Big) \\
&= \frac{1}{5} \cdot (4 + 2.25 + 0.25 + 0.25 + 1 + 6.25) = \frac{14}{5} = 2.8, \\
\boldsymbol{S}_{10}^{(2)} &= \frac{1}{n_{10}^{(2)} - 1} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=2 \\ i \le 10}} \left(\boldsymbol{x}_i - \boldsymbol{m}_{n_{10}^{(2)}}^{(2)} \right)^2 \\
&= \frac{1}{3} \cdot \Big((11-7.75)^2 + (9-7.75)^2 + (6-7.75)^2 + (5-7.75)^2 \Big) \\
&= \frac{1}{3} \cdot (10.5625 + 1.5625 + 3.0625 + 7.5625) = \frac{22.75}{3} = \frac{91}{12} = 7.58\bar{3}.
\end{aligned}$$

1. Prognose Die Prognose der Beobachtung $\boldsymbol{x}_{11} = 4$ lautet mithilfe der initialen Klassifikationsregel (vgl. (4.30))

$$\begin{aligned}
\hat{c}_{11} &= \operatorname*{arg\,max}_{c=1,2} \left(\boldsymbol{x}_{11}^T \boldsymbol{S}_{10}^{-1} \boldsymbol{m}_{n_{10}^{(c)}}^{(c)} - \frac{1}{2} \left(\boldsymbol{m}_{n_{10}^{(c)}}^{(c)} \right)^T \boldsymbol{S}_{10}^{-1} \boldsymbol{m}_{n_{10}^{(c)}}^{(c)} + \log P_{10}^{(c)} \right) \\
&\approx \operatorname*{arg\,max}_{c=1,2} \Bigg(4 \cdot 0.2721 \cdot 3 - \frac{1}{2} \cdot 3 \cdot 0.2721 \cdot 3 + \log(0.6), \\
&\qquad\qquad 4 \cdot 0.2721 \cdot 7.75 - \frac{1}{2} \cdot 7.75 \cdot 0.2721 \cdot 7.75 + \log(0.4) \Bigg) \\
&\approx \operatorname*{arg\,max}_{c=1,2} (1.1245, -0.2472) = 1.
\end{aligned}$$

Natürlich kann an dieser Stelle die wahre Klasse $c_{11} = 3$ nicht prognostiziert werden, da diese erst zum Zeitpunkt $t = 11$ zum ersten Mal neu auftritt.

1. Aktualisierung Im Folgenden werden die benötigten Größen auf Basis der folgenden drei Beobachtungen aktualisiert, wobei eine feste Lernrate von $\lambda = 1/2$ betrachtet wird. Zunächst erfolgt eine Aktualisierung durch Beobachtung $\boldsymbol{x}_{11} = 4$ mit neu auftretender Klasse $c_{11} = 3$, also $g(\boldsymbol{x}_{11}) = 3$.

Die Anzahlen in den Klassen betragen nach Aktualisierung

$$n_{11}^{(1)} = n_{10}^{(1)} = 6, \quad n_{11}^{(2)} = n_{10}^{(2)} = 4 \quad \text{und} \quad n_{11}^{(3)} = 1.$$

Die neuen Mittelwerte der Klassen lauten mithilfe von (4.27)/(4.8):

$$\boldsymbol{m}_{n_{11}^{(1)}}^{(1)} = \boldsymbol{m}_{n_{10}^{(1)}}^{(1)} = 3, \quad \boldsymbol{m}_{n_{11}^{(2)}}^{(2)} = \boldsymbol{m}_{n_{10}^{(2)}}^{(2)} = 7.75, \quad \boldsymbol{m}_{n_{11}^{(3)}}^{(3)} = \boldsymbol{x}_{11} = 4.$$

Mithilfe von (4.28) lassen sich die geschätzten a-priori Wahrscheinlichkeiten aktualisieren:

$$P_{11}^{(1)} = \frac{(1-\lambda)n_{10}^{(1)}}{(1-\lambda)n_{10}+\lambda} = \frac{(1-\frac{1}{2})\cdot 6}{(1-\frac{1}{2})\cdot 10+\frac{1}{2}} = \frac{6}{11} = 0.\overline{54} \approx 0.5455,$$
$$P_{11}^{(2)} = \frac{(1-\lambda)n_{10}^{(2)}}{(1-\lambda)n_{10}+\lambda} = \frac{(1-\frac{1}{2})\cdot 4}{(1-\frac{1}{2})\cdot 10+\frac{1}{2}} = \frac{4}{11} = 0.\overline{36} \approx 0.3636,$$
$$P_{11}^{(3)} = \frac{\lambda}{(1-\lambda)n_{10}+\lambda} = \frac{\frac{1}{2}}{(1-\frac{1}{2})\cdot 10+\frac{1}{2}} = \frac{1}{11} = 0.\overline{09} \approx 0.0909.$$

Die neue inverse gepoolte Kovarianzmatrix innerhalb der Klassen (hier Varianz) zum Zeitpunkt $t = 11$ nach Aktualisierung durch Beobachtung $\boldsymbol{x}_{11}$ hat folgende Form (nach (4.29)):

$$\begin{aligned}\boldsymbol{S}_{11}^{-1} &= \frac{(1-\lambda)n_{10}+\lambda}{(1-\lambda)n_{10}} \cdot \boldsymbol{S}_{10}^{-1} \\ &= \frac{(1-\frac{1}{2})\cdot 10+\frac{1}{2}}{(1-\frac{1}{2})\cdot 10} \cdot \frac{1}{3.675} = \frac{11}{10}\cdot\frac{1}{3.675} = \frac{11}{36.75} = \frac{44}{147} \approx 0.2993.\end{aligned}$$

2. Prognose Die Klasse der folgenden Beobachtung $\boldsymbol{x}_{12} = 6.5$ im Datenstrom kann durch die angepasste Klassifikationsregel prognostiziert werden:

$$\begin{aligned}\hat{c}_{12} &= \underset{c=1,2,3}{\arg\max}\left(\boldsymbol{x}_{12}^T \boldsymbol{S}_{11}^{-1} \boldsymbol{m}_{n_{11}^{(c)}}^{(c)} - \frac{1}{2}\left(\boldsymbol{m}_{n_{11}^{(c)}}^{(c)}\right)^T \boldsymbol{S}_{11}^{-1} \boldsymbol{m}_{n_{11}^{(c)}}^{(c)} + \log P_{11}^{(c)}\right) \\ &\approx \underset{c=1,2,3}{\arg\max}\Bigg(6.5\cdot 0.2993\cdot 3 - \frac{1}{2}\cdot 3\cdot 0.2993\cdot 3 + \log(0.5455), \\ &\qquad\qquad 6.5\cdot 0.2993\cdot 7.75 - \frac{1}{2}\cdot 7.75\cdot 0.2993\cdot 7.75 + \log(0.3636), \\ &\qquad\qquad 6.5\cdot 0.2993\cdot 4 - \frac{1}{2}\cdot 4\cdot 0.2993\cdot 4 + \log(0.0909)\Bigg) \\ &\approx \underset{c=1,2,3}{\arg\max}(3.8834, 5.0772, 2.9894) = 2.\end{aligned}$$

2. Aktualisierung Die nächste Beobachtung $\boldsymbol{x}_{12} = 6.5$ im Datenstrom weist die Klasse $c_{12} = 1$ auf, es gilt also $g(\boldsymbol{x}_{12}) = 1$. Erneut wird die Lernrate $\lambda = 1/2$ betrachtet. Mithilfe derselben Update-Formeln (4.10), (4.27), (4.28) und (4.29) ergeben sich die aktualisierten Größen folgendermaßen:

$$n_{12}^{(1)} = n_{11}^{(1)} + 1 = 6 + 1 = 7, \quad n_{12}^{(2)} = n_{11}^{(2)} = 4, \quad n_{12}^{(3)} = n_{11}^{(3)} = 1,$$
$$\boldsymbol{m}_{n_{12}^{(1)}}^{(1)} = \frac{(1-\lambda)n_{11}^{(1)}\boldsymbol{m}_{n_{11}^{(1)}}^{(1)} + \lambda\boldsymbol{x}_{12}}{(1-\lambda)n_{11}^{(1)}+\lambda} = \frac{(1-\frac{1}{2})\cdot 6\cdot 3 + \frac{1}{2}\cdot 6.5}{(1-\frac{1}{2})\cdot 6+\frac{1}{2}} = \frac{6\cdot 3+6.5}{6+1} = \frac{24.5}{7} = 3.5,$$
$$\boldsymbol{m}_{n_{12}^{(2)}}^{(2)} = \boldsymbol{m}_{n_{11}^{(2)}}^{(2)} = 7.75, \quad \boldsymbol{m}_{n_{12}^{(3)}}^{(3)} = \boldsymbol{m}_{n_{11}^{(3)}}^{(3)} = 4,$$

$$P_{12}^{(1)} = \frac{(1-\lambda)n_{11}^{(1)} + \lambda}{(1-\lambda)n_{11} + \lambda} = \frac{(1-\frac{1}{2})\cdot 6 + \frac{1}{2}}{(1-\frac{1}{2})\cdot 11 + \frac{1}{2}} = \frac{7}{12} = 0.58\overline{3} \approx 0.5833,$$

$$P_{12}^{(2)} = \frac{(1-\lambda)n_{11}^{(2)}}{(1-\lambda)n_{11} + \lambda} = \frac{(1-\frac{1}{2})\cdot 4}{(1-\frac{1}{2})\cdot 11 + \frac{1}{2}} = \frac{4}{12} = 0.\overline{3} \approx 0.3333,$$

$$P_{12}^{(3)} = \frac{(1-\lambda)n_{11}^{(3)}}{(1-\lambda)n_{11} + \lambda} = \frac{(1-\frac{1}{2})\cdot 1}{(1-\frac{1}{2})\cdot 11 + \frac{1}{2}} = \frac{1}{12} = 0.08\overline{3} \approx 0.0833,$$

$$\begin{aligned}
\boldsymbol{z}^* &= \boldsymbol{x}_{12} - \frac{\left((1-\lambda)n_{11}^{(1)} + \lambda\right)\boldsymbol{m}^{(1)}_{n_{12}^{(1)}} - \lambda\boldsymbol{x}_{12}}{(1-\lambda)n_{11}^{(1)}} \\
&= 6.5 - \frac{\left((1-\frac{1}{2})\cdot 6 + \frac{1}{2}\right)\cdot 3.5 - \frac{1}{2}\cdot 6.5}{(1-\frac{1}{2})\cdot 6} = 3.5,
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{S}_{12}^{-1} &= \frac{(1-\lambda)n_{11} + \lambda}{(1-\lambda)n_{11}}\left(\boldsymbol{S}_{11}^{-1} - \frac{\left(\boldsymbol{S}_{11}^{-1}\right)^2(\boldsymbol{z}^*)^2}{\frac{(1-\lambda)n_{11}\left(n_{11}^{(1)}+1\right)}{\lambda n_{11}^{(1)}} + (\boldsymbol{z}^*)^2\boldsymbol{S}_{11}^{-1}}\right) \\
&= \frac{(1-\frac{1}{2})\cdot 11 + \frac{1}{2}}{(1-\frac{1}{2})\cdot 11}\left(\frac{11}{36.75} - \frac{\left(\frac{11}{36.75}\right)^2\cdot 3.5^2}{\frac{(1-\frac{1}{2})\cdot 11\cdot(6+1)}{\frac{1}{2}\cdot 6} + 3.5^2\cdot\frac{11}{36.75}}\right) \\
&= \frac{12}{11}\cdot\left(\frac{11}{36.75} - \frac{\frac{121}{1350.5625}\cdot 12.25}{\frac{77}{6} + 12.25\cdot\frac{11}{36.75}}\right) = \frac{12}{11}\cdot\left(\frac{11}{36.75} - \frac{\frac{1482.25}{1350.5625}}{\frac{77}{6} + \frac{22}{6}}\right) \\
&= 12\cdot\left(\frac{\frac{100}{49}}{75} - \frac{\frac{612.5}{1350.5625}}{75}\right) = 12\cdot\left(\frac{\frac{9}{7}}{47.25} - \frac{\frac{2}{7}}{47.25}\right) = 12\cdot\left(\frac{1}{47.25}\right) = \frac{48}{189} \approx 0.2540.
\end{aligned}$$

3. Prognose Die Prognose der folgenden Beobachtung $\boldsymbol{x}_{13} = 1$ durch die angepasste Klassifikationsregel lautet:

$$\begin{aligned}
\hat{c}_{13} &= \underset{c=1,2,3}{\arg\max}\left(\boldsymbol{x}_{13}^T\boldsymbol{S}_{12}^{-1}\boldsymbol{m}^{(c)}_{n_{12}^{(c)}} - \frac{1}{2}\left(\boldsymbol{m}^{(c)}_{n_{12}^{(c)}}\right)^T\boldsymbol{S}_{12}^{-1}\boldsymbol{m}^{(c)}_{n_{12}^{(c)}} + \log P_{12}^{(c)}\right) \\
&\approx \underset{c=1,2,3}{\arg\max}\Bigg(1\cdot 0.2540\cdot 3.5 - \frac{1}{2}\cdot 3.5\cdot 0.2540\cdot 3.5 + \log(0.5833), \\
&\qquad\qquad 1\cdot 0.2540\cdot 7.75 - \frac{1}{2}\cdot 7.75\cdot 0.2540\cdot 7.75 + \log(0.3333), \\
&\qquad\qquad 1\cdot 0.2540\cdot 4 - \frac{1}{2}\cdot 4\cdot 0.2540\cdot 4 + \log(0.0833)\Bigg) \\
&\approx \underset{c=1,2,3}{\arg\max}(-1.2058, -6.7581, -3.5013) = 1.
\end{aligned}$$

3. Aktualisierung Aktualisierung der Größen mittels der Beobachtung $\boldsymbol{x}_{13} = 1$ aus Klasse 2, also $g(\boldsymbol{x}_{13}) = 2$ führt unter Beachtung von $\lambda = 1/2$ entsprechend zu folgenden Ergebnissen:

$$n_{13}^{(1)} = n_{12}^{(1)} = 7, \quad n_{13}^{(2)} = n_{12}^{(2)} + 1 = 4 + 1 = 5, \quad n_{13}^{(3)} = n_{12}^{(3)} = 1,$$

$$
\begin{aligned}
\boldsymbol{m}^{(1)}_{n^{(1)}_{13}} &= \boldsymbol{m}^{(1)}_{n^{(1)}_{12}} = 3.5,\\
\boldsymbol{m}^{(2)}_{n^{(2)}_{13}} &= \frac{(1-\lambda)n^{(2)}_{12}\boldsymbol{m}^{(2)}_{n^{(2)}_{12}} + \lambda\boldsymbol{x}_{13}}{(1-\lambda)n^{(2)}_{12}+\lambda} = \frac{(1-\tfrac{1}{2})\cdot 4\cdot 7.75 + \tfrac{1}{2}\cdot 1}{(1-\tfrac{1}{2})\cdot 4 + \tfrac{1}{2}} = \frac{4\cdot 7.75+1}{4+1} = \frac{32}{5} = 6.4,\\
\boldsymbol{m}^{(3)}_{n^{(3)}_{13}} &= \boldsymbol{m}^{(3)}_{n^{(3)}_{12}} = 4,\\
P^{(1)}_{13} &= \frac{(1-\lambda)n^{(1)}_{12}}{(1-\lambda)n_{12}+\lambda} = \frac{(1-\tfrac{1}{2})\cdot 7}{(1-\tfrac{1}{2})\cdot 12 + \tfrac{1}{2}} = \frac{7}{13} \approx 0.5385,\\
P^{(2)}_{13} &= \frac{(1-\lambda)n^{(2)}_{12}+\lambda}{(1-\lambda)n_{12}+\lambda} = \frac{(1-\tfrac{1}{2})\cdot 4 + \tfrac{1}{2}}{(1-\tfrac{1}{2})\cdot 12 + \tfrac{1}{2}} = \frac{5}{13} \approx 0.3846,\\
P^{(3)}_{13} &= \frac{(1-\lambda)n^{(3)}_{12}}{(1-\lambda)n_{12}+\lambda} = \frac{(1-\tfrac{1}{2})\cdot 1}{(1-\tfrac{1}{2})\cdot 12 + \tfrac{1}{2}} = \frac{1}{13} \approx 0.0769,\\
\boldsymbol{z}^* &= \boldsymbol{x}_{13} - \frac{\left((1-\lambda)n^{(2)}_{12}+\lambda\right)\boldsymbol{m}^{(2)}_{n^{(2)}_{13}} - \lambda\boldsymbol{x}_{13}}{(1-\lambda)n^{(2)}_{12}}\\
&= 1 - \frac{\left((1-\tfrac{1}{2})\cdot 4 + \tfrac{1}{2}\right)\cdot 6.4 - \tfrac{1}{2}\cdot 1}{(1-\tfrac{1}{2})\cdot 4} = -6.75,\\
\boldsymbol{S}^{-1}_{13} &= \frac{(1-\lambda)n_{12}+\lambda}{(1-\lambda)n_{12}}\left(\boldsymbol{S}^{-1}_{12} - \frac{\left(\boldsymbol{S}^{-1}_{12}\right)^2(\boldsymbol{z}^*)^2}{\frac{(1-\lambda)n_{12}\left(n^{(2)}_{12}+1\right)}{\lambda n^{(2)}_{12}} + (\boldsymbol{z}^*)^2\boldsymbol{S}^{-1}_{12}}\right)\\
&= \frac{(1-\tfrac{1}{2})\cdot 12 + \tfrac{1}{2}}{(1-\tfrac{1}{2})\cdot 12}\left(\frac{48}{189} - \frac{\left(\frac{48}{189}\right)^2\cdot(-6.75)^2}{\frac{(1-\frac{1}{2})\cdot 12\cdot(4+1)}{\frac{1}{2}\cdot 4} + (-6.75)^2\cdot\frac{48}{189}}\right)\\
&= \frac{13}{12}\cdot\left(\frac{48}{189} - \frac{\frac{2304}{35721}\cdot 45.5625}{15 + 45.5625\cdot\frac{48}{189}}\right) = \frac{13}{12}\cdot\left(\frac{48}{189} - \frac{\frac{104976}{35721}}{15+\frac{2187}{189}}\right)\\
&= \frac{13}{12}\cdot\left(\frac{48}{189} - \frac{\frac{144}{49}}{\frac{186}{7}}\right) = 13\cdot\left(\frac{\frac{62}{35}}{83.7} - \frac{\frac{27}{35}}{83.7}\right) = 13\cdot\left(\frac{1}{83.7}\right) = 13\cdot\left(\frac{1}{83.7}\right) = \frac{130}{837}\\
&\approx 0.1553.
\end{aligned}
$$

Im Falle der Betrachtung von $\lambda = 1/2$ sind alle einzelnen Parameter und damit die Diskriminanzkomponenten der Aktualisierungen zu einem Zeitpunkt t im Datenstrom identisch wie jene, welche durch das LDA-Modell resultieren, das auf Basis aller Beobachtungen von Zeitpunkt 1 bis t gleichzeitig erstellt wird. Werden nämlich alle Beobachtungen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{13}$ ohne zwischenzeitliche Aktualisierungen gleichzeitig betrachtet, so ergeben sich

$$
\begin{aligned}
n^{(1)}_{13} &= 7, \quad n^{(2)}_{13} = 5 \quad \text{und} \quad n^{(3)}_{13} = 1,\\
\boldsymbol{m}^{(1)}_{n^{(1)}_{13}} &= \frac{1}{n^{(1)}_{13}}\sum_{\substack{i:\, g(\boldsymbol{x}_i)=1\\ i\le 13}}\boldsymbol{x}_i = \frac{1}{7}\cdot(1+1.5+2.5+3.5+4+5.5+6.5) = \frac{24.5}{7} = 3.5,
\end{aligned}
$$

$$\boldsymbol{m}^{(2)}_{n^{(2)}_{13}} = \frac{1}{n^{(2)}_{13}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=2 \\ i\le 13}} \boldsymbol{x}_i = \frac{1}{5}\cdot(11+9+6+5+1) = \frac{32}{5} = 6.4,$$

$$\boldsymbol{m}^{(3)}_{n^{(3)}_{13}} = \frac{1}{n^{(3)}_{13}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=3 \\ i\le 13}} \boldsymbol{x}_i = \frac{1}{1}\cdot 4 = 4,$$

$$P^{(1)}_{13} = \frac{n^{(1)}_{13}}{n_{13}} = \frac{7}{13} \approx 0.5385, \quad P^{(2)}_{13} = \frac{n^{(2)}_{13}}{n_{13}} = \frac{5}{13} \approx 0.3846, \quad P^{(3)}_{13} = \frac{n^{(3)}_{13}}{n_{13}} = \frac{1}{13} \approx 0.0769,$$

$$\begin{aligned}
\boldsymbol{S}^{-1}_{13} &= \left(\frac{1}{n_{13}} \sum_{c=1}^{3} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le 13}} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{13}} \right)^2 \right)^{-1} \\
&= \Bigg(\frac{1}{13} \cdot \Big((1-3.5)^2 + (1.5-3.5)^2 + (2.5-3.5)^2 + (3.5-3.5)^2 + (4-3.5)^2 \\
&\qquad + (5.5-3.5)^2 + (6.5-3.5)^2 + (11-6.4)^2 + (9-6.4)^2 + (6-6.4)^2 \\
&\qquad + (5-6.4)^2 + (1-6.4)^2 + (4-4)^2 \Big) \Bigg)^{-1} \\
&= \left(\frac{1}{13} \cdot 83.7 \right)^{-1} = \frac{13}{83.7} = \frac{130}{837} \approx 0.1553.
\end{aligned}$$

A.3 QDA-AF (Abschnitt 4.4)

Für das Beispiel wird wieder der eindimensionale Datenstrom aus Tabelle A.1 (s. Seite 403) betrachtet. Der Algorithmus *QDA-AF* (vgl. Seite 87) bzw. die Teilalgorithmen *M-AF* und *G-AF* werden anhand dieses Beispiel-Datenstroms schrittweise berechnet.

Die Initialisierung erfolgt aus praktischen Gründen erneut auf Basis der ersten zehn Beobachtungen. Bis zu diesem Zeitpunkt sind bereits Beobachtungen aus zwei verschiedenen Klassen aufgetreten. Es wird davon ausgegangen, dass die Gesamtzahl aller Klassen nicht bekannt ist. Es ist allerdings zu beachten, dass die Anzahl der Beobachtungen bei Initialisierung bei diesem Algorithmus irrelevant ist, da die Initialisierung datenunabhängig erfolgt. Vielmehr werden daher die ersten zehn Beobachtungen ignoriert, es wird lediglich vermerkt, welche Klassen bisher aufgetreten sind.

Initialisierung Zum Zeitpunkt der Initialisierung beträgt die Anzahl an Klassen somit $M = 2$, da $c \in \{1, 2\}$. Es müssen die Parameter für den *M-AF* und den *G-AF* Algorithmus initialisiert werden. Ersterer benötigt die folgenden Parameter:

$$
\begin{aligned}
N_0^{(0)} &= 1 \quad \text{(Normierungskonstante für die Gewichte nach (4.55))},\\
\tilde{P}_0^{(1)} &= \tilde{P}_0^{(2)} = \frac{1}{2} \quad \text{(a-priori Wahrscheinlichkeiten nach (4.54))},\\
\left(N_0^{(0)}\right)' &= 0,\\
\left(\tilde{P}_0^{(1)}\right)' &= \left(\tilde{P}_0^{(2)}\right)' = 0,\\
\left(J_0^{(0)}\right)' &= 0 \quad \text{(Gradient der NLL)},\\
\lambda_0^{(0)} &= 0.9537;\ \text{Zufallszahl aus } \lambda_- = 0.7 \leq \lambda \leq \lambda_+ = 0.999 \quad \text{(Faktor)},\\
&\qquad \text{alternativ könnte man auch die Mitte des Intervalls wählen},\\
\alpha_0^{(0)} &= 10^{-8};\ \text{Zufallszahl aus } \alpha_{\min} = 10^{-8} \leq \alpha \leq \alpha_{\max} = 10^{-6} \quad \text{(Schrittweite)}.
\end{aligned}
$$

Der Teilalgorithmus *G-AF* zur Modellierung der klassenbedingten Verteilungen basiert auf den folgenden Startwerten für $c \in \{1, 2\}$:

$$
\begin{aligned}
\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)} &= \mathbf{0}_1 = 0 \quad \text{(Mittelwertvektor der Klasse } c \text{ nach (4.38))},\\
\tilde{\boldsymbol{\Pi}}_0^{(c)} &= \mathbf{0}_{1\times 1} = 0 \quad \text{(nach (4.39))},\\
\tilde{\boldsymbol{\Sigma}}_0^{(c)} &= \mathbf{0}_{1\times 1} = 0 \quad \text{(Kovarianzmatrix von Klasse } c \text{ nach (4.40))},\\
N_0^{(c)} &= 0 \quad \text{(Normierungskonstante für die Gewichte nach (4.41))},\\
d_0^{(c)} &= -1000 \quad (= \log|\tilde{\boldsymbol{\Sigma}}_0^{(c)}|;\\
&\qquad -\delta \text{ mit } \delta \text{ groß nach Anagnostopoulos et al. (2012, S. 146) (**))},\\
\boldsymbol{G}_0^{(c)} &= 1000 \quad \left(= \left(\tilde{\boldsymbol{\Sigma}}_0^{(c)}\right)^{-1};\right.\\
&\qquad \left.\delta \boldsymbol{I}_{p\times p} \text{ mit } \delta \text{ groß nach Anagnostopoulos et al. (2012, S. 146))},
\end{aligned}
$$

$$\left(\tilde{\boldsymbol{m}}^{(c)}_{n_0^{(c)}}\right)' = \mathbf{0}_1 = 0 \quad \text{(Gradient des Mittelwertvektors der Klasse } c \text{ nach (4.44))},$$
$$\left(\tilde{\boldsymbol{\Pi}}_0^{(c)}\right)' = \mathbf{0}_{1\times 1} = 0 \quad \text{(nach (4.45))},$$
$$\left(\tilde{\boldsymbol{\Sigma}}_0^{(c)}\right)' = \mathbf{0}_{1\times 1} = 0 \quad \text{(nach (4.46))},$$
$$\left(N_0^{(c)}\right)' = 0 \quad \text{(nach (4.47))},$$
$$\left(d_0^{(c)}\right)' = 0,$$
$$\left(\boldsymbol{G}_0^{(c)}\right)' = \mathbf{0}_{1\times 1} = 0,$$
$$\alpha_0^{(c)} = 9.8 \cdot 10^{-7};\ \text{Zufallszahl aus } \alpha_{\min} = 10^{-8} \leq \alpha \leq \alpha_{\max} = 10^{-6} \quad \text{(Schrittweite)},$$
$$\lambda^{(c)}_{(n_0^{(c)})} = 0.9326;\ \text{Zufallszahl aus } \lambda_- = 0.7 \leq \lambda \leq \lambda_+ = 0.999 \quad \text{(Faktor)},$$
alternativ könnte man auch die Mitte des Intervalls wählen,
$$\left(J_0^{(c)}\right)' = 0 \quad \text{(Gradient der NLL)}.$$

1. Aktualisierung Als Erstes wird ein Update durch Beobachtung $\boldsymbol{x}_{11} = 4$ mit neuer Klasse $c_{11} = 3$, also $g(\boldsymbol{x}_{11}) = 3$ durchgeführt. Bei der Aktualisierung durch die neue Beobachtung werden die Parameter für die beiden Teilalgorithmen getrennt aktualisiert.

Sei zunächst der *M-AF* Algorithmus (vgl. Seite 83) betrachtet. Die Normierungskonstanten für die Gewichte sowie die Schätzer für die a-priori Wahrscheinlichkeiten der Klassen werden durch (4.55) und (4.54) (bzw. (4.68)) aktualisiert:

$$N_{11}^{(0)} = \lambda_0^{(0)} N_0^{(0)} + 1 = 0.9537 \cdot 1 + 1 = 1.9537,$$
$$\tilde{P}_{11}^{(1)} = \left(1 - \frac{1}{N_{11}^{(0)}}\right) \tilde{P}_0^{(1)} + \frac{1}{N_{11}^{(0)}} \cdot \mathbb{1}_{\{c_{11}=1\}} = \left(1 - \frac{1}{1.9537}\right) \cdot \frac{1}{2} + 0 \approx 0.2441,$$
$$\tilde{P}_{11}^{(2)} = \tilde{P}_{11}^{(1)}, \quad \tilde{P}_{11}^{(3)} = \frac{1}{N_{11}^{(0)}} = \frac{1}{1.9537} \approx 0.5118.$$

Die entsprechenden Gradienten können durch (4.60) und (4.59) (bzw. (4.69)) aktualisiert werden:

$$\left(N_{11}^{(0)}\right)' = \lambda_0^{(0)} \left(N_0^{(0)}\right)' + N_0^{(0)} = 0.9537 \cdot 0 + 1 = 1,$$
$$\left(\tilde{P}_{11}^{(1)}\right)' = \frac{N_{11}^{(0)} - 1}{N_{11}^{(0)}} \cdot \left(\tilde{P}_0^{(1)}\right)' - \frac{\left(N_{11}^{(0)}\right)'}{\left(N_{11}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{11}=1\}} - \tilde{P}_0^{(1)}\right)$$
$$= \frac{1.9537 - 1}{1.9537} \cdot 0 - \frac{1}{1.9537^2} \cdot \left(0 - \frac{1}{2}\right) \approx 0.1310, \quad \left(\tilde{P}_{11}^{(2)}\right)' = \left(\tilde{P}_{11}^{(1)}\right)',$$
$$\left(\tilde{P}_{11}^{(3)}\right)' = 0.$$

Der neue Gradient der NLL lautet mithilfe von (4.58)

$$\left(J_{11}^{(0)}\right)' = -\sum_{c=1}^{M}\left(\mathbb{1}_{\{c_{11}=c\}} - \tilde{P}_0^{(c)}\right)\frac{\left(\tilde{P}_0^{(c)}\right)'}{\tilde{P}_0^{(c)}} = -\sum_{c=1}^{2}\left(\mathbb{1}_{\{3=c\}} - \frac{1}{2}\right)\cdot\frac{0}{\frac{1}{2}} = 0.$$

Mithilfe dieses Gradienten ergibt sich der neue Faktor durch den Gradientenabstieg (4.49):

$$\lambda_{11}^{(0)} = \left[\lambda_0^{(0)} - \alpha_0^{(0)}\left(J_{11}^{(0)}\right)'\right]_{\lambda_-}^{\lambda_+} = \left[0.9537 - 10^{-8}\cdot 0\right]_{0.7}^{0.999} = 0.9537.$$

Für die Schrittweite wird der *RPROP*-Algorithmus (4.50) herangezogen:

$$\alpha_{11}^{(0)} = \alpha_0^{(0)} = 10^{-8}, \text{ da } \left|\left(J_{11}^{(0)}\right)'\right| = 0 \leq 10^{-7}.$$

Bei dem *G-AF* Algorithmus (vgl. Seite 82) werden die einzelnen Parameter pro Klasse c aktualisiert. Zunächst werden auch hier die aktuellen Normierungskonstanten für die Gewichte benötigt, welche durch (4.41) bestimmt werden können:

$$N_{11}^{(1)} = N_0^{(1)} = 0, \quad N_{11}^{(2)} = N_0^{(2)} = 0, \quad N_{11}^{(3)} = 1.$$

Die Mittelwertvektoren können durch (4.38) aktualisiert bzw. initialisiert werden:

$$\tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)} = \tilde{\boldsymbol{m}}_{n_0^{(1)}}^{(1)} = 0, \quad \tilde{\boldsymbol{m}}_{n_{11}^{(2)}}^{(2)} = \tilde{\boldsymbol{m}}_{n_0^{(2)}}^{(2)} = 0, \quad \tilde{\boldsymbol{m}}_{n_{11}^{(3)}}^{(3)} = \boldsymbol{x}_{11} = 4.$$

Die neuen Kovarianzmatrizen berechnen sich durch (4.39) und (4.40):

$$\begin{aligned}
\tilde{\boldsymbol{\Pi}}_{11}^{(1)} &= \tilde{\boldsymbol{\Pi}}_0^{(1)} = 0, & \tilde{\boldsymbol{\Pi}}_{11}^{(2)} &= \tilde{\boldsymbol{\Pi}}_0^{(2)} = 0, & \tilde{\boldsymbol{\Pi}}_{11}^{(3)} &= 0,\\
\tilde{\boldsymbol{\Sigma}}_{11}^{(1)} &= \tilde{\boldsymbol{\Sigma}}_0^{(1)} = 0, & \tilde{\boldsymbol{\Sigma}}_{11}^{(2)} &= \tilde{\boldsymbol{\Sigma}}_0^{(2)} = 0, & \tilde{\boldsymbol{\Sigma}}_{11}^{(3)} &= 0.
\end{aligned}$$

Für die jeweiligen Gradienten können die Aktualisierungsformeln (4.44)–(4.47) (bzw. (4.70)–(4.73)) herangezogen werden:

$$\begin{aligned}
\left(N_{11}^{(1)}\right)' &= \left(N_0^{(1)}\right)' = 0, & \left(N_{11}^{(2)}\right)' &= \left(N_0^{(2)}\right)' = 0, & \left(N_{11}^{(3)}\right)' &= 0,\\
\left(\tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)' &= \left(\tilde{\boldsymbol{m}}_{n_0^{(1)}}^{(1)}\right)' = 0, & \left(\tilde{\boldsymbol{m}}_{n_{11}^{(2)}}^{(2)}\right)' &= \left(\tilde{\boldsymbol{m}}_{n_0^{(2)}}^{(2)}\right)' = 0, & \left(\tilde{\boldsymbol{m}}_{n_{11}^{(3)}}^{(3)}\right)' &= 0,\\
\left(\tilde{\boldsymbol{\Pi}}_{11}^{(1)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_0^{(1)}\right)' = 0, & \left(\tilde{\boldsymbol{\Pi}}_{11}^{(2)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_0^{(2)}\right)' = 0, & \left(\tilde{\boldsymbol{\Pi}}_{11}^{(3)}\right)' &= 0,\\
\left(\tilde{\boldsymbol{\Sigma}}_{11}^{(1)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_0^{(1)}\right)' = 0, & \left(\tilde{\boldsymbol{\Sigma}}_{11}^{(2)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_0^{(2)}\right)' = 0, & \left(\tilde{\boldsymbol{\Sigma}}_{11}^{(3)}\right)' &= 0.
\end{aligned}$$

Zur Aktualisierung von $d_t^{(c)} := \log\left|\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right|$, $\boldsymbol{G}_t^{(c)} := \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}$, $\left(d_t^{(c)}\right)'$ und $\left(\boldsymbol{G}_t^{(c)}\right)'$ werden im Allgemeinen weitere Hilfsfunktionen herangezogen. Für die Formeln sei auf Anagnostopoulos et al. (2012, S. 145) verwiesen.

Hier ist es jedoch so, dass $d_t^{(c)}$ und $\boldsymbol{G}_t^{(c)}$ für den Fall $N_t^{(c)} = 1$ nicht definiert sind. $N_t^{(c)}$ nimmt genau dann den Wert 1 an, wenn erst eine Beobachtung aus der jeweiligen Klasse c im Datenstrom aufgetreten ist, da mit $N_0^{(c)} = 0$ initialisiert wird (vgl. (4.41)).

Die Größen $d_t^{(c)}$, $\left(d_t^{(c)}\right)'$, $\boldsymbol{G}_t^{(c)}$ und $\left(\boldsymbol{G}_t^{(c)}\right)'$ bleiben daher für die Klassen $c = 1, 2$ zunächst erhalten und werden für Klasse $c = 3$ initialisiert:

$$\begin{aligned}
d_{11}^{(1)} &= d_0^{(1)} = -1000, & d_{11}^{(2)} &= d_0^{(2)} = -1000, & d_{11}^{(3)} &= -1000,\\
\boldsymbol{G}_{11}^{(1)} &= \boldsymbol{G}_0^{(1)} = 1000, & \boldsymbol{G}_{11}^{(2)} &= \boldsymbol{G}_0^{(2)} = 1000, & \boldsymbol{G}_{11}^{(3)} &= 1000,\\
\left(d_{11}^{(1)}\right)' &= \left(d_0^{(1)}\right)' = 0, & \left(d_{11}^{(2)}\right)' &= \left(d_0^{(2)}\right)' = 0, & \left(d_{11}^{(3)}\right)' &= 0,\\
\left(\boldsymbol{G}_{11}^{(1)}\right)' &= \left(\boldsymbol{G}_0^{(1)}\right)' = 0, & \left(\boldsymbol{G}_{11}^{(2)}\right)' &= \left(\boldsymbol{G}_0^{(2)}\right)' = 0, & \left(\boldsymbol{G}_{11}^{(3)}\right)' &= 0.
\end{aligned}$$

Die Gradienten der NLL je Klasse bleiben erhalten bzw. jener für Klasse 3 wird initialisiert:

$$\left(J_{11}^{(1)}\right)' = \left(J_0^{(1)}\right)' = 0, \quad \left(J_{11}^{(2)}\right)' = \left(J_0^{(2)}\right)' = 0, \quad \left(J_{11}^{(3)}\right)' = 0.$$

Die einzelnen Faktoren sehen folgendermaßen aus:

$$\begin{aligned}
&\lambda_{(n_{11}^{(1)})}^{(1)} = \lambda_{(n_0^{(1)})}^{(1)} = 0.9326, \quad \lambda_{(n_{11}^{(2)})}^{(2)} = \lambda_{(n_0^{(2)})}^{(2)} = 0.9326,\\
&\lambda_{(n_{11}^{(3)})}^{(3)} = 0.9567 \text{ (Zufallszahl aus } \lambda_- = 0.7 \leq \lambda \leq \lambda_+ = 0.999;\\
&\qquad\qquad\qquad\qquad \text{alternativ könnte man auch die Mitte des Intervalls wählen).}
\end{aligned}$$

Für die Schrittweiten gilt hier:

$$\begin{aligned}
&\alpha_{11}^{(1)} = \alpha_0^{(1)} = 9.8 \cdot 10^{-7}, \quad \alpha_{11}^{(2)} = \alpha_0^{(2)} = 9.8 \cdot 10^{-7},\\
&\alpha_{11}^{(3)} = 2.2 \cdot 10^{-7} \text{ (Zufallszahl aus } \alpha_{\min} = 10^{-8} \leq \alpha \leq \alpha_{\max} = 10^{-6}).
\end{aligned}$$

1. Prognose Sobald im Datenstrom beide Teilalgorithmen auf Basis einer neuen Beobachtung durchgeführt sind, kann die Klassifikationsregel der Quadratischen Diskriminanzanalyse aufgestellt werden. Mithilfe der rekursiv bestimmten Schätzer erfolgt die Vorhersage einer neuen Beobachtung mithilfe der Klassifikationsregel (4.63). Sei die nächste Beobachtung im Datenstrom $\boldsymbol{x}_{12} = 6.5$ mit $c_{12} = 1$ betrachtet. Die prognostizierte Klasse beträgt

$$\begin{aligned}
\tilde{c}_{12} &= \underset{c=1,2,3}{\arg\min} \left(\frac{1}{2} d_{11}^{(c)} + \frac{1}{2} \left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{11}^{(c)}}^{(c)} \right)^T \boldsymbol{G}_{11}^{(c)} \left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{11}^{(c)}}^{(c)} \right) - \log \tilde{P}_{11}^{(c)} \right)\\
&\approx \underset{c=1,2,3}{\arg\min} \Bigg(\frac{1}{2} \cdot (-1000) + \frac{1}{2} \cdot (6.5 - 0)^2 \cdot 1000 - \log(0.2441),\\
&\qquad\qquad \frac{1}{2} \cdot (-1000) + \frac{1}{2} \cdot (6.5 - 0)^2 \cdot 1000 - \log(0.2441),\\
&\qquad\qquad \frac{1}{2} \cdot (-1000) + \frac{1}{2} \cdot (6.5 - 4)^2 \cdot 1000 - \log(0.5118) \Bigg)\\
&\approx \underset{c=1,2,3}{\arg\min} \left(20626.4102, 20626.4102, 2625.6698\right) = 3.
\end{aligned}$$

Es ist klar, dass hier zunächst im ersten Schritt die Klasse $\tilde{c}_{12} = 3$ prognostiziert wird. Es wurde vor der ersten Aktualisierung lediglich initialisiert, wobei wie oben beschrieben bei dieser Methode dazu keine Beobachtungen herangezogen werden. Im Datenstrom wird dann bei der ersten Aktualisierung bzw. Aufstellung der Klassifikationsregel durch die Beobachtung $\boldsymbol{x}_{11}$ zunächst lediglich die Klasse 3 beobachtet.

Die Anpassung an einen vorliegenden concept drift wird erst im Laufe des Datenstroms deutlich, wenn bereits einige Beobachtungen zum Training der Regel herangezogen werden konnten. Dies zeigt sich in den Simulationsergebnissen in Abschnitt 9.6.

2. Aktualisierung In einem zweiten Schritt erfolgt eine Aktualisierung durch die Beobachtung $\boldsymbol{x}_{12} = 6.5$ mit $c_{12} = 1$. Das bedeutet, dass nun $M = 3$ Klassen vorliegen.

Für den *M-AF* Algorithmus werden zunächst die Normierungskonstante für die Gewichte sowie die Schätzer für die a-priori Wahrscheinlichkeiten der Klassen durch (4.55) und (4.54) (bzw. (4.68)) aktualisiert:

$$
\begin{aligned}
N_{12}^{(0)} &= \lambda_{11}^{(0)} N_{11}^{(0)} + 1 = 0.9537 \cdot 1.9537 + 1 \approx 2.8632, \\
\tilde{P}_{12}^{(1)} &= \left(1 - \frac{1}{N_{12}^{(0)}}\right) \tilde{P}_{11}^{(1)} + \frac{1}{N_{12}^{(0)}} \cdot \mathbb{1}_{\{c_{12}=1\}} \approx \left(1 - \frac{1}{2.8632}\right) \cdot 0.2441 + \frac{1}{2.8632} \cdot 1 \approx 0.5081, \\
\tilde{P}_{12}^{(2)} &= \left(1 - \frac{1}{N_{12}^{(0)}}\right) \tilde{P}_{11}^{(2)} + \frac{1}{N_{12}^{(0)}} \cdot \mathbb{1}_{\{c_{12}=2\}} \approx \left(1 - \frac{1}{2.8632}\right) \cdot 0.2441 + 0 \approx 0.1588, \\
\tilde{P}_{12}^{(3)} &= \left(1 - \frac{1}{N_{12}^{(0)}}\right) \tilde{P}_{11}^{(3)} + \frac{1}{N_{12}^{(0)}} \cdot \mathbb{1}_{\{c_{12}=3\}} \approx \left(1 - \frac{1}{2.8632}\right) \cdot 0.5118 + 0 \approx 0.3330.
\end{aligned}
$$

Die entsprechenden Gradienten können durch (4.60) und (4.59) (bzw. (4.69)) aktualisiert werden:

$$
\begin{aligned}
\left(N_{12}^{(0)}\right)' &= \lambda_{11}^{(0)} \left(N_{11}^{(0)}\right)' + N_{11}^{(0)} = 0.9537 \cdot 1 + 1.9537 = 2.9074, \\
\left(\tilde{P}_{12}^{(1)}\right)' &= \frac{N_{12}^{(0)} - 1}{N_{12}^{(0)}} \cdot \left(\tilde{P}_{11}^{(1)}\right)' - \frac{\left(N_{12}^{(0)}\right)'}{\left(N_{12}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{12}=1\}} - \tilde{P}_{11}^{(1)}\right) \\
&\approx \frac{2.8632 - 1}{2.8632} \cdot 0.1310 - \frac{2.9074}{2.8632^2} \cdot (1 - 0.2441) \approx -0.1828, \\
\left(\tilde{P}_{12}^{(2)}\right)' &= \frac{N_{12}^{(0)} - 1}{N_{12}^{(0)}} \cdot \left(\tilde{P}_{11}^{(2)}\right)' - \frac{\left(N_{12}^{(0)}\right)'}{\left(N_{12}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{12}=2\}} - \tilde{P}_{11}^{(2)}\right) \\
&\approx \frac{2.8632 - 1}{2.8632} \cdot 0.1310 - \frac{2.9074}{2.8632^2} \cdot (0 - 0.2441) \approx 0.1718, \\
\left(\tilde{P}_{12}^{(3)}\right)' &= \frac{N_{12}^{(0)} - 1}{N_{12}^{(0)}} \cdot \left(\tilde{P}_{11}^{(3)}\right)' - \frac{\left(N_{12}^{(0)}\right)'}{\left(N_{12}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{12}=3\}} - \tilde{P}_{11}^{(3)}\right) \\
&\approx \frac{2.8632 - 1}{2.8632} \cdot 0 - \frac{2.9074}{2.8632^2} \cdot (0 - 0.5118) \approx 0.1815.
\end{aligned}
$$

Der neue Gradient der NLL lautet mithilfe von (4.58)

$$\begin{aligned}\left(J_{12}^{(0)}\right)' &= -\sum_{c=1}^{M}\left(\mathbb{1}_{\{c_{12}=c\}} - \tilde{P}_{11}^{(c)}\right)\frac{\left(\tilde{P}_{11}^{(c)}\right)'}{\tilde{P}_{11}^{(c)}} = -\sum_{c=1}^{3}\left(\mathbb{1}_{\{1=c\}} - \tilde{P}_{11}^{(c)}\right)\frac{\left(\tilde{P}_{11}^{(c)}\right)'}{\tilde{P}_{11}^{(c)}} \\ &\approx -\left((1-0.2441)\cdot\frac{0.1310}{0.2441} + (0-0.2441)\cdot\frac{0.1310}{0.2441} + (0-0.5118)\cdot\frac{0}{0.5118}\right) \\ &\approx -(0.4057 + (-0.1310) + 0) = -0.2747.\end{aligned}$$

Mithilfe dieses Gradienten ergibt sich der neue Faktor durch den Gradientenabstieg (4.49):

$$\lambda_{12}^{(0)} = \left[\lambda_{11}^{(0)} - \alpha_{11}^{(0)}\left(J_{12}^{(0)}\right)'\right]_{\lambda_-}^{\lambda_+} \approx \left[0.9537 - 10^{-8}\cdot(-0.2747)\right]_{0.7}^{0.999} \approx 0.9537.$$

Für die Schrittweite wird der *RPROP*-Algorithmus (4.50) herangezogen:

$$\begin{aligned}&\alpha_{12}^{(0)} = \left[0.99\,\alpha_{11}^{(0)}\right]_{\alpha_{\min}}^{\alpha_{\max}} = \left[0.99\cdot 10^{-8}\right]_{10^{-8}}^{10^{-6}} = 10^{-8}, \\ &\text{da } \left|\left(J_{12}^{(0)}\right)'\right| \approx 0.2747 > 10^{-7} \text{ und } \left(J_{12}^{(0)}\right)'\left(J_{11}^{(0)}\right)' \approx -0.2747\cdot 0 = 0.\end{aligned}$$

Beim *G-AF* Algorithmus werden auch zunächst aktuelle Normierungskonstanten für die Gewichte benötigt, welche durch (4.41) bestimmt werden können:

$$N_{12}^{(1)} = \lambda_{(n_{11}^{(1)})}^{(1)} N_{11}^{(1)} + 1 = 0.9326\cdot 0 + 1 = 1, \quad N_{12}^{(2)} = N_{11}^{(2)} = 0, \quad N_{12}^{(3)} = N_{11}^{(3)} = 1.$$

Die Mittelwertvektoren können durch (4.38) aktualisiert werden:

$$\begin{aligned}&\tilde{\boldsymbol{m}}_{n_{12}^{(1)}}^{(1)} = \left(1 - \frac{1}{N_{12}^{(1)}}\right)\tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)} + \frac{1}{N_{12}^{(1)}}\cdot\boldsymbol{x}_{12} = \left(1-\frac{1}{1}\right)\cdot 0 + \frac{1}{1}\cdot 6.5 = 6.5, \\ &\tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)} = \tilde{\boldsymbol{m}}_{n_{11}^{(2)}}^{(2)} = 0, \quad \tilde{\boldsymbol{m}}_{n_{12}^{(3)}}^{(3)} = \tilde{\boldsymbol{m}}_{n_{11}^{(3)}}^{(3)} = 4.\end{aligned}$$

Die neuen Kovarianzmatrizen berechnen sich durch (4.39) und (4.40):

$$\begin{aligned}&\tilde{\boldsymbol{\Pi}}_{12}^{(1)} = \left(1 - \frac{1}{N_{12}^{(1)}}\right)\tilde{\boldsymbol{\Pi}}_{11}^{(1)} + \frac{1}{N_{12}^{(1)}}\cdot\boldsymbol{x}_{12}\boldsymbol{x}_{12}^T = \left(1-\frac{1}{1}\right)\cdot 0 + \frac{1}{1}\cdot 6.5^2 = 42.25, \\ &\tilde{\boldsymbol{\Pi}}_{12}^{(2)} = \tilde{\boldsymbol{\Pi}}_{11}^{(2)} = 0, \quad \tilde{\boldsymbol{\Pi}}_{12}^{(3)} = \tilde{\boldsymbol{\Pi}}_{11}^{(3)} = 0, \\ &\tilde{\boldsymbol{\Sigma}}_{12}^{(1)} = \tilde{\boldsymbol{\Pi}}_{12}^{(1)} - \tilde{\boldsymbol{m}}_{n_{12}^{(1)}}^{(1)}\left(\tilde{\boldsymbol{m}}_{n_{12}^{(1)}}^{(1)}\right)^T = 42.25 - 6.5^2 = 0, \quad \tilde{\boldsymbol{\Sigma}}_{12}^{(2)} = \tilde{\boldsymbol{\Sigma}}_{11}^{(2)} = 0, \quad \tilde{\boldsymbol{\Sigma}}_{12}^{(3)} = \tilde{\boldsymbol{\Sigma}}_{11}^{(3)} = 0.\end{aligned}$$

Für die jeweiligen Gradienten können die Aktualisierungsformeln (4.44)–(4.47) (bzw. (4.70)–(4.73)) herangezogen werden:

$$\begin{aligned}
\left(N_{12}^{(1)}\right)' &= \lambda^{(1)}_{(n_{11}^{(1)})}\left(N_{11}^{(1)}\right)' + N_{11}^{(1)} = 0.9326 \cdot 0 + 0 = 0,\\
\left(N_{12}^{(2)}\right)' &= \left(N_{11}^{(2)}\right)' = 0, \quad \left(N_{12}^{(3)}\right)' = \left(N_{11}^{(3)}\right)' = 0,\\
\left(\tilde{\boldsymbol{m}}^{(1)}_{n_{12}^{(1)}}\right)' &= \left(1 - \frac{1}{N_{12}^{(1)}}\right)\left(\tilde{\boldsymbol{m}}^{(1)}_{n_{11}^{(1)}}\right)' - \frac{\left(N_{12}^{(1)}\right)'}{\left(N_{12}^{(1)}\right)^2}\left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}^{(1)}_{n_{11}^{(1)}}\right)\\
&= \left(1 - \frac{1}{1}\right) \cdot 0 - \frac{0}{1^2} \cdot (6.5 - 0) = 0,\\
\left(\tilde{\boldsymbol{m}}^{(2)}_{n_{12}^{(2)}}\right)' &= \left(\tilde{\boldsymbol{m}}^{(2)}_{n_{11}^{(2)}}\right)' = 0, \quad \left(\tilde{\boldsymbol{m}}^{(3)}_{n_{12}^{(3)}}\right)' = \left(\tilde{\boldsymbol{m}}^{(3)}_{n_{11}^{(3)}}\right)' = 0,\\
\left(\tilde{\boldsymbol{\Pi}}_{12}^{(1)}\right)' &= \left(1 - \frac{1}{N_{12}^{(1)}}\right)\left(\tilde{\boldsymbol{\Pi}}_{11}^{(1)}\right)' - \frac{\left(N_{12}^{(1)}\right)'}{\left(N_{12}^{(1)}\right)^2}\left(\boldsymbol{x}_{12}\boldsymbol{x}_{12}^T - \tilde{\boldsymbol{\Pi}}_{11}^{(1)}\right)\\
&= \left(1 - \frac{1}{1}\right) \cdot 0 - \frac{0}{1^2} \cdot \left(6.5^2 - 0\right) = 0,\\
\left(\tilde{\boldsymbol{\Pi}}_{12}^{(2)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_{11}^{(2)}\right)' = 0, \quad \left(\tilde{\boldsymbol{\Pi}}_{12}^{(3)}\right)' = \left(\tilde{\boldsymbol{\Pi}}_{11}^{(3)}\right)' = 0,\\
\left(\tilde{\boldsymbol{\Sigma}}_{12}^{(1)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_{12}^{(1)}\right)' - \left(\tilde{\boldsymbol{m}}^{(1)}_{n_{12}^{(1)}}\right)'\left(\tilde{\boldsymbol{m}}^{(1)}_{n_{12}^{(1)}}\right)^T - \tilde{\boldsymbol{m}}^{(1)}_{n_{12}^{(1)}}\left(\left(\tilde{\boldsymbol{m}}^{(1)}_{n_{12}^{(1)}}\right)'\right)^T\\
&= 0 - 0 \cdot 6.5 - 6.5 \cdot 0 = 0,\\
\left(\tilde{\boldsymbol{\Sigma}}_{12}^{(2)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_{11}^{(2)}\right)' = 0, \quad \left(\tilde{\boldsymbol{\Sigma}}_{12}^{(3)}\right)' = \left(\tilde{\boldsymbol{\Sigma}}_{11}^{(3)}\right)' = 0.
\end{aligned}$$

Zur Aktualisierung von $d_t^{(c)} := \log\left|\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right|$, $\boldsymbol{G}_t^{(c)} := \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)}\right)^{-1}$, $\left(d_t^{(c)}\right)'$ und $\left(\boldsymbol{G}_t^{(c)}\right)'$ werden im Allgemeinen weitere Hilfsfunktionen herangezogen. Für die Formeln sei auf Anagnostopoulos et al. (2012, S. 145) verwiesen.

Hier ist es jedoch so, dass $d_t^{(c)}$ und $\boldsymbol{G}_t^{(c)}$ für den Fall $N_t^{(c)} = 1$ nicht definiert sind. $N_t^{(c)}$ nimmt genau dann den Wert 1 an, wenn erst eine Beobachtung aus der jeweiligen Klasse c im Datenstrom aufgetreten ist, da mit $N_0^{(c)} = 0$ initialisiert wird (vgl. (4.41)).

Da in diesem Fall jedoch auch weiterhin $\tilde{\boldsymbol{\Sigma}}_t^{(c)} = 0$, sind $-\delta$ und $\delta \boldsymbol{I}_{p\times p}$ für einen großen Wert δ gute Approximationen für $d_t^{(c)}$ und $\boldsymbol{G}_t^{(c)}$. Die Aktualisierung dieser Größen wird daher erst mit der zweiten auftretenden Beobachtung aus einer Klasse aufgenommen, während bei der ersten weiterhin die initialen Werte betrachtet werden. Daher gilt für die aktualisierten Werte:

$$\begin{aligned}
d_{12}^{(1)} &= d_{11}^{(1)} = -1000, & d_{12}^{(2)} &= d_{11}^{(2)} = -1000, & d_{12}^{(3)} &= d_{11}^{(3)} = -1000,\\
\boldsymbol{G}_{12}^{(1)} &= \boldsymbol{G}_{11}^{(1)} = 1000, & \boldsymbol{G}_{12}^{(2)} &= \boldsymbol{G}_{11}^{(2)} = 1000, & \boldsymbol{G}_{12}^{(3)} &= \boldsymbol{G}_{11}^{(3)} = 1000,\\
\left(d_{12}^{(1)}\right)' &= \left(d_{11}^{(1)}\right)' = 0, & \left(d_{12}^{(2)}\right)' &= \left(d_{11}^{(2)}\right)' = 0, & \left(d_{12}^{(3)}\right)' &= \left(d_{11}^{(3)}\right)' = 0,\\
\left(\boldsymbol{G}_{12}^{(1)}\right)' &= \left(\boldsymbol{G}_{11}^{(1)}\right)' = 0, & \left(\boldsymbol{G}_{12}^{(2)}\right)' &= \left(\boldsymbol{G}_{11}^{(2)}\right)' = 0, & \left(\boldsymbol{G}_{12}^{(3)}\right)' &= \left(\boldsymbol{G}_{11}^{(3)}\right)' = 0.
\end{aligned}$$

Der Gradient der NLL der Klasse 1 berechnet sich durch (4.43) bzw. (4.61):

$$\begin{aligned}\left(J_{12}^{(1)}\right)' &= \frac{1}{2}\left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)^T \left(-2\,\boldsymbol{G}_{11}^{(1)}\left(\tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)' + \left(\boldsymbol{G}_{11}^{(1)}\right)'\left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)\right) + \frac{1}{2}\left(d_{11}^{(1)}\right)' \\ &= \frac{1}{2}\cdot(6.5-0)\cdot(-2\cdot 1000\cdot 0 + 0\cdot(6.5-0)) + \frac{1}{2}\cdot 0 = 0, \\ \left(J_{12}^{(2)}\right)' &= \left(J_{11}^{(2)}\right)' = 0, \quad \left(J_{12}^{(3)}\right)' = \left(J_{11}^{(3)}\right)' = 0.\end{aligned}$$

Der Gradientenabstieg (4.49) dient zur Aktualisierung des Faktors

$$\begin{aligned}\lambda_{(n_{12}^{(1)})}^{(1)} &= \left[\lambda_{(n_{11}^{(1)})}^{(1)} - \alpha_{11}^{(1)}\left(J_{12}^{(1)}\right)'\right]_{\lambda_-}^{\lambda_+} = \left[0.9326 - 9.8\cdot 10^{-7}\cdot 0\right]_{0.7}^{0.999} = 0.9326, \\ \lambda_{(n_{12}^{(2)})}^{(2)} &= \lambda_{(n_{11}^{(2)})}^{(2)} = 0.9326, \quad \lambda_{(n_{12}^{(3)})}^{(3)} = \lambda_{(n_{11}^{(3)})}^{(3)} = 0.9567\end{aligned}$$

und zur Aktualisierung der Schrittweite wird der *RPROP*-Algorithmus (4.50) herangezogen:

$$\begin{aligned}\alpha_{12}^{(1)} &= \alpha_{11}^{(1)} = 9.8\cdot 10^{-7}, \quad \text{da } \left|\left(J_{12}^{(1)}\right)'\right| = 0 \leq 10^{-7}, \\ \alpha_{12}^{(2)} &= \alpha_{11}^{(2)} = 9.8\cdot 10^{-7}, \quad \alpha_{12}^{(3)} = \alpha_{11}^{(3)} = 2.2\cdot 10^{-7}.\end{aligned}$$

2. Prognose Mithilfe der rekursiv bestimmten Schätzer erfolgt die Vorhersage einer neuen Beobachtung mithilfe der Klassifikationsregel (4.63). Sei die nächste Beobachtung im Datenstrom $\boldsymbol{x}_{13} = 1$ mit $c_{13} = 2$ betrachtet. Die prognostizierte Klasse beträgt

$$\begin{aligned}\tilde{c}_{13} &= \underset{c=1,2,3}{\arg\min}\left(\frac{1}{2}\,d_{12}^{(c)} + \frac{1}{2}\left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{12}^{(c)}}^{(c)}\right)^T \boldsymbol{G}_{12}^{(c)}\left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{12}^{(c)}}^{(c)}\right) - \log\tilde{P}_{12}^{(c)}\right) \\ &\approx \underset{c=1,2,3}{\arg\min}\Bigg(\frac{1}{2}\cdot(-1000) + \frac{1}{2}\cdot(1-6.5)^2\cdot 1000 - \log(0.5081), \\ &\qquad\qquad \frac{1}{2}\cdot(-1000) + \frac{1}{2}\cdot(1-0)^2\cdot 1000 - \log(0.1588), \\ &\qquad\qquad \frac{1}{2}\cdot(-1000) + \frac{1}{2}\cdot(1-4)^2\cdot 1000 - \log(0.3330)\Bigg) \\ &\approx \underset{c=1,2,3}{\arg\min}\,(14625.6771, 1.8401, 4001.09961) = 2.\end{aligned}$$

3. Aktualisierung Die neuen Größen durch Aktualisierung durch die Beobachtung $\boldsymbol{x}_{13} = 1$ mit $c_{13} = 2$ berechnen sich folgendermaßen analog:

- *M-AF:*

$$\begin{aligned}N_{13}^{(0)} &= \lambda_{12}^{(0)} N_{12}^{(0)} + 1 \approx 0.9537\cdot 2.8632 + 1 \approx 3.7306, \\ \tilde{P}_{13}^{(1)} &= \left(1 - \frac{1}{N_{13}^{(0)}}\right)\tilde{P}_{12}^{(1)} + \frac{1}{N_{13}^{(0)}}\cdot \mathbb{1}_{\{c_{13}=1\}} \approx \left(1 - \frac{1}{3.7306}\right)\cdot 0.5081 + 0 \approx 0.3719,\end{aligned}$$

$$\begin{aligned}
\tilde{P}_{13}^{(2)} &= \left(1-\frac{1}{N_{13}^{(0)}}\right)\tilde{P}_{12}^{(2)}+\frac{1}{N_{13}^{(0)}}\cdot\mathbb{1}_{\{c_{13}=2\}}\\
&\approx\left(1-\frac{1}{3.7306}\right)\cdot 0.1588+\frac{1}{3.7306}\cdot 1\approx 0.3843,\\
\tilde{P}_{13}^{(3)} &= \left(1-\frac{1}{N_{13}^{(0)}}\right)\tilde{P}_{12}^{(3)}+\frac{1}{N_{13}^{(0)}}\cdot\mathbb{1}_{\{c_{13}=3\}}\approx\left(1-\frac{1}{3.7306}\right)\cdot 0.3330+0\approx 0.2438,\\
\left(N_{13}^{(0)}\right)' &= \lambda_{12}^{(0)}\left(N_{12}^{(0)}\right)'+N_{12}^{(0)}\approx 0.9537\cdot 2.9074+2.8632\approx 5.6360,\\
\left(\tilde{P}_{13}^{(1)}\right)' &= \frac{N_{13}^{(0)}-1}{N_{13}^{(0)}}\cdot\left(\tilde{P}_{12}^{(1)}\right)'-\frac{\left(N_{13}^{(0)}\right)'}{\left(N_{13}^{(0)}\right)^2}\left(\mathbb{1}_{\{c_{13}=1\}}-\tilde{P}_{12}^{(1)}\right)\\
&\approx\frac{3.7306-1}{3.7306}\cdot(-0.1828)-\frac{5.6360}{3.7306^2}\cdot(0-0.5081)\approx 0.0720,\\
\left(\tilde{P}_{13}^{(2)}\right)' &= \frac{N_{13}^{(0)}-1}{N_{13}^{(0)}}\cdot\left(\tilde{P}_{12}^{(2)}\right)'-\frac{\left(N_{13}^{(0)}\right)'}{\left(N_{13}^{(0)}\right)^2}\left(\mathbb{1}_{\{c_{13}=2\}}-\tilde{P}_{12}^{(2)}\right)\\
&\approx\frac{3.7306-1}{3.7306}\cdot 0.1718-\frac{5.6360}{3.7306^2}\cdot(1-0.1588)\approx -0.2149,\\
\left(\tilde{P}_{13}^{(3)}\right)' &= \frac{N_{13}^{(0)}-1}{N_{13}^{(0)}}\cdot\left(\tilde{P}_{12}^{(3)}\right)'-\frac{\left(N_{13}^{(0)}\right)'}{\left(N_{13}^{(0)}\right)^2}\left(\mathbb{1}_{\{c_{13}=3\}}-\tilde{P}_{12}^{(3)}\right)\\
&\approx\frac{3.7306-1}{3.7306}\cdot 0.1815-\frac{5.6360}{3.7306^2}\cdot(0-0.3330)\approx 0.2677.
\end{aligned}$$

Der neue Gradient der NLL lautet mithilfe von (4.58)

$$\begin{aligned}
\left(J_{13}^{(0)}\right)' &= -\sum_{c=1}^{M}\left(\mathbb{1}_{\{c_{13}=c\}}-\tilde{P}_{12}^{(c)}\right)\frac{\left(\tilde{P}_{12}^{(c)}\right)'}{\tilde{P}_{12}^{(c)}}=-\sum_{c=1}^{3}\left(\mathbb{1}_{\{2=c\}}-\tilde{P}_{12}^{(c)}\right)\frac{\left(\tilde{P}_{12}^{(c)}\right)'}{\tilde{P}_{12}^{(c)}}\\
&\approx-\Bigg((0-0.5081)\cdot\frac{-0.1828}{0.5081}+(1-0.1588)\cdot\frac{0.1718}{0.1588}\\
&\qquad+(0-0.3330)\cdot\frac{0.1815}{0.3330}\Bigg)\\
&\approx-(0.1828+0.9101+(-0.1815))=-0.9114.
\end{aligned}$$

Mithilfe dieses Gradienten ergibt sich der neue Faktor durch den Gradientenabstieg (4.49):

$$\lambda_{13}^{(0)}=\left[\lambda_{12}^{(0)}-\alpha_{12}^{(0)}\left(J_{13}^{(0)}\right)'\right]_{\lambda_-}^{\lambda_+}\approx\left[0.9537-10^{-8}\cdot(-0.9114)\right]_{0.7}^{0.999}\approx 0.9537.$$

Für die Schrittweite wird der *RPROP*-Algorithmus (4.50) herangezogen:

$$\alpha_{13}^{(0)}=\left[1.01\,\alpha_{12}^{(0)}\right]_{\alpha_{\min}}^{\alpha_{\max}}=\left[1.01\cdot 10^{-8}\right]_{10^{-8}}^{10^{-6}}=1.01\cdot 10^{-8},$$

da $\left|\left(J_{13}^{(0)}\right)'\right|\approx 0.9114>10^{-7}$ u. $\left(J_{13}^{(0)}\right)'\left(J_{12}^{(0)}\right)'\approx -0.9114\cdot(-0.2747)\approx 0.2504>0.$

- *G-AF:*

$$N_{13}^{(1)} = N_{12}^{(1)} = 1, \quad N_{13}^{(2)} = \lambda^{(2)}_{(n_{12}^{(2)})} N_{12}^{(2)} + 1 = 0.9326 \cdot 0 + 1 = 1, \quad N_{13}^{(3)} = N_{12}^{(3)} = 1,$$

$$\tilde{\boldsymbol{m}}^{(1)}_{n_{13}^{(1)}} = \tilde{\boldsymbol{m}}^{(1)}_{n_{12}^{(1)}} = 6.5,$$

$$\tilde{\boldsymbol{m}}^{(2)}_{n_{13}^{(2)}} = \left(1 - \frac{1}{N_{13}^{(2)}}\right) \tilde{\boldsymbol{m}}^{(2)}_{n_{12}^{(2)}} + \frac{1}{N_{13}^{(2)}} \cdot \boldsymbol{x}_{13} = \left(1 - \frac{1}{1}\right) \cdot 0 + \frac{1}{1} \cdot 1 = 1,$$

$$\tilde{\boldsymbol{m}}^{(3)}_{n_{13}^{(3)}} = \tilde{\boldsymbol{m}}^{(3)}_{n_{12}^{(3)}} = 4,$$

$$\tilde{\boldsymbol{\Pi}}_{13}^{(1)} = \tilde{\boldsymbol{\Pi}}_{12}^{(1)} = 42.25,$$

$$\tilde{\boldsymbol{\Pi}}_{13}^{(2)} = \left(1 - \frac{1}{N_{13}^{(2)}}\right) \tilde{\boldsymbol{\Pi}}_{12}^{(2)} + \frac{1}{N_{13}^{(2)}} \cdot \boldsymbol{x}_{13}\boldsymbol{x}_{13}^T = \left(1 - \frac{1}{1}\right) \cdot 0 + \frac{1}{1} \cdot 1^2 = 1,$$

$$\tilde{\boldsymbol{\Pi}}_{13}^{(3)} = \tilde{\boldsymbol{\Pi}}_{12}^{(3)} = 0,$$

$$\tilde{\boldsymbol{\Sigma}}_{13}^{(1)} = \tilde{\boldsymbol{\Sigma}}_{12}^{(1)} = 0,$$

$$\tilde{\boldsymbol{\Sigma}}_{13}^{(2)} = \tilde{\boldsymbol{\Pi}}_{13}^{(2)} - \tilde{\boldsymbol{m}}^{(2)}_{n_{13}^{(2)}} \left(\tilde{\boldsymbol{m}}^{(2)}_{n_{13}^{(2)}}\right)^T = 1 - 1^2 = 0, \quad \tilde{\boldsymbol{\Sigma}}_{13}^{(3)} = \tilde{\boldsymbol{\Sigma}}_{12}^{(3)} = 0.$$

Für die jeweiligen Gradienten gilt:

$$\left(N_{13}^{(1)}\right)' = \left(N_{12}^{(1)}\right)' = 0,$$

$$\left(N_{13}^{(2)}\right)' = \lambda^{(2)}_{(n_{12}^{(2)})} \left(N_{12}^{(2)}\right)' + N_{12}^{(2)} = 0.9326 \cdot 0 + 0 = 0,$$

$$\left(N_{13}^{(3)}\right)' = \left(N_{12}^{(3)}\right)' = 0,$$

$$\left(\tilde{\boldsymbol{m}}^{(1)}_{n_{13}^{(1)}}\right)' = \left(\tilde{\boldsymbol{m}}^{(1)}_{n_{12}^{(1)}}\right)' = 0,$$

$$\begin{aligned}\left(\tilde{\boldsymbol{m}}^{(2)}_{n_{13}^{(2)}}\right)' &= \left(1 - \frac{1}{N_{13}^{(2)}}\right) \left(\tilde{\boldsymbol{m}}^{(2)}_{n_{12}^{(2)}}\right)' - \frac{\left(N_{13}^{(2)}\right)'}{\left(N_{13}^{(2)}\right)^2} \left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}^{(2)}_{n_{12}^{(2)}}\right) \\ &= \left(1 - \frac{1}{1}\right) \cdot 0 - \frac{0}{1^2} \cdot (1 - 0) = 0,\end{aligned}$$

$$\left(\tilde{\boldsymbol{m}}^{(3)}_{n_{13}^{(3)}}\right)' = \left(\tilde{\boldsymbol{m}}^{(3)}_{n_{12}^{(3)}}\right)' = 0,$$

$$\left(\tilde{\boldsymbol{\Pi}}_{13}^{(1)}\right)' = \left(\tilde{\boldsymbol{\Pi}}_{12}^{(1)}\right)' = 0,$$

$$\begin{aligned}\left(\tilde{\boldsymbol{\Pi}}_{13}^{(2)}\right)' &= \left(1 - \frac{1}{N_{13}^{(2)}}\right) \left(\tilde{\boldsymbol{\Pi}}_{12}^{(2)}\right)' - \frac{\left(N_{13}^{(2)}\right)'}{\left(N_{13}^{(2)}\right)^2} \left(\boldsymbol{x}_{13}\boldsymbol{x}_{13}^T - \tilde{\boldsymbol{\Pi}}_{12}^{(2)}\right) \\ &= \left(1 - \frac{1}{1}\right) \cdot 0 - \frac{0}{1^2} \cdot \left(1^2 - 0\right) = 0,\end{aligned}$$

$$\left(\tilde{\boldsymbol{\Pi}}_{13}^{(3)}\right)' = \left(\tilde{\boldsymbol{\Pi}}_{12}^{(3)}\right)' = 0,$$

$$\left(\tilde{\boldsymbol{\Sigma}}_{13}^{(1)}\right)' = \left(\tilde{\boldsymbol{\Sigma}}_{12}^{(1)}\right)' = 0,$$

$$\begin{aligned}
\left(\tilde{\boldsymbol{\Sigma}}_{13}^{(2)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_{13}^{(2)}\right)' - \left(\tilde{\boldsymbol{m}}_{n_{13}^{(2)}}^{(2)}\right)' \left(\tilde{\boldsymbol{m}}_{n_{13}^{(2)}}^{(2)}\right)^T - \tilde{\boldsymbol{m}}_{n_{13}^{(2)}}^{(2)} \left(\left(\tilde{\boldsymbol{m}}_{n_{13}^{(2)}}^{(2)}\right)'\right)^T \\
&= 0 - 0 \cdot 1 - 1 \cdot 0 = 0, \\
\left(\tilde{\boldsymbol{\Sigma}}_{13}^{(3)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_{12}^{(3)}\right)' = 0.
\end{aligned}$$

Auch zu diesem Zeitpunkt liegt erst eine Beobachtung aus jeder Klasse vor, d. h. $n_{13}^{(c)} = 1,\ c = 1, 2, 3,$ (und auch $N_{13}^{(c)} = 1$). Bei der ersten Beobachtung werden weiterhin die initialen Werte für d, d', $\boldsymbol{G}$ und $\boldsymbol{G}'$ betrachtet (s. o.). Daher auch hier:

$$\begin{aligned}
d_{13}^{(1)} &= d_{12}^{(1)} = -1000, & d_{13}^{(2)} &= d_{12}^{(2)} = -1000, & d_{13}^{(3)} &= d_{12}^{(3)} = -1000, \\
\boldsymbol{G}_{13}^{(1)} &= \boldsymbol{G}_{12}^{(1)} = 1000, & \boldsymbol{G}_{13}^{(2)} &= \boldsymbol{G}_{12}^{(2)} = 1000, & \boldsymbol{G}_{13}^{(2)} &= \boldsymbol{G}_{12}^{(2)} = 1000, \\
\left(d_{13}^{(1)}\right)' &= \left(d_{12}^{(1)}\right)' = 0, & \left(d_{13}^{(2)}\right)' &= \left(d_{12}^{(2)}\right)' = 0, & \left(d_{13}^{(2)}\right)' &= \left(d_{12}^{(2)}\right)' = 0, \\
\left(\boldsymbol{G}_{13}^{(1)}\right)' &= \left(\boldsymbol{G}_{12}^{(1)}\right)' = 0, & \left(\boldsymbol{G}_{13}^{(2)}\right)' &= \left(\boldsymbol{G}_{12}^{(2)}\right)' = 0, & \left(\boldsymbol{G}_{13}^{(3)}\right)' &= \left(\boldsymbol{G}_{12}^{(3)}\right)' = 0.
\end{aligned}$$

Der neue Gradient der NLL der Klasse 2 berechnet sich durch (4.43) bzw. (4.61):

$$\begin{aligned}
\left(J_{13}^{(1)}\right)' &= \left(J_{12}^{(1)}\right)' = 0, \\
\left(J_{13}^{(2)}\right)' &= \frac{1}{2}\left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)}\right)^T \left(-2\boldsymbol{G}_{12}^{(2)} \left(\tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)}\right)' + \left(\boldsymbol{G}_{12}^{(2)}\right)' \left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)}\right)\right) \\
&\quad + \frac{1}{2}\left(d_{12}^{(2)}\right)' \\
&= \frac{1}{2} \cdot (1-1) \cdot (-2 \cdot 1000 \cdot 0 + 0 \cdot (1-1)) + \frac{1}{2} \cdot 0 = 0, \\
\left(J_{13}^{(3)}\right)' &= \left(J_{12}^{(3)}\right)' = 0.
\end{aligned}$$

Der Gradientenabstieg (4.49) dient zur Aktualisierung des Faktors

$$\begin{aligned}
\lambda_{(n_{13}^{(1)})}^{(1)} &= \lambda_{(n_{12}^{(1)})}^{(1)} = 0.9326, \\
\lambda_{(n_{13}^{(2)})}^{(2)} &= \left[\lambda_{(n_{12}^{(2)})}^{(2)} - \alpha_{12}^{(2)} \left(J_{13}^{(2)}\right)'\right]_{\lambda_-}^{\lambda_+} = \left[0.9326 - 9.8 \cdot 10^{-7} \cdot 0\right]_{0.7}^{0.999} = 0.9326, \\
\lambda_{(n_{13}^{(3)})}^{(3)} &= \lambda_{(n_{12}^{(3)})}^{(3)} = 0.9567
\end{aligned}$$

und zur Aktualisierung der Schrittweite wird auch hier der *RPROP*-Algorithmus (4.50) herangezogen:

$$\begin{aligned}
\alpha_{13}^{(1)} &= \alpha_{12}^{(1)} = 9.8 \cdot 10^{-7}, \\
\alpha_{13}^{(2)} &= \alpha_{12}^{(2)} = 9.8 \cdot 10^{-7}, \quad \text{da } \left|\left(J_{13}^{(2)}\right)'\right| = 0 \leq 10^{-7}, \\
\alpha_{13}^{(3)} &= \alpha_{12}^{(3)} = 2.2 \cdot 10^{-7}.
\end{aligned}$$

A.4 LDA-AF (Abschnitt 4.4)

Auch für den Algorithmus *LDA-AF* (vgl. Seite 88) müssen schrittweise die Parameter mittels der Teilalgorithmen *G-AF* und *M-AF* aktualisiert werden, wobei hier im Gegensatz zu *QDA-AF* immer noch ein weiterer Durchlauf von *G-AF* erfolgt, um die gepoolte Kovarianzmatrix anzupassen.

Initialisierung Als Initialisierung werden die folgenden Werte für den *M-AF* Algorithmus betrachtet:

$$
\begin{aligned}
N_0^{(0)} &= 1 \quad \text{(Normierungskonstante für die Gewichte nach (4.55))},\\
\tilde{P}_0^{(1)} &= \tilde{P}_0^{(2)} = \frac{1}{2} \quad \text{(a-priori Wahrscheinlichkeiten nach (4.54))},\\
\left(N_0^{(0)}\right)' &= 0,\\
\left(\tilde{P}_0^{(1)}\right)' &= \left(\tilde{P}_0^{(2)}\right)' = 0,\\
\left(J_0^{(0)}\right)' &= 0 \quad \text{(Gradient der NLL)},\\
\lambda_0^{(0)} &= 0.9537;\ \text{Zufallszahl aus } \lambda_- = 0.7 \leq \lambda \leq \lambda_+ = 0.999 \quad \text{(Faktor)},\\
&\qquad \text{alternativ könnte man auch die Mitte des Intervalls wählen},\\
\alpha_0^{(0)} &= 10^{-8};\ \text{Zufallszahl aus } \alpha_{\min} = 10^{-8} \leq \alpha \leq \alpha_{\max} = 10^{-6} \quad \text{(Schrittweite)}.
\end{aligned}
$$

Der Teilalgorithmus *G-AF* zur Modellierung der klassenbedingten Verteilungen basiert auf den folgenden Startwerten für $c \in \{1, 2\}$:

$$
\begin{aligned}
N_0^{(c)} &= 0 \quad \text{(Normierungskonstante für die Gewichte nach (4.41))},\\
\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)} &= \mathbf{0}_1 = 0 \quad \text{(Mittelwertvektor der Klasse } c \text{ nach (4.38))},\\
\tilde{\boldsymbol{\Pi}}_0^{(c)} &= \mathbf{0}_{1\times 1} = 0 \quad \text{(nach (4.39))},\\
\tilde{\boldsymbol{\Sigma}}_0^{(c)} &= \mathbf{0}_{1\times 1} = 0 \quad \text{(Kovarianzmatrix von Klasse } c \text{ nach (4.40))},\\
d_0^{(c)} &= -1000 \quad (= \log|\tilde{\boldsymbol{\Sigma}}_0^{(c)}|;\\
&\qquad -\delta \text{ mit } \delta \text{ groß nach Anagnostopoulos et al. (2012, S. 146) } (**)),\\
\boldsymbol{G}_0^{(c)} &= 1000 \quad \left(= \left(\tilde{\boldsymbol{\Sigma}}_0^{(c)}\right)^{-1};\right.\\
&\qquad \left.\delta \boldsymbol{I}_{p\times p} \text{ mit } \delta \text{ groß nach Anagnostopoulos et al. (2012, S. 146))},\\
\left(N_0^{(c)}\right)' &= 0 \quad \text{(nach (4.47))},\\
\left(\tilde{\boldsymbol{m}}_{n_0^{(c)}}^{(c)}\right)' &= \mathbf{0}_1 = 0 \quad \text{(Gradient des Mittelwertvektors der Klasse } c \text{ nach (4.44))},\\
\left(\tilde{\boldsymbol{\Pi}}_0^{(c)}\right)' &= \mathbf{0}_{1\times 1} = 0 \quad \text{(nach (4.45))},\\
\left(\tilde{\boldsymbol{\Sigma}}_0^{(c)}\right)' &= \mathbf{0}_{1\times 1} = 0 \quad \text{(nach (4.46))},
\end{aligned}
$$

$$\left(d_0^{(c)}\right)' = 0,$$
$$\left(\boldsymbol{G}_0^{(c)}\right)' = \mathbf{0}_{1\times 1} = 0,$$
$$\alpha_0^{(c)} = 8.6 \cdot 10^{-7};\ \text{Zufallszahl aus } \alpha_{\min} = 10^{-8} \leq \alpha \leq \alpha_{\max} = 10^{-6} \quad \text{(Schrittweite)},$$
$$\lambda^{(c)}_{(n_0^{(c)})} = 0.9356;\ \text{Zufallszahl aus } \lambda_- = 0.7 \leq \lambda \leq \lambda_+ = 0.999 \quad \text{(Faktor)},$$
alternativ könnte man auch die Mitte des Intervalls wählen,
$$\left(J_0^{(c)}\right)' = 0 \quad \text{(Gradient der NLL)}.$$

Für den Durchlauf des *G-AF* Algorithmus zur Anpassung der gepoolten Kovarianzmatrix werden analog die folgenden Startwerte betrachtet:

$$N_0^{(P)} = 0,$$
$$\tilde{\boldsymbol{m}}_{n_0}^{(P)} = 0,$$
$$\tilde{\boldsymbol{\Pi}}_0^{(P)} = \mathbf{0}_{1\times 1} = 0,$$
$$\tilde{\boldsymbol{\Sigma}}_0^{(P)} = \mathbf{0}_{1\times 1} = 0,$$
$$d_0^{(P)} = -1000,$$
$$\boldsymbol{G}_0^{(P)} = 1000,$$
$$\left(N_0^{(P)}\right)' = 0,$$
$$\left(\tilde{\boldsymbol{m}}_{n_0}^{(P)}\right)' = \mathbf{0}_1 = 0,$$
$$\left(\tilde{\boldsymbol{\Pi}}_0^{(P)}\right)' = \mathbf{0}_{1\times 1} = 0,$$
$$\left(\tilde{\boldsymbol{\Sigma}}_0^{(P)}\right)' = \mathbf{0}_{1\times 1} = 0,$$
$$\left(d_0^{(P)}\right)' = 0,$$
$$\left(\boldsymbol{G}_0^{(P)}\right)' = \mathbf{0}_{1\times 1} = 0,$$
$$\alpha_0^{(P)} = 9.8 \cdot 10^{-7},$$
$$\lambda_0^{(P)} = 0.9326,$$
$$\left(J_0^{(P)}\right)' = 0.$$

1. Aktualisierung Als Erstes wird ein Update durch Beobachtung $\boldsymbol{x}_{11} = 4$ mit neuer Klasse $c_{11} = 3$, also $g(\boldsymbol{x}_{11}) = 3$ durchgeführt. Bei der Aktualisierung durch die neue Beobachtung werden die beiden Teilalgorithmen nacheinander betrachtet. Zunächst werden die entsprechenden Parameter mittels des *M-AF* Algorithmus (vgl. Seite 83) aktualisiert.

Die Normierungskonstanten für die Gewichte sowie die Schätzer für die a-priori Wahrscheinlichkeiten der Klassen werden durch (4.55) und (4.54) (bzw. (4.68)) aktualisiert:

$$N_{11}^{(0)} = \lambda_0^{(0)} N_0^{(0)} + 1 = 0.9537 \cdot 1 + 1 = 1.9537,$$
$$\tilde{P}_{11}^{(1)} = \left(1 - \frac{1}{N_{11}^{(0)}}\right) \tilde{P}_0^{(1)} + \frac{1}{N_{11}^{(0)}} \cdot \mathbb{1}_{\{c_{11}=1\}} = \left(1 - \frac{1}{1.9537}\right) \cdot \frac{1}{2} + 0 \approx 0.2441,$$
$$\tilde{P}_{11}^{(2)} = \tilde{P}_{11}^{(1)}, \quad \tilde{P}_{11}^{(3)} = \frac{1}{N_{11}^{(0)}} = \frac{1}{1.9537} \approx 0.5118.$$

Die entsprechenden Gradienten können durch (4.60) und (4.59) (bzw. (4.69)) aktualisiert werden:

$$\left(N_{11}^{(0)}\right)' = \lambda_0^{(0)} \left(N_0^{(0)}\right)' + N_0^{(0)} = 0.9537 \cdot 0 + 1 = 1,$$
$$\left(\tilde{P}_{11}^{(1)}\right)' = \frac{N_{11}^{(0)} - 1}{N_{11}^{(0)}} \cdot \left(\tilde{P}_0^{(1)}\right)' - \frac{\left(N_{11}^{(0)}\right)'}{\left(N_{11}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{11}=1\}} - \tilde{P}_0^{(1)}\right)$$
$$= \frac{1.9537 - 1}{1.9537} \cdot 0 - \frac{1}{1.9537^2} \cdot \left(0 - \frac{1}{2}\right) \approx 0.1310, \quad \left(\tilde{P}_{11}^{(2)}\right)' = \left(\tilde{P}_{11}^{(1)}\right)',$$

$$\left(\tilde{P}_{11}^{(3)}\right)' = 0.$$

Der neue Gradient der NLL lautet mithilfe von (4.58)

$$\left(J_{11}^{(0)}\right)' = -\sum_{c=1}^{M}\left(\mathbb{1}_{\{c_{11}=c\}} - \tilde{P}_0^{(c)}\right)\frac{\left(\tilde{P}_0^{(c)}\right)'}{\tilde{P}_0^{(c)}} = -\sum_{c=1}^{2}\left(\mathbb{1}_{\{3=c\}} - \frac{1}{2}\right)\cdot\frac{0}{\frac{1}{2}} = 0.$$

Mithilfe dieses Gradienten ergibt sich der neue Faktor durch den Gradientenabstieg (4.49):

$$\lambda_{11}^{(0)} = \left[\lambda_0^{(0)} - \alpha_0^{(0)}\left(J_{11}^{(0)}\right)'\right]_{\lambda_-}^{\lambda_+} = \left[0.9537 - 10^{-8}\cdot 0\right]_{0.7}^{0.999} = 0.9537.$$

Für die Schrittweite wird der *RPROP*-Algorithmus (4.50) herangezogen:

$$\alpha_{11}^{(0)} = \alpha_0^{(0)} = 10^{-8},\ \text{da}\ \left|\left(J_{11}^{(0)}\right)'\right| = 0 \leq 10^{-7}.$$

Bei dem *G-AF* Algorithmus (vgl. Seite 82) zur Aktualisierung der klassenbedingten Verteilungen werden die einzelnen Parameter pro Klasse c aktualisiert. Die Kovarianzmatrix wird dabei festgehalten, daher:

$$\text{fix}_{\boldsymbol{\Sigma}} = \texttt{TRUE},\ \boldsymbol{\Sigma}_0 = \tilde{\boldsymbol{\Sigma}}_0^{(P)} = 0,\ \boldsymbol{\Pi}_0 = \tilde{\boldsymbol{\Pi}}_0^{(P)} = 0,\ d_0 = d_0^{(P)} = -1000,\ \boldsymbol{G}_0 = \boldsymbol{G}_0^{(P)} = 1000.$$

Zunächst werden auch hier die aktuellen Normierungskonstanten für die Gewichte benötigt, welche durch (4.41) bestimmt werden können:

$$N_{11}^{(1)} = N_0^{(1)} = 0, \quad N_{11}^{(2)} = N_0^{(2)} = 0, \quad N_{11}^{(3)} = 1.$$

Die Mittelwertvektoren können durch (4.38) aktualisiert bzw. initialisiert werden:

$$\tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)} = \tilde{\boldsymbol{m}}_{n_0^{(1)}}^{(1)} = 0, \quad \tilde{\boldsymbol{m}}_{n_{11}^{(2)}}^{(2)} = \tilde{\boldsymbol{m}}_{n_0^{(2)}}^{(2)} = 0, \quad \tilde{\boldsymbol{m}}_{n_{11}^{(3)}}^{(3)} = \boldsymbol{x}_{11} = 4.$$

Für die jeweiligen Gradienten gilt (vgl. (4.70) und (4.73)):

$$\left(N_{11}^{(1)}\right)' = \left(N_0^{(1)}\right)' = 0, \qquad \left(N_{11}^{(2)}\right)' = \left(N_0^{(2)}\right)' = 0, \qquad \left(N_{11}^{(3)}\right)' = 0,$$
$$\left(\tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)' = \left(\tilde{\boldsymbol{m}}_{n_0^{(1)}}^{(1)}\right)' = 0, \qquad \left(\tilde{\boldsymbol{m}}_{n_{11}^{(2)}}^{(2)}\right)' = \left(\tilde{\boldsymbol{m}}_{n_0^{(2)}}^{(2)}\right)' = 0, \qquad \left(\tilde{\boldsymbol{m}}_{n_{11}^{(3)}}^{(3)}\right)' = 0.$$

Da $\text{fix}_{\boldsymbol{\Sigma}} = \texttt{TRUE}$, gilt für die Kovarianzmatrizen und die Gradienten hier (vgl. Algorithmen 1 und 4 auf Seiten 82/88; die Größen für Klasse 1 und 2 bleiben erhalten):

$$
\begin{aligned}
\tilde{\boldsymbol{\Pi}}_{11}^{(1)} &= \tilde{\boldsymbol{\Pi}}_{0}^{(1)} = 0, & \tilde{\boldsymbol{\Pi}}_{11}^{(2)} &= \tilde{\boldsymbol{\Pi}}_{0}^{(2)} = 0, & \tilde{\boldsymbol{\Pi}}_{11}^{(3)} &= \boldsymbol{\Pi}_{0} = 0,\\
\tilde{\boldsymbol{\Sigma}}_{11}^{(1)} &= \tilde{\boldsymbol{\Sigma}}_{0}^{(1)} = 0, & \tilde{\boldsymbol{\Sigma}}_{11}^{(2)} &= \tilde{\boldsymbol{\Sigma}}_{0}^{(2)} = 0, & \tilde{\boldsymbol{\Sigma}}_{11}^{(3)} &= \boldsymbol{\Sigma}_{0} = 0,\\
d_{11}^{(1)} &= d_{0}^{(1)} = -1000, & d_{11}^{(2)} &= d_{0}^{(2)} = -1000, & d_{11}^{(3)} &= d_{0} = -1000,\\
\boldsymbol{G}_{11}^{(1)} &= \boldsymbol{G}_{0}^{(1)} = 1000, & \boldsymbol{G}_{11}^{(2)} &= \boldsymbol{G}_{0}^{(2)} = 1000, & \boldsymbol{G}_{11}^{(3)} &= \boldsymbol{G}_{0} = 1000,\\
\left(\tilde{\boldsymbol{\Pi}}_{11}^{(1)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_{0}^{(1)}\right)' = 0, & \left(\tilde{\boldsymbol{\Pi}}_{11}^{(2)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_{0}^{(2)}\right)' = 0, & \left(\tilde{\boldsymbol{\Pi}}_{11}^{(3)}\right)' &= 0,\\
\left(\tilde{\boldsymbol{\Sigma}}_{11}^{(1)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_{0}^{(1)}\right)' = 0, & \left(\tilde{\boldsymbol{\Sigma}}_{11}^{(2)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_{0}^{(2)}\right)' = 0, & \left(\tilde{\boldsymbol{\Sigma}}_{11}^{(3)}\right)' &= 0,\\
\left(d_{11}^{(1)}\right)' &= \left(d_{0}^{(1)}\right)' = 0, & \left(d_{11}^{(2)}\right)' &= \left(d_{0}^{(2)}\right)' = 0, & \left(d_{11}^{(3)}\right)' &= 0,\\
\left(\boldsymbol{G}_{11}^{(1)}\right)' &= \left(\boldsymbol{G}_{0}^{(1)}\right)' = 0, & \left(\boldsymbol{G}_{11}^{(2)}\right)' &= \left(\boldsymbol{G}_{0}^{(2)}\right)' = 0, & \left(\boldsymbol{G}_{11}^{(3)}\right)' &= 0.
\end{aligned}
$$

Die Gradienten in Klasse 3 werden dabei auf 0 gesetzt.

Die Gradienten der NLL je Klasse bleiben erhalten bzw. jener für Klasse 3 wird initialisiert:

$$\left(J_{11}^{(1)}\right)' = \left(J_{0}^{(1)}\right)' = 0, \quad \left(J_{11}^{(2)}\right)' = \left(J_{0}^{(2)}\right)' = 0, \quad \left(J_{11}^{(3)}\right)' = 0.$$

Die einzelnen Faktoren sehen dann folgendermaßen aus:

$$
\begin{aligned}
&\lambda_{(n_{11}^{(1)})}^{(1)} = \lambda_{(n_{0}^{(1)})}^{(1)} = 0.9356, \quad \lambda_{(n_{11}^{(2)})}^{(2)} = \lambda_{(n_{0}^{(2)})}^{(2)} = 0.9356,\\
&\lambda_{(n_{11}^{(3)})}^{(3)} = 0.8903 \text{ (Zufallszahl aus } \lambda_{-} = 0.7 \leq \lambda \leq \lambda_{+} = 0.999;\\
&\qquad\qquad \text{alternativ könnte man auch die Mitte des Intervalls wählen).}
\end{aligned}
$$

Für die Schrittweiten gilt hier:

$$
\begin{aligned}
&\alpha_{11}^{(1)} = \alpha_{0}^{(1)} = 8.6 \cdot 10^{-7}, \quad \alpha_{11}^{(2)} = \alpha_{0}^{(2)} = 8.6 \cdot 10^{-7},\\
&\alpha_{11}^{(3)} = 9.3 \cdot 10^{-7} \text{ (Zufallszahl aus } \alpha_{\min} = 10^{-8} \leq \alpha \leq \alpha_{\max} = 10^{-6}).
\end{aligned}
$$

Zuletzt wird der neue Datenpunkt zentriert durch $\boldsymbol{\xi}_{11} = \boldsymbol{x}_{11} - \tilde{\boldsymbol{m}}_{n_{11}^{(3)}}^{(3)} = 4 - 4 = 0$.

Zur Aktualisierung der gepoolten Kovarianzmatrix wird ein weiterer Durchlauf des *G-AF* Algorithmus mit $\boldsymbol{\xi}_{11} = 0$ betrachtet, wobei hier $\text{fix}_{\boldsymbol{m}} = \texttt{TRUE}$ und $\boldsymbol{m}_0 = 0$ gesetzt wird.

Die aktuelle Normierungskonstante für die Gewichte wird auch hier durch (4.41) bestimmt, der entsprechende Gradient durch (4.73):

$$
\begin{aligned}
N_{11}^{(P)} &= \lambda_{0}^{(P)} N_{0}^{(P)} + 1 = 0.9326 \cdot 0 + 1 = 1,\\
\left(N_{11}^{(P)}\right)' &= \lambda_{0}^{(P)} \left(N_{0}^{(P)}\right)' + N_{0}^{(P)} = 0.9326 \cdot 0 + 0 = 0.
\end{aligned}
$$

Da $\text{fix}_{\boldsymbol{m}} = \texttt{TRUE}$, gilt für den neuen Mittelwertvektor sowie den Gradienten:

$$\tilde{\boldsymbol{m}}_{n_{11}}^{(P)} = \boldsymbol{m}_0 = 0, \quad \left(\tilde{\boldsymbol{m}}_{n_{11}}^{(P)}\right)' = 0.$$

Die neue Kovarianzmatrix berechnet sich durch (4.39) und (4.40):

$$\tilde{\boldsymbol{\Pi}}_{11}^{(P)} = \left(1 - \frac{1}{N_{11}^{(P)}}\right) \tilde{\boldsymbol{\Pi}}_{0}^{(P)} + \frac{1}{N_{11}^{(P)}} \cdot \boldsymbol{\xi}_{11}\, \boldsymbol{\xi}_{11}^{T} = \left(1 - \frac{1}{1}\right) \cdot 0 + \frac{1}{1} \cdot 0^2 = 0,$$
$$\tilde{\boldsymbol{\Sigma}}_{11}^{(P)} = \tilde{\boldsymbol{\Pi}}_{11}^{(P)} - \tilde{\boldsymbol{m}}_{n_{11}}^{(P)} \left(\tilde{\boldsymbol{m}}_{n_{11}}^{(P)}\right)^{T} = 0 - 0^2 = 0.$$

Wie bei *QDA-AF* wird die Aktualisierung der Größen d und $\boldsymbol{G}$ erst mit der zweiten auftretenden Beobachtung aus einer Klasse aufgenommen, während bei der ersten weiterhin die initialen Werte betrachtet werden. Daher:

$$d_{11}^{(P)} = d_{0}^{(P)} = -1000, \qquad \boldsymbol{G}_{11}^{(P)} = \boldsymbol{G}_{0}^{(P)} = 1000,$$
$$\left(d_{11}^{(P)}\right)' = \left(d_{0}^{(P)}\right)' = 0, \qquad \left(\boldsymbol{G}_{11}^{(P)}\right)' = \left(\boldsymbol{G}_{0}^{(P)}\right)' = 0.$$

Der Gradient der NLL ergibt sich durch (4.43) bzw. (4.61):

$$\begin{aligned}\left(J_{11}^{(P)}\right)' &= \frac{1}{2} \left(\boldsymbol{\xi}_{11} - \tilde{\boldsymbol{m}}_{n_0}^{(P)}\right)^{T} \left(-2\, \boldsymbol{G}_{0}^{(P)} \left(\tilde{\boldsymbol{m}}_{n_0}^{(P)}\right)' + \left(\boldsymbol{G}_{0}^{(P)}\right)' \left(\boldsymbol{\xi}_{11} - \tilde{\boldsymbol{m}}_{n_0}^{(P)}\right)\right) + \frac{1}{2} \left(d_{0}^{(P)}\right)' \\ &= \frac{1}{2} \cdot (0 - 0) \cdot (-2 \cdot 1000 \cdot 0 + 0 \cdot (0 - 0)) + \frac{1}{2} \cdot 0 = 0.\end{aligned}$$

Für den neuen Faktor gilt mit (4.49):

$$\lambda_{11}^{(P)} = \left[\lambda_{0}^{(P)} - \alpha_{0}^{(P)} \left(J_{11}^{(P)}\right)'\right]_{\lambda_-}^{\lambda_+} = \left[0.9326 - 9.8 \cdot 10^{-7} \cdot 0\right]_{0.7}^{0.999} = 0.9326.$$

Die neue Schrittweite ergibt sich hier durch (4.50):

$$\alpha_{11}^{(P)} = \alpha_{0}^{(P)} = 9.8 \cdot 10^{-7}, \quad \text{da } \left|\left(J_{11}^{(P)}\right)'\right| = 0 < 10^{-7}.$$

1. Prognose Sobald im Datenstrom alle nötigen Durchläufe der Teilalgorithmen auf Basis einer neuen Beobachtung durchgeführt sind, kann die Klassifikationsregel der Linearen Diskriminanzanalyse aufgestellt werden und neue Beobachtungen können prognostiziert werden. Mithilfe der rekursiv bestimmten Schätzer erfolgt die Vorhersage einer neuen Beobachtung dabei mithilfe der Klassifikationsregel (4.64). Sei die nächste Beobachtung im Datenstrom $\boldsymbol{x}_{12} = 6.5$ mit $c_{12} = 1$ betrachtet. Die prognostizierte Klasse beträgt

$$\begin{aligned}\tilde{c}_{12} &= \underset{c=1,2,3}{\arg\min} \left(\frac{1}{2}\, d_{11}^{(P)} + \frac{1}{2} \left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{11}^{(c)}}^{(c)}\right)^{T} \boldsymbol{G}_{11}^{(P)} \left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{11}^{(c)}}^{(c)}\right) - \log \tilde{P}_{11}^{(c)}\right) \\ &\approx \underset{c=1,2,3}{\arg\min} \Bigg(\frac{1}{2} \cdot (-1000) + \frac{1}{2} \cdot (6.5 - 0)^2 \cdot 1000 - \log(0.2441), \\ &\qquad \frac{1}{2} \cdot (-1000) + \frac{1}{2} \cdot (6.5 - 0)^2 \cdot 1000 - \log(0.2441), \\ &\qquad \frac{1}{2} \cdot (-1000) + \frac{1}{2} \cdot (6.5 - 4)^2 \cdot 1000 - \log(0.5118)\Bigg) \\ &\approx \underset{c=1,2,3}{\arg\min} \left(20626.4102, 20626.4102, 2625.6698\right) = 3.\end{aligned}$$

2. Aktualisierung In einem zweiten Schritt erfolgt eine Aktualisierung durch die Beobachtung $\boldsymbol{x}_{12} = 6.5$ mit $c_{12} = 1$. Inzwischen liegen bereits $M = 3$ Klassen vor.

Im Durchlauf des *M-AF* Algorithmus werden zunächst die Normierungskonstante für die Gewichte sowie die Schätzer für die a-priori Wahrscheinlichkeiten der Klassen durch (4.55) und (4.54) (bzw. (4.68)) aktualisiert:

$$
\begin{aligned}
N_{12}^{(0)} &= \lambda_{11}^{(0)} N_{11}^{(0)} + 1 = 0.9537 \cdot 1.9537 + 1 \approx 2.8632, \\
\tilde{P}_{12}^{(1)} &= \left(1 - \frac{1}{N_{12}^{(0)}}\right) \tilde{P}_{11}^{(1)} + \frac{1}{N_{12}^{(0)}} \cdot \mathbb{1}_{\{c_{12}=1\}} \approx \left(1 - \frac{1}{2.8632}\right) \cdot 0.2441 + \frac{1}{2.8632} \cdot 1 \approx 0.5081, \\
\tilde{P}_{12}^{(2)} &= \left(1 - \frac{1}{N_{12}^{(0)}}\right) \tilde{P}_{11}^{(2)} + \frac{1}{N_{12}^{(0)}} \cdot \mathbb{1}_{\{c_{12}=2\}} \approx \left(1 - \frac{1}{2.8632}\right) \cdot 0.2441 + 0 \approx 0.1588, \\
\tilde{P}_{12}^{(3)} &= \left(1 - \frac{1}{N_{12}^{(0)}}\right) \tilde{P}_{11}^{(3)} + \frac{1}{N_{12}^{(0)}} \cdot \mathbb{1}_{\{c_{12}=3\}} \approx \left(1 - \frac{1}{2.8632}\right) \cdot 0.5118 + 0 \approx 0.3330.
\end{aligned}
$$

Die entsprechenden Gradienten können durch (4.60) und (4.59) (bzw. (4.69)) aktualisiert werden:

$$
\begin{aligned}
\left(N_{12}^{(0)}\right)' &= \lambda_{11}^{(0)} \left(N_{11}^{(0)}\right)' + N_{11}^{(0)} = 0.9537 \cdot 1 + 1.9537 = 2.9074, \\
\left(\tilde{P}_{12}^{(1)}\right)' &= \frac{N_{12}^{(0)} - 1}{N_{12}^{(0)}} \cdot \left(\tilde{P}_{11}^{(1)}\right)' - \frac{\left(N_{12}^{(0)}\right)'}{\left(N_{12}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{12}=1\}} - \tilde{P}_{11}^{(1)}\right) \\
&\approx \frac{2.8632 - 1}{2.8632} \cdot 0.1310 - \frac{2.9074}{2.8632^2} \cdot (1 - 0.2441) \approx -0.1828, \\
\left(\tilde{P}_{12}^{(2)}\right)' &= \frac{N_{12}^{(0)} - 1}{N_{12}^{(0)}} \cdot \left(\tilde{P}_{11}^{(2)}\right)' - \frac{\left(N_{12}^{(0)}\right)'}{\left(N_{12}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{12}=2\}} - \tilde{P}_{11}^{(2)}\right) \\
&\approx \frac{2.8632 - 1}{2.8632} \cdot 0.1310 - \frac{2.9074}{2.8632^2} \cdot (0 - 0.2441) \approx 0.1718, \\
\left(\tilde{P}_{12}^{(3)}\right)' &= \frac{N_{12}^{(0)} - 1}{N_{12}^{(0)}} \cdot \left(\tilde{P}_{11}^{(3)}\right)' - \frac{\left(N_{12}^{(0)}\right)'}{\left(N_{12}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{12}=3\}} - \tilde{P}_{11}^{(3)}\right) \\
&\approx \frac{2.8632 - 1}{2.8632} \cdot 0 - \frac{2.9074}{2.8632^2} \cdot (0 - 0.5118) \approx 0.1815.
\end{aligned}
$$

Der neue Gradient der NLL lautet mithilfe von (4.58)

$$
\begin{aligned}
\left(J_{12}^{(0)}\right)' &= -\sum_{c=1}^{M} \left(\mathbb{1}_{\{c_{12}=c\}} - \tilde{P}_{11}^{(c)}\right) \frac{\left(\tilde{P}_{11}^{(c)}\right)'}{\tilde{P}_{11}^{(c)}} = -\sum_{c=1}^{3} \left(\mathbb{1}_{\{1=c\}} - \tilde{P}_{11}^{(c)}\right) \frac{\left(\tilde{P}_{11}^{(c)}\right)'}{\tilde{P}_{11}^{(c)}} \\
&\approx -\left((1 - 0.2441) \cdot \frac{0.1310}{0.2441} + (0 - 0.2441) \cdot \frac{0.1310}{0.2441} + (0 - 0.5118) \cdot \frac{0}{0.5118}\right) \\
&\approx -(0.4057 + (-0.1310) + 0) = -0.2747.
\end{aligned}
$$

Mithilfe dieses Gradienten ergibt sich der neue Faktor durch den Gradientenabstieg (4.49):

$$\lambda_{12}^{(0)} = \left[\lambda_{11}^{(0)} - \alpha_{11}^{(0)} \left(J_{12}^{(0)}\right)'\right]_{\lambda_-}^{\lambda_+} = \left[0.9537 - 10^{-8} \cdot (-0.2747)\right]_{0.7}^{0.999} \approx 0.9537.$$

Für die Schrittweite wird der *RPROP*-Algorithmus (4.50) herangezogen:

$$\alpha_{12}^{(0)} = \left[0.99\, \alpha_{11}^{(0)}\right]_{\alpha_{\min}}^{\alpha_{\max}} = \left[0.99 \cdot 10^{-8}\right]_{10^{-8}}^{10^{-6}} = 10^{-8},$$
$$\text{da } \left|\left(J_{12}^{(0)}\right)'\right| \approx 0.2747 > 10^{-7} \text{ und } \left(J_{12}^{(0)}\right)' \left(J_{11}^{(0)}\right)' \approx -0.2747 \cdot 0 = 0.$$

Die $M = 3$ Durchläufe des *G-AF* Algorithmus zur Aktualisierung der klassenbedingten Verteilungen erfolgen erneut mit den Parametern

$$\text{fix}_{\boldsymbol{\Sigma}} = \texttt{TRUE},\ \boldsymbol{\Sigma}_0 = \tilde{\boldsymbol{\Sigma}}_{11}^{(P)} = 0,\ \boldsymbol{\Pi}_0 = \tilde{\boldsymbol{\Pi}}_{11}^{(P)} = 0,\ d_0 = d_{11}^{(P)} = -1000,\ \boldsymbol{G}_0 = \boldsymbol{G}_{11}^{(P)} = 1000.$$

Zunächst werden auch hier jeweils aktuelle Normierungskonstanten für die Gewichte benötigt, welche durch (4.41) bestimmt werden können:

$$N_{12}^{(1)} = \lambda_{(n_{11}^{(1)})}^{(1)} N_{11}^{(1)} + 1 = 0.9356 \cdot 0 + 1 = 1, \quad N_{12}^{(2)} = N_{11}^{(2)} = 0, \quad N_{12}^{(3)} = N_{11}^{(3)} = 1.$$

Die Mittelwertvektoren können durch (4.38) aktualisiert werden:

$$\tilde{\boldsymbol{m}}_{n_{12}^{(1)}}^{(1)} = \left(1 - \frac{1}{N_{12}^{(1)}}\right) \tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)} + \frac{1}{N_{12}^{(1)}} \cdot \boldsymbol{x}_{12} = \left(1 - \frac{1}{1}\right) \cdot 0 + \frac{1}{1} \cdot 6.5 = 6.5,$$
$$\tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)} = \tilde{\boldsymbol{m}}_{n_{11}^{(2)}}^{(2)} = 0, \quad \tilde{\boldsymbol{m}}_{n_{12}^{(3)}}^{(3)} = \tilde{\boldsymbol{m}}_{n_{11}^{(3)}}^{(3)} = 4.$$

Für die jeweiligen Gradienten gilt mit (4.73) bzw. (4.70):

$$\left(N_{12}^{(1)}\right)' = \lambda_{(n_{11}^{(1)})}^{(1)} \left(N_{11}^{(1)}\right)' + N_{11}^{(1)} = 0.9356 \cdot 0 + 0 = 0,$$
$$\left(N_{12}^{(2)}\right)' = \left(N_{11}^{(2)}\right)' = 0, \quad \left(N_{12}^{(3)}\right)' = \left(N_{11}^{(3)}\right)' = 0,$$
$$\left(\tilde{\boldsymbol{m}}_{n_{12}^{(1)}}^{(1)}\right)' = \left(1 - \frac{1}{N_{12}^{(1)}}\right) \left(\tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)' - \frac{\left(N_{12}^{(1)}\right)'}{\left(N_{12}^{(1)}\right)^2} \left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)$$
$$= \left(1 - \frac{1}{1}\right) \cdot 0 - \frac{0}{1^2} \cdot (6.5 - 0) = 0,$$
$$\left(\tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)}\right)' = \left(\tilde{\boldsymbol{m}}_{n_{11}^{(2)}}^{(2)}\right)' = 0, \quad \left(\tilde{\boldsymbol{m}}_{n_{12}^{(3)}}^{(3)}\right)' = \left(\tilde{\boldsymbol{m}}_{n_{11}^{(3)}}^{(3)}\right)' = 0.$$

Da $\text{fix}_{\boldsymbol{\Sigma}} = \texttt{TRUE}$, gilt für die Kovarianzmatrizen und die Gradienten hier ebenfalls (vgl. Algorithmen 1 und 4 auf Seiten 82/88; die Größen der Klassen 2 und 3 bleiben unverändert):

$$
\begin{aligned}
\tilde{\boldsymbol{\Pi}}_{12}^{(1)} &= \boldsymbol{\Pi}_0 = 0, & \tilde{\boldsymbol{\Pi}}_{12}^{(2)} &= \tilde{\boldsymbol{\Pi}}_{11}^{(2)} = 0, & \tilde{\boldsymbol{\Pi}}_{12}^{(3)} &= \tilde{\boldsymbol{\Pi}}_{11}^{(3)} = 0,\\
\tilde{\boldsymbol{\Sigma}}_{12}^{(1)} &= \boldsymbol{\Sigma}_0 = 0, & \tilde{\boldsymbol{\Sigma}}_{12}^{(2)} &= \tilde{\boldsymbol{\Sigma}}_{11}^{(2)} = 0, & \tilde{\boldsymbol{\Sigma}}_{12}^{(3)} &= \tilde{\boldsymbol{\Sigma}}_{11}^{(3)} = 0,\\
d_{12}^{(1)} &= d_0 = -1000, & d_{12}^{(2)} &= d_{11}^{(2)} = -1000, & d_{12}^{(3)} &= d_{11}^{(3)} = -1000,\\
\boldsymbol{G}_{12}^{(1)} &= \boldsymbol{G}_0 = 1000, & \boldsymbol{G}_{12}^{(2)} &= \boldsymbol{G}_{11}^{(2)} = 1000, & \boldsymbol{G}_{12}^{(3)} &= \boldsymbol{G}_{11}^{(3)} = 1000,\\
\left(\tilde{\boldsymbol{\Pi}}_{12}^{(1)}\right)' &= 0, & \left(\tilde{\boldsymbol{\Pi}}_{12}^{(2)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_{11}^{(2)}\right)' = 0, & \left(\tilde{\boldsymbol{\Pi}}_{12}^{(3)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_{11}^{(3)}\right)' = 0,\\
\left(\tilde{\boldsymbol{\Sigma}}_{12}^{(1)}\right)' &= 0, & \left(\tilde{\boldsymbol{\Sigma}}_{12}^{(2)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_{11}^{(2)}\right)' = 0, & \left(\tilde{\boldsymbol{\Sigma}}_{12}^{(3)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_{11}^{(3)}\right)' = 0,\\
\left(d_{12}^{(1)}\right)' &= 0, & \left(d_{12}^{(2)}\right)' &= \left(d_{11}^{(2)}\right)' = 0, & \left(d_{12}^{(3)}\right)' &= \left(d_{11}^{(3)}\right)' = 0,\\
\left(\boldsymbol{G}_{12}^{(1)}\right)' &= 0, & \left(\boldsymbol{G}_{12}^{(2)}\right)' &= \left(\boldsymbol{G}_{11}^{(2)}\right)' = 0, & \left(\boldsymbol{G}_{12}^{(3)}\right)' &= \left(\boldsymbol{G}_{11}^{(3)}\right)' = 0.
\end{aligned}
$$

Der neue Gradient der NLL der Klasse 1 berechnet sich durch (4.43) bzw. (4.61):

$$
\begin{aligned}
\left(J_{12}^{(1)}\right)' &= \frac{1}{2}\left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)^T \left(-2\,\boldsymbol{G}_{11}^{(1)}\left(\tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)' + \left(\boldsymbol{G}_{11}^{(1)}\right)'\left(\boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{11}^{(1)}}^{(1)}\right)\right) + \frac{1}{2}\left(d_{11}^{(1)}\right)'\\
&= \frac{1}{2}\cdot(6.5-0)\cdot(-2\cdot 1000\cdot 0 + 0\cdot(6.5-0)) + \frac{1}{2}\cdot 0 = 0,\\
\left(J_{12}^{(2)}\right)' &= \left(J_{11}^{(2)}\right)' = 0, \quad \left(J_{12}^{(3)}\right)' = \left(J_{11}^{(3)}\right)' = 0.
\end{aligned}
$$

Der Gradientenabstieg (4.49) dient zur Aktualisierung des Faktors:

$$
\begin{aligned}
\lambda_{(n_{12}^{(1)})}^{(1)} &= \left[\lambda_{(n_{11}^{(1)})}^{(1)} - \alpha_{11}^{(1)}\left(J_{12}^{(1)}\right)'\right]_{\lambda_-}^{\lambda_+} = \left[0.9356 - 8.6\cdot 10^{-7}\cdot 0\right]_{0.7}^{0.999} = 0.9356,\\
\lambda_{(n_{12}^{(2)})}^{(2)} &= \lambda_{(n_{11}^{(2)})}^{(2)} = 0.9356, \quad \lambda_{(n_{12}^{(3)})}^{(3)} = \lambda_{(n_{11}^{(3)})}^{(3)} = 0.8903.
\end{aligned}
$$

Zur Aktualisierung der Schrittweite wird auch hier der *RPROP*-Algorithmus (4.50) herangezogen:

$$
\begin{aligned}
\alpha_{12}^{(1)} &= \alpha_{11}^{(1)} = 8.6\cdot 10^{-7}, \quad \text{da } \left|\left(J_{12}^{(1)}\right)'\right| = 0 \le 10^{-7},\\
\alpha_{12}^{(2)} &= \alpha_{11}^{(2)} = 8.6\cdot 10^{-7}, \quad \alpha_{12}^{(3)} = \alpha_{11}^{(3)} = 9.3\cdot 10^{-7}.
\end{aligned}
$$

Zuletzt wird der neue Datenpunkt zentriert durch $\boldsymbol{\xi}_{12} = \boldsymbol{x}_{12} - \tilde{\boldsymbol{m}}_{n_{12}^{(1)}}^{(1)} = 6.5 - 6.5 = 0$.

Zur Aktualisierung der gepoolten Kovarianzmatrix wird ein weiterer Durchlauf des *G-AF* Algorithmus mit $\boldsymbol{\xi}_{12} = 0$ betrachtet, wobei hier $\text{fix}_{\boldsymbol{m}} = \texttt{TRUE}$ und $\boldsymbol{m}_0 = 0$ gesetzt wird.

Die aktuelle Normierungskonstante für die Gewichte wird auch hier durch (4.41) bestimmt, der entsprechende Gradient durch (4.73):

$$
\begin{aligned}
N_{12}^{(P)} &= \lambda_{11}^{(P)} N_{11}^{(P)} + 1 = 0.9326\cdot 1 + 1 = 1.9326,\\
\left(N_{12}^{(P)}\right)' &= \lambda_{11}^{(P)}\left(N_{11}^{(P)}\right)' + N_{11}^{(P)} = 0.9326\cdot 0 + 1 = 1.
\end{aligned}
$$

Da $\text{fix}_{\boldsymbol{m}} = \texttt{TRUE}$, gilt für den neuen Mittelwertvektor sowie den Gradienten:

$$\tilde{\boldsymbol{m}}_{n_{12}}^{(P)} = \boldsymbol{m}_0 = 0, \quad \left(\tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)' = 0.$$

Die neue Kovarianzmatrix berechnet sich durch (4.39) und (4.40):

$$\begin{aligned}
\tilde{\boldsymbol{\Pi}}_{12}^{(P)} &= \left(1 - \frac{1}{N_{12}^{(P)}}\right) \tilde{\boldsymbol{\Pi}}_{11}^{(P)} + \frac{1}{N_{12}^{(P)}} \cdot \boldsymbol{\xi}_{12}\, \boldsymbol{\xi}_{12}^T = \left(1 - \frac{1}{1.9326}\right) \cdot 0 + \frac{1}{1.9326} \cdot 0^2 = 0, \\
\tilde{\boldsymbol{\Sigma}}_{12}^{(P)} &= \tilde{\boldsymbol{\Pi}}_{12}^{(P)} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)} \left(\tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)^T = 0 - 0^2 = 0.
\end{aligned}$$

Zur Aktualisierung von $d_{11}^{(P)} := \log \left|\tilde{\boldsymbol{\Sigma}}_{11}^{(P)}\right|$ und $\boldsymbol{G}_{11}^{(P)} := \left(\tilde{\boldsymbol{\Sigma}}_{11}^{(P)}\right)^{-1}$ werden im Allgemeinen weitere Hilfsfunktionen herangezogen. Für die Formeln sei auf Anagnostopoulos et al. (2012, S. 145) verwiesen:

$$\begin{aligned}
\boldsymbol{H}_{12}^{(P)} &= \boldsymbol{G}_{11}^{(P)} \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right) \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)^T \boldsymbol{G}_{11}^{(P)} = 1000 \cdot (0-0)^2 \cdot 1000 = 0, \\
\gamma_{12}^{(P)} &= \frac{\left(N_{12}^{(P)} - 1\right)^2}{N_{12}^{(P)}} + \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)^T \boldsymbol{G}_{11}^{(P)} \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right) \\
&= \frac{(1.9326 - 1)^2}{1.9326} + (0-0) \cdot 1000 \cdot (0-0) \approx 0.4500.
\end{aligned}$$

Damit:

$$\begin{aligned}
\boldsymbol{G}_{12}^{(P)} &= \left(\frac{N_{12}^{(P)}}{N_{12}^{(P)} - 1}\right) \left(\boldsymbol{G}_{11}^{(P)} - \frac{\boldsymbol{H}_{12}^{(P)}}{\gamma_{12}^{(P)}}\right) \approx \frac{1.9326}{1.9326 - 1} \cdot \left(1000 - \frac{0}{0.4500}\right) \approx 2072.2711, \\
d_{12}^{(P)} &= (p-2) \log\left(N_{12}^{(P)} - 1\right) + (1-p) \log N_{12}^{(P)} + \log \gamma_{12}^{(P)} + d_{11}^{(P)} \\
&\approx (1-2) \cdot \log(1.9326 - 1) + (1-1) \cdot \log(1.9326) + \log(0.4500) - 1000 \\
&\approx -1000.7287.
\end{aligned}$$

Die entsprechenden aktuellen Gradienten berechnen sich durch

$$\begin{aligned}
\left(\boldsymbol{H}_{12}^{(P)}\right)' = {} & \left(\boldsymbol{G}_{11}^{(P)}\right)' \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right) \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)^T \boldsymbol{G}_{11}^{(P)} \\
& - \boldsymbol{G}_{11}^{(P)} \left(\tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)' \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)^T \boldsymbol{G}_{11}^{(P)} \\
& - \boldsymbol{G}_{11}^{(P)} \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right) \left(\left(\tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)'\right)^T \boldsymbol{G}_{11}^{(P)} \\
& + \boldsymbol{G}_{11}^{(P)} \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right) \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)^T \left(\boldsymbol{G}_{11}^{(P)}\right)' \\
= {} & 0 \cdot (0-0)^2 \cdot 1000 - 1000 \cdot 0 \cdot (0-0) \cdot 1000 - 1000 \cdot (0-0) \cdot 0 \cdot 1000 \\
& + 1000 \cdot (0-0)^2 \cdot 0 = 0,
\end{aligned}$$

$$\begin{aligned}
\left(\gamma_{12}^{(P)}\right)' &= \left(N_{12}^{(P)}\right)' \cdot \frac{\left(N_{12}^{(P)}\right)^2 - 1}{\left(N_{12}^{(P)}\right)^2} + \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)^T \left(\boldsymbol{G}_{11}^{(P)}\right)' \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right) \\
&\quad - 2\left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)^T \boldsymbol{G}_{11}^{(P)} \left(\tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)' \\
&= 1 \cdot \frac{1.9326^2 - 1}{1.9326^2} + (0-0) \cdot 0 \cdot (0-0) - 2 \cdot (0-0) \cdot 1000 \cdot 0 \approx 0.7323, \\
\left(\boldsymbol{G}_{12}^{(P)}\right)' &= -\frac{\left(N_{12}^{(P)}\right)'}{\left(N_{12}^{(P)} - 1\right)^2} \left(\boldsymbol{G}_{11}^{(P)} - \frac{1}{\gamma_{12}^{(P)}} \cdot \boldsymbol{H}_{12}^{(P)}\right) \\
&\quad + \frac{N_{12}^{(P)}}{N_{12}^{(P)} - 1} \left(\left(\boldsymbol{G}_{11}^{(P)}\right)' + \frac{\left(\gamma_{12}^{(P)}\right)'}{\left(\gamma_{12}^{(P)}\right)^2} \cdot \boldsymbol{H}_{12}^{(P)} - \frac{1}{\gamma_{12}^{(P)}} \cdot \left(\boldsymbol{H}_{12}^{(P)}\right)'\right) \\
&\approx -\frac{1}{(1.9326-1)^2} \left(1000 - \frac{1}{0.4500} \cdot 0\right) \\
&\quad + \frac{1.9326}{1.9326-1} \left(0 + \frac{0.7323}{0.4500^2} \cdot 0 - \frac{1}{0.4500} \cdot 0\right) \approx -1149.7652, \\
\left(d_{12}^{(P)}\right)' &= \frac{(p-2)\left(N_{12}^{(P)}\right)'}{N_{12}^{(P)} - 1} + \frac{(1-p)\left(N_{12}^{(P)}\right)'}{N_{12}^{(P)}} + \frac{\left(\gamma_{12}^{(P)}\right)'}{\gamma_{12}^{(P)}} + \left(d_{11}^{(P)}\right)' \\
&\approx \frac{(1-2)\cdot 1}{1.9326-1} + \frac{(1-1)\cdot 1}{1.9326} + \frac{0.7323}{0.4500} + 0 \approx 0.5550.
\end{aligned}$$

Der Gradient der NLL ergibt sich durch (4.43) bzw. (4.61):

$$\begin{aligned}
\left(J_{12}^{(P)}\right)' &= \frac{1}{2}\left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{11}}^{(P)}\right)^T \left(-2\,\boldsymbol{G}_{11}^{(P)} \left(\tilde{\boldsymbol{m}}_{n_{11}}^{(P)}\right)' + \left(\boldsymbol{G}_{11}^{(P)}\right)' \left(\boldsymbol{\xi}_{12} - \tilde{\boldsymbol{m}}_{n_{11}}^{(P)}\right)\right) + \frac{1}{2}\left(d_{11}^{(P)}\right)' \\
&= \frac{1}{2} \cdot (0-0) \cdot (-2 \cdot 1000 \cdot 0 + 0 \cdot (0-0)) + \frac{1}{2} \cdot 0 = 0.
\end{aligned}$$

Für den Faktor gilt mit (4.49):

$$\lambda_{12}^{(P)} = \left[\lambda_{11}^{(P)} - \alpha_{11}^{(P)} \left(J_{12}^{(P)}\right)'\right]_{\lambda_-}^{\lambda_+} = \left[0.9326 - 9.8 \cdot 10^{-7} \cdot 0\right]_{0.7}^{0.999} = 0.9326.$$

Für die Schrittweite gilt hier nach (4.50):

$$\alpha_{12}^{(P)} = \alpha_{11}^{(P)} = 9.8 \cdot 10^{-7}, \quad \text{da } \left|\left(J_{12}^{(P)}\right)'\right| = 0 < 10^{-7}.$$

2. Prognose Mithilfe der rekursiv bestimmten Schätzer erfolgt die Vorhersage einer neuen Beobachtung mithilfe der Klassifikationsregel (4.64). Sei die nächste Beobachtung im Datenstrom $\boldsymbol{x}_{13} = 1$ mit $c_{13} = 2$ betrachtet. Die prognostizierte Klasse beträgt

$$\tilde{c}_{13} = \underset{c=1,2,3}{\arg\min} \left(\frac{1}{2}\, d_{12}^{(P)} + \frac{1}{2}\left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{12}^{(c)}}^{(c)}\right)^T \boldsymbol{G}_{12}^{(P)} \left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{12}^{(c)}}^{(c)}\right) - \log \tilde{P}_{12}^{(c)}\right)$$

$$
\begin{aligned}
&\approx \underset{c=1,2,3}{\arg\min} \Bigg(\frac{1}{2} \cdot (-1000.7287) + \frac{1}{2} \cdot (1-6.5)^2 \cdot 2072.2711 - \log(0.5081), \\
&\qquad\qquad \frac{1}{2} \cdot (-1000.7287) + \frac{1}{2} \cdot (1-0)^2 \cdot 2072.2711 - \log(0.1588), \\
&\qquad\qquad \frac{1}{2} \cdot (-1000.7287) + \frac{1}{2} \cdot (1-4)^2 \cdot 2072.2711 - \log(0.3330) \Bigg) \\
&\approx \underset{c=1,2,3}{\arg\min} \left(30843.4131, 537.6113, 8825.9552\right) = 2.
\end{aligned}
$$

3. Aktualisierung Die neuen Größen durch Aktualisierung durch die Beobachtung $\boldsymbol{x}_{13} = 1$ mit $c_{13} = 2$ berechnen sich folgendermaßen:

- *M-AF:*

$$
\begin{aligned}
N_{13}^{(0)} &= \lambda_{12}^{(0)} N_{12}^{(0)} + 1 \approx 0.9537 \cdot 2.8632 + 1 \approx 3.7306, \\
\tilde{P}_{13}^{(1)} &= \left(1 - \frac{1}{N_{13}^{(0)}}\right) \tilde{P}_{12}^{(1)} + \frac{1}{N_{13}^{(0)}} \cdot \mathbb{1}_{\{c_{13}=1\}} \\
&\approx \left(1 - \frac{1}{3.7306}\right) \cdot 0.5081 + 0 \approx 0.3719, \\
\tilde{P}_{13}^{(2)} &= \left(1 - \frac{1}{N_{13}^{(0)}}\right) \tilde{P}_{12}^{(2)} + \frac{1}{N_{13}^{(0)}} \cdot \mathbb{1}_{\{c_{13}=2\}} \\
&\approx \left(1 - \frac{1}{3.7306}\right) \cdot 0.1588 + \frac{1}{3.7306} \cdot 1 \approx 0.3843, \\
\tilde{P}_{13}^{(3)} &= \left(1 - \frac{1}{N_{13}^{(0)}}\right) \tilde{P}_{12}^{(3)} + \frac{1}{N_{13}^{(0)}} \cdot \mathbb{1}_{\{c_{13}=3\}} \\
&\approx \left(1 - \frac{1}{3.7306}\right) \cdot 0.3330 + 0 \approx 0.2438, \\
\left(N_{13}^{(0)}\right)' &= \lambda_{12}^{(0)} \left(N_{12}^{(0)}\right)' + N_{12}^{(0)} \approx 0.9537 \cdot 2.9074 + 2.8632 \approx 5.6360, \\
\left(\tilde{P}_{13}^{(1)}\right)' &= \frac{N_{13}^{(0)} - 1}{N_{13}^{(0)}} \cdot \left(\tilde{P}_{12}^{(1)}\right)' - \frac{\left(N_{13}^{(0)}\right)'}{\left(N_{13}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{13}=1\}} - \tilde{P}_{12}^{(1)}\right) \\
&\approx \frac{3.7306 - 1}{3.7306} \cdot (-0.1828) - \frac{5.6360}{3.7306^2} \cdot (0 - 0.5081) \approx 0.0720, \\
\left(\tilde{P}_{13}^{(2)}\right)' &= \frac{N_{13}^{(0)} - 1}{N_{13}^{(0)}} \cdot \left(\tilde{P}_{12}^{(2)}\right)' - \frac{\left(N_{13}^{(0)}\right)'}{\left(N_{13}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{13}=2\}} - \tilde{P}_{12}^{(2)}\right) \\
&\approx \frac{3.7306 - 1}{3.7306} \cdot 0.1718 - \frac{5.6360}{3.7306^2} \cdot (1 - 0.1588) \approx -0.2149, \\
\left(\tilde{P}_{13}^{(3)}\right)' &= \frac{N_{13}^{(0)} - 1}{N_{13}^{(0)}} \cdot \left(\tilde{P}_{12}^{(3)}\right)' - \frac{\left(N_{13}^{(0)}\right)'}{\left(N_{13}^{(0)}\right)^2} \left(\mathbb{1}_{\{c_{13}=3\}} - \tilde{P}_{12}^{(3)}\right) \\
&\approx \frac{3.7306 - 1}{3.7306} \cdot 0.1815 - \frac{5.6360}{3.7306^2} \cdot (0 - 0.3330) \approx 0.2677.
\end{aligned}
$$

Der neue Gradient der NLL lautet mithilfe von (4.58)

$$\begin{aligned}\left(J_{13}^{(0)}\right)' &= -\sum_{c=1}^{M}\left(\mathbb{1}_{\{c_{13}=c\}} - \tilde{P}_{12}^{(c)}\right)\frac{\left(\tilde{P}_{12}^{(c)}\right)'}{\tilde{P}_{12}^{(c)}} = -\sum_{c=1}^{3}\left(\mathbb{1}_{\{2=c\}} - \tilde{P}_{12}^{(c)}\right)\frac{\left(\tilde{P}_{12}^{(c)}\right)'}{\tilde{P}_{12}^{(c)}} \\ &\approx -\Bigg((0-0.5081)\cdot\frac{-0.1828}{0.5081} + (1-0.1588)\cdot\frac{0.1718}{0.1588} \\ &\qquad + (0-0.3330)\cdot\frac{0.1815}{0.3330}\Bigg) \\ &\approx -(0.1828 + 0.9101 + (-0.1815)) = -0.9114.\end{aligned}$$

Mithilfe dieses Gradienten ergibt sich der neue Faktor durch den Gradientenabstieg (4.49):

$$\lambda_{13}^{(0)} = \left[\lambda_{12}^{(0)} - \alpha_{12}^{(0)}\left(J_{13}^{(0)}\right)'\right]_{\lambda_-}^{\lambda_+} \approx \left[0.9537 - 10^{-8}\cdot(-0.9114)\right]_{0.7}^{0.999} \approx 0.9537.$$

Für die Schrittweite wird der *RPROP*-Algorithmus (4.50) herangezogen:

$$\begin{aligned}&\alpha_{13}^{(0)} = \left[1.01\,\alpha_{12}^{(0)}\right]_{\alpha_{\min}}^{\alpha_{\max}} = \left[1.01\cdot 10^{-8}\right]_{10^{-8}}^{10^{-6}} = 1.01\cdot 10^{-8}, \\ &\text{da } \left|\left(J_{13}^{(0)}\right)'\right| \approx 0.9114 > 10^{-7} \\ &\text{und } \left(J_{13}^{(0)}\right)'\left(J_{12}^{(0)}\right)' \approx -0.9114\cdot(-0.2747) \approx 0.2504 > 0.\end{aligned}$$

- *G-AF für klassenbedingte Verteilungen:*

  Es wird

  $$\begin{aligned}&\text{fix}_{\boldsymbol{\Sigma}} = \texttt{TRUE},\ \boldsymbol{\Sigma}_0 = \tilde{\boldsymbol{\Sigma}}_{12}^{(P)} = 0,\ \boldsymbol{\Pi}_0 = \tilde{\boldsymbol{\Pi}}_{12}^{(P)} = 0,\ d_0 = d_{12}^{(P)} \approx -1000.7287, \\ &\boldsymbol{G}_0 = \boldsymbol{G}_{12}^{(P)} \approx 2072.2711\end{aligned}$$

  gesetzt.

  Beim *G-AF* Algorithmus werden auch zunächst aktuelle Normierungskonstanten für die Gewichte benötigt, welche durch (4.41) bestimmt werden:

  $$N_{13}^{(1)} = N_{12}^{(1)} = 1, \quad N_{13}^{(2)} = \lambda_{(n_{12}^{(2)})}^{(2)} N_{12}^{(2)} + 1 = 0.9356\cdot 0 + 1 = 1, \quad N_{13}^{(3)} = N_{12}^{(3)} = 1.$$

  Die Mittelwertvektoren können durch (4.38) aktualisiert werden:

  $$\begin{aligned}&\tilde{\boldsymbol{m}}_{n_{13}^{(1)}}^{(1)} = \tilde{\boldsymbol{m}}_{n_{12}^{(1)}}^{(1)} = 6.5, \\ &\tilde{\boldsymbol{m}}_{n_{13}^{(2)}}^{(2)} = \left(1 - \frac{1}{N_{13}^{(2)}}\right)\tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)} + \frac{1}{N_{13}^{(2)}}\cdot\boldsymbol{x}_{13} = \left(1-\frac{1}{1}\right)\cdot 0 + \frac{1}{1}\cdot 1 = 1, \\ &\tilde{\boldsymbol{m}}_{n_{13}^{(3)}}^{(3)} = \tilde{\boldsymbol{m}}_{n_{12}^{(3)}}^{(3)} = 4.\end{aligned}$$

Für die jeweiligen Gradienten gilt mit (4.73) bzw. (4.70):

$$
\begin{aligned}
\left(N_{13}^{(1)}\right)' &= \left(N_{12}^{(1)}\right)' = 0,\\
\left(N_{13}^{(2)}\right)' &= \lambda_{(n_{12}^{(2)})}^{(2)} \left(N_{12}^{(2)}\right)' + N_{12}^{(2)} = 0.9356 \cdot 0 + 0 = 0,\\
\left(N_{13}^{(3)}\right)' &= \left(N_{12}^{(3)}\right)' = 0,\\
\left(\tilde{\boldsymbol{m}}_{n_{13}^{(1)}}^{(1)}\right)' &= \left(\tilde{\boldsymbol{m}}_{n_{12}^{(1)}}^{(1)}\right)' = 0,\\
\left(\tilde{\boldsymbol{m}}_{n_{13}^{(2)}}^{(2)}\right)' &= \left(1 - \frac{1}{N_{13}^{(2)}}\right) \left(\tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)}\right)' - \frac{\left(N_{13}^{(2)}\right)'}{\left(N_{13}^{(2)}\right)^2} \left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)}\right)\\
&= \left(1 - \frac{1}{1}\right) \cdot 0 - \frac{0}{1^2} \cdot (1 - 0) = 0,\\
\left(\tilde{\boldsymbol{m}}_{n_{13}^{(3)}}^{(3)}\right)' &= \left(\tilde{\boldsymbol{m}}_{n_{12}^{(3)}}^{(3)}\right)' = 0.
\end{aligned}
$$

Da $\text{fix}_{\boldsymbol{\Sigma}}$ = `TRUE`, gilt für die Kovarianzmatrizen und die Gradienten hier ebenfalls (vgl. Algorithmen 1 und 4 auf Seiten 82/88; die Größen der Klassen 1 und 3 bleiben unverändert):

$$
\begin{aligned}
\tilde{\boldsymbol{\Pi}}_{13}^{(1)} &= \tilde{\boldsymbol{\Pi}}_{12}^{(1)} = 0, & \tilde{\boldsymbol{\Pi}}_{13}^{(2)} &= \boldsymbol{\Pi}_0 = 0, & \tilde{\boldsymbol{\Pi}}_{13}^{(3)} &= \tilde{\boldsymbol{\Pi}}_{12}^{(3)} = 0,\\
\tilde{\boldsymbol{\Sigma}}_{13}^{(1)} &= \tilde{\boldsymbol{\Sigma}}_{12}^{(1)} = 0, & \tilde{\boldsymbol{\Sigma}}_{13}^{(2)} &= \boldsymbol{\Sigma}_0 = 0, & \tilde{\boldsymbol{\Sigma}}_{13}^{(3)} &= \tilde{\boldsymbol{\Sigma}}_{12}^{(3)} = 0,\\
d_{13}^{(1)} &= d_{12}^{(1)} = -1000, & d_{13}^{(2)} &= d_0 \approx -1000.7287, & d_{13}^{(3)} &= d_{12}^{(3)} = -1000,\\
\boldsymbol{G}_{13}^{(1)} &= \boldsymbol{G}_{12}^{(1)} = 1000, & \boldsymbol{G}_{13}^{(2)} &= \boldsymbol{G}_0 \approx 2072.2711, & \boldsymbol{G}_{13}^{(3)} &= \boldsymbol{G}_{12}^{(3)} = 1000,\\
\left(\tilde{\boldsymbol{\Pi}}_{13}^{(1)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_{12}^{(1)}\right)' = 0, & \left(\tilde{\boldsymbol{\Pi}}_{13}^{(2)}\right)' &= 0, & \left(\tilde{\boldsymbol{\Pi}}_{13}^{(3)}\right)' &= \left(\tilde{\boldsymbol{\Pi}}_{12}^{(3)}\right)' = 0,\\
\left(\tilde{\boldsymbol{\Sigma}}_{13}^{(1)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_{12}^{(1)}\right)' = 0, & \left(\tilde{\boldsymbol{\Sigma}}_{13}^{(2)}\right)' &= 0, & \left(\tilde{\boldsymbol{\Sigma}}_{13}^{(3)}\right)' &= \left(\tilde{\boldsymbol{\Sigma}}_{12}^{(3)}\right)' = 0,\\
\left(d_{13}^{(1)}\right)' &= \left(d_{12}^{(1)}\right)' = 0, & \left(d_{13}^{(2)}\right)' &= 0, & \left(d_{13}^{(3)}\right)' &= \left(d_{12}^{(3)}\right)' = 0,\\
\left(\boldsymbol{G}_{13}^{(1)}\right)' &= \left(\boldsymbol{G}_{12}^{(1)}\right)' = 0, & \left(\boldsymbol{G}_{13}^{(2)}\right)' &= 0, & \left(\boldsymbol{G}_{13}^{(3)}\right)' &= \left(\boldsymbol{G}_{12}^{(3)}\right)' = 0.
\end{aligned}
$$

Der aktuelle Gradient der NLL der Klasse 2 berechnet sich durch (4.43) bzw. (4.61):

$$
\begin{aligned}
\left(J_{13}^{(1)}\right)' &= \left(J_{12}^{(1)}\right)' = 0,\\
\left(J_{13}^{(2)}\right)' &= \frac{1}{2} \left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)}\right)^T \left(-2\, \boldsymbol{G}_{12}^{(2)} \left(\tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)}\right)' + \left(\boldsymbol{G}_{12}^{(2)}\right)' \left(\boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{12}^{(2)}}^{(2)}\right)\right)\\
&\quad + \frac{1}{2} \left(d_{12}^{(2)}\right)'\\
&= \frac{1}{2} \cdot (1 - 0) \cdot (-2 \cdot 1000 \cdot 0 + 0 \cdot (1 - 0)) + \frac{1}{2} \cdot 0 = 0,
\end{aligned}
$$

$$\left(J_{13}^{(3)}\right)' = \left(J_{12}^{(3)}\right)' = 0.$$

Der Gradientenabstieg (4.49) dient zur Aktualisierung des Faktors:

$$\begin{aligned}
\lambda_{(n_{13}^{(1)})}^{(1)} &= \lambda_{(n_{12}^{(1)})}^{(1)} = 0.9356,\\
\lambda_{(n_{13}^{(2)})}^{(2)} &= \left[\lambda_{(n_{12}^{(2)})}^{(2)} - \alpha_{12}^{(2)}\left(J_{13}^{(2)}\right)'\right]_{\lambda_-}^{\lambda_+} = \left[0.9356 - 8.6\cdot 10^{-7}\cdot 0\right]_{0.7}^{0.999} = 0.9356,\\
\lambda_{(n_{13}^{(3)})}^{(3)} &= \lambda_{(n_{12}^{(3)})}^{(3)} = 0.8903.
\end{aligned}$$

Zur Aktualisierung der Schrittweite wird auch hier der *RPROP*-Algorithmus (4.50) herangezogen:

$$\begin{aligned}
\alpha_{13}^{(1)} &= \alpha_{12}^{(1)} = 8.6\cdot 10^{-7},\\
\alpha_{13}^{(2)} &= \alpha_{12}^{(2)} = 8.6\cdot 10^{-7}, \quad \text{da } \left|\left(J_{13}^{(2)}\right)'\right| = 0 \leq 10^{-7},\\
\alpha_{13}^{(3)} &= \alpha_{12}^{(3)} = 9.3\cdot 10^{-7}.
\end{aligned}$$

Zuletzt wird der neue Datenpunkt zentriert durch $\boldsymbol{\xi}_{13} = \boldsymbol{x}_{13} - \tilde{\boldsymbol{m}}_{n_{13}^{(2)}}^{(2)} = 1 - 1 = 0$.

- *G-AF für gepoolte Kovarianzmatrix:*

  Zur Aktualisierung der gepoolten Kovarianzmatrix wird ein weiterer Durchlauf des *G-AF* Algorithmus mit $\boldsymbol{\xi}_{13} = 0$ betrachtet, wobei hier $\text{fix}_{\boldsymbol{m}} = \texttt{TRUE}$ und $\boldsymbol{m}_0 = 0$ gesetzt wird.

  Die aktuelle Normierungskonstante für die Gewichte wird auch hier durch (4.41) bestimmt, der entsprechende Gradient durch (4.73):

$$\begin{aligned}
N_{13}^{(P)} &= \lambda_{12}^{(P)} N_{12}^{(P)} + 1 = 0.9326\cdot 1.9326 + 1 \approx 2.8023,\\
\left(N_{13}^{(P)}\right)' &= \lambda_{12}^{(P)}\left(N_{12}^{(P)}\right)' + N_{12}^{(P)} = 0.9326\cdot 1 + 1.9326 = 2.8652.
\end{aligned}$$

  Da $\text{fix}_{\boldsymbol{m}} = \texttt{TRUE}$, gilt für den neuen Mittelwertvektor sowie den Gradienten:

$$\tilde{\boldsymbol{m}}_{n_{13}}^{(P)} = \boldsymbol{m}_0 = 0, \quad \left(\tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)' = 0.$$

  Die neue Kovarianzmatrix berechnet sich durch (4.39) und (4.40):

$$\begin{aligned}
\tilde{\boldsymbol{\Pi}}_{13}^{(P)} &= \left(1 - \frac{1}{N_{13}^{(P)}}\right)\tilde{\boldsymbol{\Pi}}_{12}^{(P)} + \frac{1}{N_{13}^{(P)}}\cdot\boldsymbol{\xi}_{13}\,\boldsymbol{\xi}_{13}^T \approx \left(1 - \frac{1}{2.8023}\right)\cdot 0 + \frac{1}{2.8023}\cdot 0^2 = 0,\\
\tilde{\boldsymbol{\Sigma}}_{13}^{(P)} &= \tilde{\boldsymbol{\Pi}}_{13}^{(P)} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\left(\tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)^T = 0 - 0^2 = 0.
\end{aligned}$$

Auch hier werden $d_{12}^{(P)} := \log\left|\tilde{\boldsymbol{\Sigma}}_{12}^{(P)}\right|$ und $\boldsymbol{G}_{12}^{(P)} := \left(\tilde{\boldsymbol{\Sigma}}_{12}^{(P)}\right)^{-1}$ mithilfe von Hilfsfunktionen aktualisiert:

$$\begin{aligned}
\boldsymbol{H}_{13}^{(P)} &= \boldsymbol{G}_{12}^{(P)}\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)^T \boldsymbol{G}_{12}^{(P)} \\
&\approx 2072.2711 \cdot (0-0) \cdot (0-0) \cdot 2072.2711 = 0, \\
\gamma_{13}^{(P)} &= \frac{\left(N_{13}^{(P)} - 1\right)^2}{N_{13}^{(P)}} + \left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)^T \boldsymbol{G}_{12}^{(P)}\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right) \\
&\approx \frac{(2.8023-1)^2}{2.8023} + (0-0) \cdot 2072.2711 \cdot (0-0) \approx 1.1592.
\end{aligned}$$

Damit:

$$\begin{aligned}
\boldsymbol{G}_{13}^{(P)} &= \left(\frac{N_{13}^{(P)}}{N_{13}^{(P)} - 1}\right)\left(\boldsymbol{G}_{12}^{(P)} - \frac{\boldsymbol{H}_{13}^{(P)}}{\gamma_{13}^{(P)}}\right) \approx \frac{2.8023}{2.8023-1} \cdot \left(2072.2711 - \frac{0}{1.1592}\right) \\
&\approx 3222.0636, \\
d_{13}^{(P)} &= (p-2)\log\left(N_{13}^{(P)} - 1\right) + (1-p)\log N_{13}^{(P)} + \log\gamma_{13}^{(P)} + d_{12}^{(P)} \\
&\approx (1-2) \cdot \log(2.8023-1) + (1-1) \cdot \log(2.8023) + \log(1.1592) - 1000.7287 \\
&\approx -1001.1700.
\end{aligned}$$

Die entsprechenden aktuellen Gradienten berechnen sich durch

$$\begin{aligned}
\left(\boldsymbol{H}_{13}^{(P)}\right)' &= \left(\boldsymbol{G}_{12}^{(P)}\right)'\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)^T \boldsymbol{G}_{12}^{(P)} \\
&\quad - \boldsymbol{G}_{12}^{(P)}\left(\tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)'\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)^T \boldsymbol{G}_{12}^{(P)} \\
&\quad - \boldsymbol{G}_{12}^{(P)}\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)\left(\left(\tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)'\right)^T \boldsymbol{G}_{12}^{(P)} \\
&\quad + \boldsymbol{G}_{12}^{(P)}\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)^T \left(\boldsymbol{G}_{12}^{(P)}\right)' \\
&\approx (-1149.7652) \cdot (0-0)^2 \cdot 2072.2711 - 2072.2711 \cdot 0 \cdot (0-0) \cdot 2072.2711 \\
&\quad - 2072.2711 \cdot (0-0) \cdot 0 \cdot 2072.2711 \\
&\quad + 2072.2711 \cdot (0-0)^2 \cdot (-1149.7652) = 0, \\
\left(\gamma_{13}^{(P)}\right)' &= \left(N_{13}^{(P)}\right)' \cdot \frac{\left(N_{13}^{(P)}\right)^2 - 1}{\left(N_{13}^{(P)}\right)^2} + \left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)^T \left(\boldsymbol{G}_{12}^{(P)}\right)'\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right) \\
&\quad - 2\left(\boldsymbol{\xi}_{13} - \tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)^T \boldsymbol{G}_{12}^{(P)}\left(\tilde{\boldsymbol{m}}_{n_{13}}^{(P)}\right)' \\
&\approx 2.8652 \cdot \frac{2.8023^2 - 1}{2.8023^2} + (0-0) \cdot (-1149.7652) \cdot (0-0) \\
&\quad - 2 \cdot (0-0) \cdot 2072.2711 \cdot 0 \approx 2.5003,
\end{aligned}$$

$$\begin{aligned}
\left(\boldsymbol{G}_{13}^{(P)}\right)' &= -\frac{\left(N_{13}^{(P)}\right)'}{\left(N_{13}^{(P)}-1\right)^2}\left(\boldsymbol{G}_{12}^{(P)}-\frac{1}{\gamma_{13}^{(3)}}\cdot\boldsymbol{H}_{13}^{(P)}\right)\\
&\quad+\frac{N_{13}^{(P)}}{N_{13}^{(P)}-1}\left(\left(\boldsymbol{G}_{12}^{(P)}\right)'+\frac{\left(\gamma_{13}^{(P)}\right)'}{\left(\gamma_{13}^{(P)}\right)^2}\cdot\boldsymbol{H}_{13}^{(P)}-\frac{1}{\gamma_{13}^{(P)}}\cdot\left(\boldsymbol{H}_{13}^{(P)}\right)'\right)\\
&\approx -\frac{2.8652}{(2.8023-1)^2}\left(2072.2711-\frac{1}{1.1592}\cdot 0\right)\\
&\quad+\frac{2.8023}{2.8023-1}\left(-1149.7652+\frac{2.5003}{1.1592^2}\cdot 0-\frac{1}{1.1592}\cdot 0\right)\approx -3615.5871,\\
\left(d_{13}^{(P)}\right)' &= \frac{(p-2)\left(N_{13}^{(P)}\right)'}{N_{13}^{(P)}-1}+\frac{(1-p)\left(N_{13}^{(P)}\right)'}{N_{13}^{(P)}}+\frac{\left(\gamma_{13}^{(P)}\right)'}{\gamma_{13}^{(P)}}+\left(d_{12}^{(P)}\right)'\\
&\approx \frac{(1-2)\cdot 2.8652}{2.8023-1}+\frac{(1-1)\cdot 2.8652}{2.8023}+\frac{2.5003}{1.1592}+0.5550\approx 1.1222.
\end{aligned}$$

Der Gradient der NLL ergibt sich durch (4.43) bzw. (4.61):

$$\begin{aligned}
\left(J_{13}^{(P)}\right)' &= \frac{1}{2}\left(\boldsymbol{\xi}_{13}-\tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)^T\left(-2\,\boldsymbol{G}_{12}^{(P)}\left(\tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)'+\left(\boldsymbol{G}_{12}^{(P)}\right)'\left(\boldsymbol{\xi}_{13}-\tilde{\boldsymbol{m}}_{n_{12}}^{(P)}\right)\right)\\
&\quad+\frac{1}{2}\left(d_{12}^{(P)}\right)'\\
&\approx \frac{1}{2}\cdot(0-0)\cdot(-2\cdot 2072.2711\cdot 0-1149.7652\cdot(0-0))+\frac{1}{2}\cdot 0.5550\\
&\approx 0.2775.
\end{aligned}$$

Für den Faktor gilt mit (4.49):

$$\lambda_{13}^{(P)}=\left[\lambda_{12}^{(P)}-\alpha_{12}^{(P)}\left(J_{13}^{(P)}\right)'\right]_{\lambda_-}^{\lambda_+}\approx\left[0.9326-9.8\cdot 10^{-7}\cdot 0.2775\right]_{0.7}^{0.999}\approx 0.9326.$$

Für die Schrittweite gilt hier nach (4.50):

$$\begin{aligned}
&\alpha_{13}^{(P)}=\left[0.99\,\alpha_{12}^{(P)}\right]_{\alpha_{\min}}^{\alpha_{\max}}=\left[0.99\cdot 9.8\cdot 10^{-7}\right]_{10^{-8}}^{10^{-6}}=9.702\cdot 10^{-7},\\
&\text{da }\left|\left(J_{13}^{(P)}\right)'\right|\approx 0.2775>10^{-7}\text{ und }\left(J_{13}^{(P)}\right)'\left(J_{12}^{(P)}\right)'\approx 0.2775\cdot 0=0.
\end{aligned}$$

B Formeln und Herleitungen

Die folgenden Abschnitte umfassen ausführliche Herleitungen aufgeführter Formeln der verschiedenen Kapitel.

B.1 Herleitungen für OLDC (Abschnitt 4.3)

Herleitung der aktualisierten gepoolten Kovarianzmatrix (4.23) bzw. (4.24) durch die neue Beobachtung $\boldsymbol{x}_{t+1}$:

$$
\begin{aligned}
\boldsymbol{S}_{t+1} &= \frac{1}{n_t+1}\left(\sum_{c=1}^{M}\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c\\ i\le t+1}}\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}\right)^T\right)\\
&\overset{g(\boldsymbol{x}_{t+1})=k}{=} \frac{1}{n_t+1}\left(\sum_{\substack{c=1\\ c\neq k}}^{M}\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c\\ i\le t}}\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}\right)^T\right.\\
&\qquad\left.+\sum_{\substack{i:\, g(\boldsymbol{x}_i)=k\\ i\le t+1}}\left(\boldsymbol{x}_i-\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}}\right)^T\right)\\
&= \frac{1}{n_t+1}\left(\sum_{\substack{c=1\\ c\neq k}}^{M}\sum_{i=1}^{t}\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t+1}}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right.\\
&\qquad\left.+\sum_{i=1}^{t+1}\left(\boldsymbol{x}_i-\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}}\right)\\
&= \frac{1}{n_t+1}\left(\sum_{\substack{c=1\\ c\neq k}}^{M}\sum_{i=1}^{t}\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t}}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t}}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right.\\
&\qquad+\sum_{i=1}^{t}\left(\boldsymbol{x}_i-\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}}\\
&\qquad\left.+\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}}\right)\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}^{(k)}_{n^{(k)}_{t+1}}\right)^T\right)\\
&\overset{(4.8)}{=} \frac{1}{n_t+1}\left(\sum_{\substack{c=1\\ c\neq k}}^{M}\sum_{i=1}^{t}\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t}}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}^{(c)}_{n^{(c)}_{t}}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}\right.\\
&\qquad\left.+\sum_{i=1}^{t}\left(\boldsymbol{x}_i-\boldsymbol{m}^{(k)}_{n^{(k)}_{t}}-\frac{\boldsymbol{x}_{t+1}-\boldsymbol{m}^{(k)}_{n^{(k)}_{t}}}{n^{(k)}_{t}+1}\right)\right.
\end{aligned}
$$

$$
\begin{aligned}
&\cdot \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}} - \frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}} \\
&+ \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}} - \frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right) \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}} - \frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right)^T \Bigg) \\
= \frac{1}{n_t + 1} \Bigg(& \sum_{\substack{c=1 \\ c \neq k}}^{M} \sum_{i=1}^{t} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \\
&+ \sum_{i=1}^{t} \Bigg(\Bigg(\Bigg(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right)^T \\
&\quad - \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right) \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right)^T \\
&\quad - \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right)^T \\
&\quad + \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right) \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right)^T \Bigg) \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}} \Bigg) \\
&+ \frac{\left(n_t^{(k)} \right)^2}{\left(n_t^{(k)} + 1 \right)^2} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right) \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right)^T \Bigg) \\
= \frac{1}{n_t + 1} & \sum_{\substack{c=1 \\ c \neq k}}^{M} \sum_{i=1}^{t} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n_t^{(c)}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \\
&+ \frac{1}{n_t + 1} \sum_{i=1}^{t} \Bigg(\Bigg(\Bigg(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right)^T \\
&\quad - \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right) \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right)^T \\
&\quad - \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}} \right)^T \\
&\quad + \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right) \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)} + 1} \right)^T \Bigg) \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}} \Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{\left(n_t^{(k)}\right)^2}{(n_t+1)\left(n_t^{(k)}+1\right)^2}\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T \\
&= \frac{1}{n_t+1}\sum_{\substack{c=1\\c\neq k}}^{M}\sum_{i=1}^{t}\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(c)}}^{(c)}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \\
&\quad + \frac{1}{n_t+1}\sum_{i=1}^{t}\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}} \\
&\quad + \frac{1}{n_t+1}\sum_{i=1}^{t}\left(\left(\frac{\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T}{\left(n_t^{(k)}+1\right)^2}\right.\right. \\
&\qquad\qquad - \left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\frac{\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}}{n_t^{(k)}+1}\right)^T \\
&\qquad\qquad \left.\left. - \left(\frac{\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}}{n_t^{(k)}+1}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T\right)\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}}\right) \\
&\quad + \frac{\left(n_t^{(k)}\right)^2}{(n_t+1)\left(n_t^{(k)}+1\right)^2}\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T \\
&= \frac{1}{n_t+1}\sum_{c=1}^{M}\sum_{i=1}^{t}\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(c)}}^{(c)}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(c)}}^{(c)}\right)^T\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \\
&\quad + \frac{1}{n_t+1}\sum_{i=1}^{t}\left(\left(\frac{\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T}{\left(n_t^{(k)}+1\right)^2}\right.\right. \\
&\qquad\qquad - \left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\frac{\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}}{n_t^{(k)}+1}\right)^T \\
&\qquad\qquad \left.\left. - \left(\frac{\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}}{n_t^{(k)}+1}\right)\left(\boldsymbol{x}_i-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T\right)\cdot\mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}}\right) \\
&\quad + \frac{\left(n_t^{(k)}\right)^2}{(n_t+1)\left(n_t^{(k)}+1\right)^2}\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)\left(\boldsymbol{x}_{t+1}-\boldsymbol{m}_{n_t^{(k)}}^{(k)}\right)^T
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{n_t}{n_t+1} \cdot \boldsymbol{S}_t + \frac{1}{n_t+1} \Bigg(\underbrace{\sum_{i=1}^{t} \frac{\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T}{\left(n_t^{(k)}+1\right)^2} \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}}}_{n_t^{(k)} \text{ Summanden}} \\
&\qquad - \sum_{i=1}^{t} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right) \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)}+1}\right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}} \\
&\qquad - \sum_{i=1}^{t} \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)}+1}\right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}} \Bigg) \\
&\quad + \frac{\left(n_t^{(k)}\right)^2}{(n_t+1)\left(n_t^{(k)}+1\right)^2} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T \\
&= \frac{n_t}{n_t+1} \cdot \boldsymbol{S}_t + \frac{1}{n_t+1} \Bigg(n_t^{(k)} \cdot \frac{\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T}{\left(n_t^{(k)}+1\right)^2} \\
&\qquad \underbrace{- \sum_{i=1}^{t} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right) \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}}}_{=\mathbf{0}} \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)}+1}\right)^T \\
&\qquad - \left(\frac{\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}}{n_t^{(k)}+1}\right) \underbrace{\sum_{i=1}^{t} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=k\}}}_{=\mathbf{0}^T} \Bigg) \\
&\quad + \frac{\left(n_t^{(k)}\right)^2}{(n_t+1)\left(n_t^{(k)}+1\right)^2} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T \\
&= \frac{n_t}{n_t+1} \cdot \boldsymbol{S}_t + \frac{n_t^{(k)}}{(n_t+1)\left(n_t^{(k)}+1\right)^2} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T \\
&\quad + \frac{\left(n_t^{(k)}\right)^2}{(n_t+1)\left(n_t^{(k)}+1\right)^2} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T \\
&= \frac{n_t}{n_t+1} \cdot \boldsymbol{S}_t + \left(\frac{n_t^{(k)} + \left(n_t^{(k)}\right)^2}{(n_t+1)\left(n_t^{(k)}+1\right)^2}\right) \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T \\
&= \frac{n_t}{n_t+1} \cdot \boldsymbol{S}_t + \left(\frac{n_t^{(k)}}{(n_t+1)\left(n_t^{(k)}+1\right)}\right) \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T \\
&= \frac{n_t}{n_t+1} \left(\boldsymbol{S}_t + \frac{n_t^{(k)}}{n_t\left(n_t^{(k)}+1\right)} \left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)\left(\boldsymbol{x}_{t+1} - \boldsymbol{m}^{(k)}_{n_t^{(k)}}\right)^T \right). \qquad \text{(B.1)}
\end{aligned}
$$

B.2 Herleitungen für Online Diskriminanzanalyse mit exponentiellem Vergessen (Abschnitt 4.4)

Zur Bestimmung des ML-Schätzers (4.34) für den Erwartungswertvektor $\boldsymbol{\mu}^{(c)}$ wird die gewichtete negative log-Likelihood (4.33) nach $\boldsymbol{\mu}^{(c)}$ abgeleitet:

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_{(1)}^{(c)}, \ldots, \boldsymbol{x}_{(n_t^{(c)})}^{(c)}) \\
&\overset{(4.33)}{=} \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} \left(\left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_i) \right) + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_t) \\
&\overset{(4.32)}{=} \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\frac{1}{2} \log \left| \boldsymbol{\Sigma}^{(c)} \right| \left(\left(\prod_{j=1}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) + \ldots + \left(\prod_{j=n_t^{(c)}-2}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) + \lambda_{(n_t^{(c)}-1)}^{(c)} + 1 \right) \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} \left(\left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) \cdot \frac{1}{2} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right) \right) \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\mu}^{(c)}} \left(\frac{1}{2} \left(\boldsymbol{x}_t - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_t - \boldsymbol{\mu}^{(c)} \right) \right) + 0 \\
&= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} \left(\left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) \cdot \frac{1}{2} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_i - \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_i + 2 \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \frac{1}{2} \left(- \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_t - \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{x}_t + 2 \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \boldsymbol{\mu}^{(c)} \right) \\
&= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} \left(\left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_i \right) \right) + \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_t \right).
\end{aligned}
\tag{B.2}
$$

Mit $v_i^{(c)} := \prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)}$ und $v_t^{(c)} := 1$ und durch Gleichsetzen von (B.2) mit $\mathbf{0}$ sowie Auflösen nach $\boldsymbol{\mu}^{(c)}$ ergibt sich der ML-Schätzer für den Erwartungswertvektor:

$$
\begin{aligned}
&\mathbf{0} \overset{!}{=} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} v_i^{(c)} \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_i \right) + v_t^{(c)} \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_t \right) \\
&\Leftrightarrow \mathbf{0} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} v_i^{(c)} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_i \right) + v_t^{(c)} \left(\boldsymbol{\mu}^{(c)} - \boldsymbol{x}_t \right) \\
&\Leftrightarrow \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} v_i^{(c)} \boldsymbol{\mu}^{(c)} + v_t^{(c)} \boldsymbol{\mu}^{(c)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} v_i^{(c)} \boldsymbol{x}_i + v_t^{(c)} \boldsymbol{x}_t \\
&\Leftrightarrow \boldsymbol{\mu}^{(c)} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} v_i^{(c)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} v_i^{(c)} \boldsymbol{x}_i
\end{aligned}
$$

$$\Leftrightarrow \boldsymbol{\mu}^{(c)} = \frac{\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t}} v_i^{(c)} \boldsymbol{x}_i}{\sum\limits_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t}} v_i^{(c)}} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t}} \frac{v_i^{(c)}}{N_t^{(c)}} \cdot \boldsymbol{x}_i$$

$$\Rightarrow \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t}} \frac{v_i^{(c)}}{N_t^{(c)}} \cdot \boldsymbol{x}_i. \tag{B.3}$$

Herleitung der rekursiven Variante des gewichteten Mittelwertschätzers (4.38):

$$\begin{aligned}
\tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}} &= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t}} \frac{v_i^{(c)}}{N_t^{(c)}} \cdot \boldsymbol{x}_i \\
&= \frac{1}{N_t^{(c)}} \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t-1}} \left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) \boldsymbol{x}_i + \boldsymbol{x}_t \right) \\
&= \frac{\lambda^{(c)}_{(n_t^{(c)}-1)}}{N_t^{(c)}} \Bigg(\underbrace{\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq \max(k \,|\, i\leq k\leq t-1 \,\wedge\, g(\boldsymbol{x}_k)=c)-1}} \left(\prod_{j=\sum_{l=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_l)=c\}}}^{n_t^{(c)}-2} \lambda_{(j)}^{(c)} \right) \boldsymbol{x}_i}_{=N_{t-1}^{(c)} \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}}} + \boldsymbol{x}^{(c)}_{(n_t^{(c)}-1)} \Bigg) \\
&\quad + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \\
&= \frac{1}{N_t^{(c)}} \left(\lambda^{(c)}_{(n_t^{(c)}-1)} N_{t-1}^{(c)} \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}} + \boldsymbol{x}_t \right) \\
&\overset{(4.37)}{=} \frac{1}{N_t^{(c)}} \left(\lambda^{(c)}_{(n_t^{(c)}-1)} \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t-1}} v_i^{(c)} \right) \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}} + \boldsymbol{x}_t \right) \\
&= \frac{1}{N_t^{(c)}} \left(\lambda^{(c)}_{(n_t^{(c)}-1)} \left(\lambda_{(1)}^{(c)} \cdots \lambda^{(c)}_{(n_t^{(c)}-2)} + \ldots + \lambda^{(c)}_{(n_t^{(c)}-2)} + 1 \right) \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}} + \boldsymbol{x}_t \right) \\
&= \frac{1}{N_t^{(c)}} \left(\lambda_{(1)} \cdots \lambda^{(c)}_{(n_t^{(c)}-1)} + \ldots + \lambda^{(c)}_{(n_t^{(c)}-2)} \lambda^{(c)}_{(n_t^{(c)}-1)} + \lambda^{(c)}_{(n_t^{(c)}-1)} + 1 - 1 \right) \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}} \\
&\quad + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \\
&= \frac{1}{N_t^{(c)}} \left(\left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\leq t}} v_i^{(c)} \right) - 1 \right) \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}} + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \\
&\overset{(4.37)}{=} \frac{1}{N_t^{(c)}} \left(N_t^{(c)} - 1 \right) \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}} + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \\
&= \left(1 - \frac{1}{N_t^{(c)}} \right) \tilde{\boldsymbol{m}}^{(c)}_{n_{t-1}^{(c)}} + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t.
\end{aligned} \tag{B.4}$$

Zur Bestimmung des ML-Schätzers (4.36) für die Kovarianzmatrix $\boldsymbol{\Sigma}^{(c)}$ wird die gewichtete negative log-Likelihood (4.33) nach $\boldsymbol{\Sigma}^{(c)}$ abgeleitet:

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \mathcal{L}^{(\boldsymbol{\lambda})}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}^{(c)}_{(1)}, \ldots, \boldsymbol{x}^{(c)}_{(n_t^{(c)})}) \\
&\overset{(4.33)}{=} \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} \left(\left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda^{(c)}_{(j)} \right) \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_i) \right) + \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \mathcal{L}(\boldsymbol{\mu}^{(c)}, \boldsymbol{\Sigma}^{(c)}; \boldsymbol{x}_t) \\
&\overset{(4.32)}{=} \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \left(\frac{1}{2} \log \left| \boldsymbol{\Sigma}^{(c)} \right| \left(\left(\prod_{j=1}^{n_t^{(c)}-1} \lambda^{(c)}_{(j)} \right) + \ldots + \left(\prod_{j=n_t^{(c)}-2}^{n_t^{(c)}-1} \lambda^{(c)}_{(j)} \right) + \lambda^{(c)}_{(n_t^{(c)}-1)} + 1 \right) \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} \left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda^{(c)}_{(j)} \right) \cdot \frac{1}{2} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right) \right) \\
&\quad + \frac{\partial}{\partial \boldsymbol{\Sigma}^{(c)}} \left(\frac{1}{2} \left(\boldsymbol{x}_t - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_t - \boldsymbol{\mu}^{(c)} \right) \right) + 0 \\
&\overset{\substack{(2),(3) \\ \text{S. } 125}}{=} \frac{1}{2} \left(\left(\prod_{j=1}^{n_t^{(c)}-1} \lambda^{(c)}_{(j)} \right) + \ldots + \left(\prod_{j=n_t^{(c)}-2}^{n_t^{(c)}-1} \lambda^{(c)}_{(j)} \right) + \lambda^{(c)}_{(n_t^{(c)}-1)} + 1 \right) \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \\
&\quad - \frac{1}{2} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t-1}} \left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda^{(c)}_{(j)} \right) \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \\
&\quad - \frac{1}{2} \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_t - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_t - \boldsymbol{\mu}^{(c)} \right)^T \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1}. \qquad \text{(B.5)}
\end{aligned}
$$

Mit $v_i^{(c)} := \prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda^{(c)}_{(j)}$ und $v_t^{(c)} := 1$ und durch Gleichsetzen von (B.5) mit der Nullmatrix, Multiplikation beider Seiten mit dem Faktor 2 sowie der Matrix $\boldsymbol{\Sigma}^{(c)}$ von rechts sowie Auflösen nach $\boldsymbol{\Sigma}^{(c)}$ ergibt sich der ML-Schätzer für die Kovarianzmatrix:

$$
\begin{aligned}
\mathbf{0} &= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} v_i^{(c)} - \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} \left(v_i^{(c)} \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right)^T \right) \\
&\Leftrightarrow \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} v_i^{(c)} = \left(\boldsymbol{\Sigma}^{(c)} \right)^{-1} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} v_i^{(c)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right)^T \\
&\Leftrightarrow \boldsymbol{\Sigma}^{(c)} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} v_i^{(c)} = \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \leq t}} v_i^{(c)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right) \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)} \right)^T
\end{aligned}
$$

$$
\begin{aligned}
\Leftrightarrow \boldsymbol{\Sigma}^{(c)} &= \frac{\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} v_i^{(c)} \left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)}\right)\left(\boldsymbol{x}_i - \boldsymbol{\mu}^{(c)}\right)^T}{\sum\limits_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} v_i^{(c)}} \\
&= \frac{\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} v_i^{(c)} \left(\boldsymbol{x}_i\, \boldsymbol{x}_i^T - 2\boldsymbol{x}_i \left(\boldsymbol{\mu}^{(c)}\right)^T + \boldsymbol{\mu}^{(c)} \left(\boldsymbol{\mu}^{(c)}\right)^T\right)}{\sum\limits_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} v_i^{(c)}} \\
&= \frac{\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} v_i^{(c)} \left(\boldsymbol{x}_i\, \boldsymbol{x}_i^T - 2\boldsymbol{x}_i \left(\boldsymbol{\mu}^{(c)}\right)^T\right)}{\sum\limits_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} v_i^{(c)}} + \boldsymbol{\mu}^{(c)} \left(\boldsymbol{\mu}^{(c)}\right)^T \\
\Rightarrow \tilde{\boldsymbol{\Sigma}}_t^{(c)} &= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} \frac{v_i^{(c)} \left(\boldsymbol{x}_i\, \boldsymbol{x}_i^T - 2\boldsymbol{x}_i \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T\right)}{\sum\limits_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j\le t}} v_j^{(c)}} + \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T \\
&= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} \frac{v_i^{(c)} \boldsymbol{x}_i\, \boldsymbol{x}_i^T}{\sum\limits_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j\le t}} v_j^{(c)}} - 2 \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} \frac{v_i^{(c)}}{\sum\limits_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j\le t}} v_j^{(c)}} \cdot \boldsymbol{x}_i \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T + \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T \\
&\overset{(4.34)}{=} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} \frac{v_i^{(c)}}{\sum\limits_{\substack{j:\, g(\boldsymbol{x}_j)=c \\ j\le t}} v_j^{(c)}} \cdot \boldsymbol{x}_i\, \boldsymbol{x}_i^T - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T \\
&\overset{(4.35)}{=} \underbrace{\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} \frac{v_i^{(c)}}{N_t^{(c)}} \cdot \boldsymbol{x}_i\, \boldsymbol{x}_i^T}_{=:\tilde{\boldsymbol{\Pi}}_t^{(c)}} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}\right)^T .
\end{aligned}
\tag{B.6}
$$

Herleitung der rekursiven Variante (4.40):

$$
\begin{aligned}
\tilde{\boldsymbol{\Pi}}_t^{(c)} &= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t}} \frac{v_i^{(c)}}{N_t^{(c)}} \cdot \boldsymbol{x}_i\, \boldsymbol{x}_i^T \\
&= \frac{1}{N_t^{(c)}} \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le t-1}} \left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda_{(j)}^{(c)} \right) \boldsymbol{x}_i\, \boldsymbol{x}_i^T + \boldsymbol{x}_t\, \boldsymbol{x}_t^T \right) \\
&= \frac{\lambda_{(n_t^{(c)}-1)}^{(c)}}{N_t^{(c)}} \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i\le \max(k \,|\, i\le k\le t-1 \,\wedge\, g(\boldsymbol{x}_k)=c)-1}} \left(\prod_{j=\sum_{l=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_l)=c\}}}^{n_t^{(c)}-2} \lambda_{(j)}^{(c)} \right) \boldsymbol{x}_i\, \boldsymbol{x}_i^T \right. \\
&\qquad \left. + \boldsymbol{x}_{(n_t^{(c)}-1)}^{(c)} \left(\boldsymbol{x}_{(n_t^{(c)}-1)}^{(c)}\right)^T \right) + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t\, \boldsymbol{x}_t^T
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{N_t^{(c)}} \left(\lambda^{(c)}_{(n_t^{(c)}-1)} N_{t-1}^{(c)} \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)} + \boldsymbol{x}_t \, \boldsymbol{x}_t^T \right) \\
&\overset{(4.37)}{=} \frac{1}{N_t^{(c)}} \left(\lambda^{(c)}_{(n_t^{(c)}-1)} \left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t-1}} v_i^{(c)} \right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)} + \boldsymbol{x}_t \boldsymbol{x}_t^T \right) \\
&\overset{(4.33)}{=} \frac{1}{N_t^{(c)}} \left(\lambda^{(c)}_{(n_t^{(c)}-1)} \left(\lambda^{(c)}_{(1)} \cdots \lambda^{(c)}_{(n_t^{(c)}-2)} + \ldots + \lambda^{(c)}_{(n_t^{(c)}-2)} + 1 \right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)} + \boldsymbol{x}_t \, \boldsymbol{x}_t^T \right) \\
&= \frac{1}{N_t^{(c)}} \left(\lambda_{(1)} \cdots \lambda^{(c)}_{(n_t^{(c)}-1)} + \ldots + \lambda^{(c)}_{(n_t^{(c)}-2)} \lambda^{(c)}_{(n_t^{(c)}-1)} + \lambda^{(c)}_{(n_t^{(c)}-1)} + 1 - 1 \right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)} \\
&\quad + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \, \boldsymbol{x}_t^T \\
&\overset{(4.33)}{=} \frac{1}{N_t^{(c)}} \left(\left(\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t}} v_i^{(c)} \right) - 1 \right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)} + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \, \boldsymbol{x}_t^T \\
&\overset{(4.37)}{=} \frac{1}{N_t^{(c)}} \left(N_t^{(c)} - 1 \right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)} + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \, \boldsymbol{x}_t^T \\
&= \left(1 - \frac{1}{N_t^{(c)}} \right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)} + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \, \boldsymbol{x}_t^T, \qquad \text{(B.7)}
\end{aligned}
$$

$$
\begin{aligned}
\tilde{\boldsymbol{\Sigma}}_t^{(c)} &= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t}} \frac{v_i^{(c)}}{N_t^{(c)}} \cdot \boldsymbol{x}_i \, \boldsymbol{x}_i^T - \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}} \left(\tilde{\boldsymbol{m}}^{(c)}_{N_t^{(c)}} \right)^T \\
&= \frac{1}{N_t^{(c)}} \left(N_t^{(c)} - 1 \right) \tilde{\boldsymbol{\Pi}}_{t-1}^{(c)} + \frac{1}{N_t^{(c)}} \cdot \boldsymbol{x}_t \, \boldsymbol{x}_t^T - \tilde{\boldsymbol{m}}^{(c)}_{n_t^{(c)}} \left(\tilde{\boldsymbol{m}}^{(c)}_{N_t^{(c)}} \right)^T . \qquad \text{(B.8)}
\end{aligned}
$$

Herleitung der rekursiven Variante (4.41) der Normierungskonstante (4.37) ((4.55) aus (4.56) analog):

$$
\begin{aligned}
N_t^{(c)} &= \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t}} v_i^{(c)} \overset{(4.33)}{=} \sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le t-1}} \left(\prod_{j=\sum_{k=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_k)=c\}}}^{n_t^{(c)}-1} \lambda^{(c)}_{(j)} \right) + 1 \\
&= \lambda^{(c)}_{(n_t^{(c)}-1)} \Bigg(\underbrace{\sum_{\substack{i:\, g(\boldsymbol{x}_i)=c \\ i \le \max(k \,|\, i \le k \le t-1 \,\wedge\, g(\boldsymbol{x}_k)=c)-1}} \left(\prod_{j=\sum_{l=1}^{i} \mathbb{1}_{\{g(\boldsymbol{x}_l)=c\}}}^{n_t^{(c)}-2} \lambda^{(c)}_{(j)} \right) + 1}_{=N_{t-1}^{(c)}} \Bigg) + 1 \\
&= \lambda^{(c)}_{(n_t^{(c)}-1)} N_{t-1}^{(c)} + 1 \\
&= \lambda^{(c)}_{(n_{t-1}^{(c)})} N_{t-1}^{(c)} + 1. \qquad \text{(B.9)}
\end{aligned}
$$

Herleitung des Gradienten (4.43) der negativen log-Likelihood (4.42):

$$
\begin{aligned}
\left(J_{t+1}^{(c)}\right)' &= \frac{\partial \mathcal{L}(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)}, \tilde{\boldsymbol{\Sigma}}_t^{(c)}; \boldsymbol{x}_{t+1})}{\partial \lambda^{(c)}} \\
&= \frac{\partial}{\partial \lambda^{(c)}} \left(\frac{1}{2} \log \left| \tilde{\boldsymbol{\Sigma}}_t^{(c)} \right| + \frac{1}{2} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)^T \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)} \right)^{-1} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right) + const. \right) \\
&= \frac{\partial}{\partial \lambda^{(c)}} \left(\frac{1}{2} \log \left| \tilde{\boldsymbol{\Sigma}}_t^{(c)} \right| \right) + \frac{\partial}{\partial \lambda^{(c)}} \left(\frac{1}{2} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)^T \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)} \right)^{-1} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right) \right) \\
&= \frac{1}{2} \left(\log \left| \tilde{\boldsymbol{\Sigma}}_t^{(c)} \right| \right)' - \frac{1}{2} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)^T \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)} \right)^{-1} \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)' \\
&\quad + \frac{1}{2} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)^T \left(\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)} \right)^{-1} \right)' \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right) \\
&\quad - \frac{1}{2} \left(\left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)^T \right)' \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)} \right)^{-1} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right) \\
&= \frac{1}{2} \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)^T \left(-2 \left(\tilde{\boldsymbol{\Sigma}}_t^{(c)} \right)^{-1} \left(\tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right)' + \left(\left(\tilde{\boldsymbol{\Sigma}}_t^{(c)} \right)^{-1} \right)' \left(\boldsymbol{x}_{t+1} - \tilde{\boldsymbol{m}}_{n_t^{(c)}}^{(c)} \right) \right) \\
&\quad + \frac{1}{2} \left(\log \left| \tilde{\boldsymbol{\Sigma}}_t^{(c)} \right| \right)'. \qquad \text{(B.10)}
\end{aligned}
$$

Zur Bestimmung des ML-Schätzers für den Parameter $p^{(c)}$ wird die gewichtete negative log-Likelihood (4.53) nach $p^{(c)}$ abgeleitet:

$$
\begin{aligned}
&\frac{\partial}{\partial p^{(c)}} \mathcal{L}^{(\boldsymbol{\lambda})}(p^{(1)}, \ldots, p^{(M)}; c_1, \ldots, c_t) \\
&\overset{(4.53)}{=} \frac{\partial}{\partial p^{(c)}} \sum_{i=1}^{t-1} \left(\left(\prod_{j=i}^{t-1} \lambda_j^{(0)} \right) \mathcal{L}(p^{(1)}, \ldots, p^{(M)}; c_i) \right) + \frac{\partial}{\partial p^{(c)}} \mathcal{L}(p^{(1)}, \ldots, p^{(M)}; c_t) \\
&\overset{(4.52)}{=} \frac{\partial}{\partial p^{(c)}} \left(-\lambda_1^{(0)} \cdots \lambda_{t-1}^{(0)} \sum_{c'=1}^{M} \left(\mathbb{1}_{\{c_1 = c'\}} \cdot \log \left(\frac{p^{(c')}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \right) + \ldots + \\
&\quad + \frac{\partial}{\partial p^{(c)}} \left(-\lambda_{t-1}^{(0)} \sum_{c'=1}^{M} \left(\mathbb{1}_{\{c_{t-1} = c'\}} \cdot \log \left(\frac{p^{(c')}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \right) \\
&\quad + \frac{\partial}{\partial p^{(c)}} \left(-\sum_{c'=1}^{M} \left(\mathbb{1}_{\{c_t = c'\}} \cdot \log \left(\frac{p^{(c')}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \right) \\
&= \frac{\partial}{\partial p^{(c)}} \left(-v_1^{(0)} \sum_{c'=1}^{M} \left(\mathbb{1}_{\{c_1 = c'\}} \cdot \log \left(\frac{p^{(c')}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \right) + \ldots + \\
&\quad + \frac{\partial}{\partial p^{(c)}} \left(-v_{t-1}^{(0)} \sum_{c'=1}^{M} \left(\mathbb{1}_{\{c_{t-1} = c'\}} \cdot \log \left(\frac{p^{(c')}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \right) \\
&\quad + \frac{\partial}{\partial p^{(c)}} \left(-\sum_{c'=1}^{M} \left(\mathbb{1}_{\{c_t = c'\}} \cdot \log \left(\frac{p^{(c')}}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \right)
\end{aligned}
$$

$$
\begin{aligned}
&= -v_1^{(0)} \left(\mathbb{1}_{\{c_1=c\}} \left(\frac{1}{p^{(c)}} - \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) - \sum_{\substack{c'=1,\\ c'\neq c}}^{M} \left(\mathbb{1}_{\{c_1=c'\}} \cdot \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \\
&\quad - \ldots - v_{t-1}^{(0)} \left(\mathbb{1}_{\{c_{t-1}=c\}} \left(\frac{1}{p^{(c)}} - \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) - \sum_{\substack{c'=1,\\ c'\neq c}}^{M} \left(\mathbb{1}_{\{c_{t-1}=c'\}} \cdot \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \\
&\quad - \left(\mathbb{1}_{\{c_t=c\}} \left(\frac{1}{p^{(c)}} - \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) - \sum_{\substack{c'=1,\\ c'\neq c}}^{M} \left(\mathbb{1}_{\{c_t=c'\}} \cdot \frac{1}{\sum_{k=1}^{M} p^{(k)}} \right) \right) \\
&= -v_1^{(0)} \left(\mathbb{1}_{\{c_1=c\}} \cdot \frac{1}{p^{(c)}} - \sum_{c'=1}^{M} \mathbb{1}_{\{c_1=c'\}} \right) - \ldots - v_{t-1}^{(0)} \left(\mathbb{1}_{\{c_{t-1}=c\}} \cdot \frac{1}{p^{(c)}} - \sum_{c'=1}^{M} \mathbb{1}_{\{c_{t-1}=c'\}} \right) \\
&\quad - \left(\mathbb{1}_{\{c_t=c\}} \cdot \frac{1}{p^{(c)}} - \sum_{c'=1}^{M} \mathbb{1}_{\{c_t=c'\}} \right) \\
&= -v_1^{(0)} \left(\mathbb{1}_{\{c_1=c\}} \cdot \frac{1}{p^{(c)}} - 1 \right) - \ldots - v_{t-1}^{(0)} \left(\mathbb{1}_{\{c_{t-1}=c\}} \cdot \frac{1}{p^{(c)}} - 1 \right) - \left(\mathbb{1}_{\{c_t=c\}} \cdot \frac{1}{p^{(c)}} - 1 \right).
\end{aligned}
\tag{B.11}
$$

Mit $v_t^{(0)} := 1$ ergibt Gleichsetzen von (B.11) mit 0 und Auflösen nach $p^{(c)}$ den ML-Schätzer:

$$
\begin{aligned}
0 &\stackrel{!}{=} -v_1^{(0)} \left(\mathbb{1}_{\{c_1=c\}} \cdot \frac{1}{p^{(c)}} - 1 \right) - \ldots - v_{t-1}^{(0)} \left(\mathbb{1}_{\{c_{t-1}=c\}} \cdot \frac{1}{p^{(c)}} - 1 \right) - \left(\mathbb{1}_{\{c_t=c\}} \cdot \frac{1}{p^{(c)}} - 1 \right) \\
&\Leftrightarrow \sum_{j=1}^{t} v_j^{(0)} = v_1^{(0)} \cdot \mathbb{1}_{\{c_1=c\}} \cdot \frac{1}{p^{(c)}} + \ldots + v_{t-1}^{(0)} \cdot \mathbb{1}_{\{c_{t-1}=c\}} \cdot \frac{1}{p^{(c)}} + \mathbb{1}_{\{c_t=c\}} \cdot \frac{1}{p^{(c)}} \\
&\Leftrightarrow p^{(c)} = \frac{v_1^{(0)} \cdot \mathbb{1}_{\{c_1=c\}} + \ldots + v_{t-1}^{(0)} \cdot \mathbb{1}_{\{c_{t-1}=c\}} + \mathbb{1}_{\{c_t=c\}}}{\sum_{j=1}^{t} v_j^{(0)}} = \sum_{i=1}^{t} \frac{v_i^{(0)}}{\sum_{j=1}^{t} v_j^{(0)}} \cdot \mathbb{1}_{\{c_i=c\}} \\
&\Rightarrow \tilde{P}_t^{(c)} = \sum_{i=1}^{t} \frac{v_i^{(0)}}{\sum_{j=1}^{t} v_j^{(0)}} \cdot \mathbb{1}_{\{c_i=c\}} \stackrel{(4.56)}{=} \sum_{i=1}^{t} \frac{v_i^{(0)}}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_i=c\}}, \quad c = 1, \ldots, M.
\end{aligned}
\tag{B.12}
$$

Herleitung der rekursiven Variante (4.54):

$$
\begin{aligned}
\tilde{P}_t^{(c)} &= \sum_{i=1}^{t} \frac{v_i^{(0)}}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_i=c\}} \\
&\stackrel{(4.53)}{=} \frac{1}{N_t^{(0)}} \left(\sum_{i=1}^{t-1} \left(\prod_{j=i}^{t-1} \lambda_j^{(0)} \right) \cdot \mathbb{1}_{\{c_i=c\}} + \mathbb{1}_{\{c_t=c\}} \right) \\
&= \frac{1}{N_t^{(0)}} \left(\lambda_1^{(0)} \cdots \lambda_{t-1}^{(0)} \cdot \mathbb{1}_{\{c_1=c\}} + \ldots + \lambda_{t-1}^{(0)} \cdot \mathbb{1}_{\{c_{t-1}=c\}} + \mathbb{1}_{\{c_t=c\}} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{N_t^{(0)}} \cdot \lambda_{t-1}^{(0)} \left(\lambda_1^{(0)} \cdots \lambda_{t-2}^{(0)} \cdot \mathbb{1}_{\{c_1=c\}} + \ldots + \lambda_{t-2}^{(0)} \cdot \mathbb{1}_{\{c_{t-2}=c\}} + \mathbb{1}_{\{c_{t-1}=c\}} \right) \\
&\quad + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t=c\}} \\
&= \frac{1}{N_t^{(0)}} \cdot \lambda_{t-1}^{(0)} \underbrace{\left(\sum_{i=1}^{t-2} \left(\left(\prod_{j=i}^{t-2} \lambda_j^{(0)} \right) \cdot \mathbb{1}_{\{c_i=c\}} \right) + \mathbb{1}_{\{c_{t-1}=c\}} \right)}_{=N_{t-1}^{(0)} \tilde{P}_{t-1}^{(c)}} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t=c\}} \\
&= \frac{1}{N_t^{(0)}} \cdot \lambda_{t-1}^{(0)} N_{t-1}^{(0)} \tilde{P}_{t-1}^{(c)} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t=c\}} \\
&\overset{(4.56)}{=} \frac{1}{N_t^{(0)}} \cdot \lambda_{t-1}^{(0)} \left(v_1^{(0)} + \ldots + v_{t-1}^{(0)} \right) \tilde{P}_{t-1}^{(c)} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t=c\}} \\
&\overset{(4.53)}{=} \frac{1}{N_t^{(0)}} \cdot \lambda_{t-1}^{(0)} \left(\lambda_1^{(0)} \cdots \lambda_{t-2}^{(0)} + \ldots + \lambda_{t-2}^{(0)} + 1 \right) \tilde{P}_{t-1}^{(c)} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t=c\}} \\
&= \frac{1}{N_t^{(0)}} \cdot \left(\lambda_1^{(0)} \cdots \lambda_{t-1}^{(0)} + \ldots + \lambda_{t-2}^{(0)} \lambda_{t-1}^{(0)} + \lambda_{t-1}^{(0)} \right) \tilde{P}_{t-1}^{(c)} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t=c\}} \\
&\overset{(4.53)}{=} \frac{1}{N_t^{(0)}} \cdot \left(v_1^{(0)} + \ldots + v_{t-1}^{(0)} + 1 - 1 \right) \tilde{P}_{t-1}^{(c)} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t=c\}} \\
&\overset{(4.56)}{=} \frac{1}{N_t^{(0)}} \cdot \left(N_t^{(0)} - 1 \right) \tilde{P}_{t-1}^{(c)} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t=c\}} \\
&= \left(1 - \frac{1}{N_t^{(0)}} \right) \tilde{P}_{t-1}^{(c)} + \frac{1}{N_t^{(0)}} \cdot \mathbb{1}_{\{c_t=c\}}, \quad c = 1, \ldots, M.
\end{aligned} \tag{B.13}
$$

Herleitung des Gradienten (4.58) der negativen log-Likelihood (4.57):

$$
\begin{aligned}
\left(J_{t+1}^{(0)} \right)' &= \frac{\partial \mathcal{L}(\tilde{P}_t^{(1)}, \ldots, \tilde{P}_t^{(M)}; c_{t+1})}{\partial \lambda^{(0)}} \\
&= \frac{\partial}{\partial \lambda^{(0)}} \left(- \sum_{c=1}^{M} \mathbb{1}_{\{c_{t+1}=c\}} \left(\log \tilde{P}_t^{(c)} - \log \left(\sum_{k=1}^{M} \tilde{P}_t^{(k)} \right) \right) \right) \\
&= - \sum_{c=1}^{M} \mathbb{1}_{\{c_{t+1}=c\}} \cdot \frac{\partial}{\partial \lambda^{(0)}} \left(\log \tilde{P}_t^{(c)} \right) + \sum_{c=1}^{M} \mathbb{1}_{\{c_{t+1}=c\}} \cdot \frac{\partial}{\partial \lambda^{(0)}} \left(\log \left(\sum_{k=1}^{M} \tilde{P}_t^{(k)} \right) \right) \\
&= - \sum_{c=1}^{M} \mathbb{1}_{\{c_{t+1}=c\}} \cdot \frac{\left(\tilde{P}_t^{(c)} \right)'}{\tilde{P}_t^{(c)}} + \sum_{c=1}^{M} \mathbb{1}_{\{c_{t+1}=c\}} \cdot \frac{\sum_{k=1}^{M} \left(\tilde{P}_t^{(k)} \right)'}{\sum_{k=1}^{M} \tilde{P}_t^{(k)}} \\
&= - \sum_{c=1}^{M} \mathbb{1}_{\{c_{t+1}=c\}} \cdot \frac{\left(\tilde{P}_t^{(c)} \right)'}{\tilde{P}_t^{(c)}} + \sum_{k=1}^{M} \left(\tilde{P}_t^{(k)} \right)' \\
&= - \sum_{c=1}^{M} \left(\mathbb{1}_{\{c_{t+1}=c\}} \cdot \frac{\left(\tilde{P}_t^{(c)} \right)'}{\tilde{P}_t^{(c)}} - \left(\tilde{P}_t^{(c)} \right)' \right) \\
&= - \sum_{c=1}^{M} \left(\left(\mathbb{1}_{\{c_{t+1}=c\}} - \tilde{P}_t^{(c)} \right) \frac{\left(\tilde{P}_t^{(c)} \right)'}{\tilde{P}_t^{(c)}} \right).
\end{aligned} \tag{B.14}
$$

B.3 Herleitungen für die Erweiterungen auf Chunks (Kapitel 5)

Die folgenden zwei Formeln sind Herleitungen für die Formel (5.4) aus Abschnitt 5.1 zur Aktualisierung der Kovarianzmatrix $\boldsymbol{S}_{t_2}$ zum Zeitpunkt t_2 durch einen Chunk von $n_{t_1:t_2}$ neuen Beobachtungen:

$$
\begin{aligned}
&\sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \frac{1}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \right) \right. \\
&\qquad \left. \cdot \left(\boldsymbol{x}_i - \frac{1}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \frac{1}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \right) \right. \\
&\qquad \left. \cdot \left(\boldsymbol{x}_i - \frac{1}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right. \right. \right. \\
&\qquad\qquad \left. \left. \left. - n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \frac{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \cdot \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \frac{n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \right) \right. \\
&\qquad \left. \cdot \left(\boldsymbol{x}_i - \frac{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \cdot \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \frac{n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \frac{n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \right) \right. \\
&\qquad \left. \cdot \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \frac{n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&\quad - \frac{n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \underbrace{\sum_{i=1}^{t_1} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}}_{=\mathbf{0}} \cdot \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \\
&\quad - \frac{n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \underbrace{\sum_{i=1}^{t_1} \left(\left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right)}_{=\mathbf{0}} \\
&\quad + \frac{\left(n^{(c)}_{t_1:t_2} \right)^2}{\left(n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2} \right)^2} \sum_{i=1}^{t_1} \left(\left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&\quad + \frac{\left(n^{(c)}_{t_1:t_2} \right)^2}{\left(n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2} \right)^2} \sum_{i=1}^{t_1} \left(\left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=1}^{t_1} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&\quad + \frac{n^{(c)}_{t_1} \left(n^{(c)}_{t_1:t_2} \right)^2}{\left(n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2} \right)^2} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T . \qquad \text{(B.15)}
\end{aligned}
$$

$$
\begin{aligned}
&\sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \frac{1}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \right) \right. \\
&\qquad \left. \cdot \left(\boldsymbol{x}_i - \frac{1}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \frac{1}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} + n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \right) \right. \\
&\qquad \left. \cdot \left(\boldsymbol{x}_i - \frac{n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} + n^{(c)}_{t_1:t_2} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} + n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - n^{(c)}_{t_1} \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \frac{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \cdot \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \frac{n^{(c)}_{t_1}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \right) \right. \\
&\qquad \left. \cdot \left(\boldsymbol{x}_i - \frac{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \cdot \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \frac{n^{(c)}_{t_1}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \frac{n^{(c)}_{t_1}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \right) \right. \\
&\qquad \left. \cdot \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} - \frac{n^{(c)}_{t_1}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&\quad - \frac{n^{(c)}_{t_1}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \underbrace{\sum_{i=t_1+1}^{t_2} \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}}}_{=\mathbf{0}} \cdot \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \\
&\quad - \frac{n^{(c)}_{t_1}}{n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2}} \underbrace{\sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right)}_{=\mathbf{0}} \\
&\quad + \frac{\left(n^{(c)}_{t_1} \right)^2}{\left(n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2} \right)^2} \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&\quad + \frac{\left(n^{(c)}_{t_1} \right)^2}{\left(n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2} \right)^2} \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&= \sum_{i=t_1+1}^{t_2} \left(\left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{x}_i - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T \cdot \mathbb{1}_{\{g(\boldsymbol{x}_i)=c\}} \right) \\
&\quad + \frac{\left(n^{(c)}_{t_1} \right)^2 n^{(c)}_{t_1:t_2}}{\left(n^{(c)}_{t_1} + n^{(c)}_{t_1:t_2} \right)^2} \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right) \left(\boldsymbol{m}^{(c)}_{n^{(c)}_{t_1}} - \boldsymbol{m}^{(c)}_{n^{(c)}_{t_1:t_2}} \right)^T . \qquad \text{(B.16)}
\end{aligned}
$$

Die folgenden Formeln sind Beweise und Herleitungen für Formeln aus Abschnitt 5.3.

Herleitung der rekursiven Variante (5.32) des gewichteten Mittelwertschätzers (5.29):

$$
\begin{aligned}
\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} &= \sum_{j=1}^{\tau} \frac{v^{(c)}_{t_j} \sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)}}{N^{(c)}_{t_\tau}} \\
&= \frac{1}{N^{(c)}_{t_\tau}} \left(\sum_{j=1}^{\tau-1} \left(\left(\prod_{k=t_j}^{t_{\tau-1}} \lambda^{(c)}_k \right) \sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)} \right) + \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)} \right) \\
&= \frac{1}{N^{(c)}_{t_\tau}} \left(\lambda^{(c)}_{t_1} \cdots \lambda^{(c)}_{t_{\tau-1}} \sum_{i=t_0+1}^{t_1} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)} + \ldots + \lambda^{(c)}_{t_{\tau-1}} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)} + \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)} \right) \\
&= \frac{1}{N^{(c)}_{t_\tau}} \left(\lambda^{(c)}_{t_{\tau-1}} \left(\lambda^{(c)}_{t_1} \cdots \lambda^{(c)}_{t_{\tau-2}} \sum_{i=t_0+1}^{t_1} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)} + \ldots + \right. \right. \\
&\qquad \left. \left. + \lambda^{(c)}_{t_{\tau-2}} \sum_{i=t_{\tau-3}+1}^{t_{\tau-2}} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)} + \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)} \right) + \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}^{(c)}_{(n^{(c)}_i)} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\lambda_{t_{\tau-1}}^{(c)}}{N_{t_\tau}^{(c)}} \underbrace{\left(\sum_{j=1}^{\tau-2} \left(\left(\prod_{k=t_j}^{t_{\tau-2}} \lambda_k^{(c)} \right) \sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) + \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)}_{= N_{t_{\tau-1}}^{(c)} \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)}} \\
&\quad + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(\lambda_{t_{\tau-1}}^{(c)} N_{t_{\tau-1}}^{(c)} \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} + \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(\lambda_{t_{\tau-1}}^{(c)} \left(v_{t_1}^{(c)} n_{t_0:t_1}^{(c)} + \ldots + v_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \right) \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} + \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(\lambda_{t_{\tau-1}}^{(c)} \left(\lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-2}}^{(c)} n_{t_0:t_1}^{(c)} + \ldots + \lambda_{t_{\tau-2}}^{(c)} n_{t_{\tau-3}:t_{\tau-2}}^{(c)} + n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \right) \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} \right. \\
&\qquad \left. + \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right) \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(\lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} n_{t_0:t_1}^{(c)} + \ldots + \lambda_{t_{\tau-2}}^{(c)} \lambda_{t_{\tau-1}}^{(c)} n_{t_{\tau-3}:t_{\tau-2}}^{(c)} + \lambda_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \right) \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} \\
&\quad + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(v_{t_1}^{(c)} n_{t_0:t_1}^{(c)} + \ldots + v_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} + n_{t_{\tau-1}:t_\tau}^{(c)} - n_{t_{\tau-1}:t_\tau}^{(c)} \right) \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} \\
&\quad + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(N_{t_\tau}^{(c)} - n_{t_{\tau-1}:t_\tau}^{(c)} \right) \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \\
&= \left(1 - \frac{n_{t_{\tau-1}:t_\tau}^{(c)}}{N_{t_\tau}^{(c)}} \right) \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)}. \qquad \text{(B.17)}
\end{aligned}
$$

Herleitung der rekursiven Variante (5.34) des Schätzers für die Kovarianzmatrix (5.31):

$$
\begin{aligned}
\tilde{\boldsymbol{\Pi}}_{t_\tau}^{(c)} &= \sum_{j=1}^{\tau} \frac{v_{t_j}^{(c)} \sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T}{N_{t_\tau}^{(c)}} \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(\sum_{j=1}^{\tau-1} \left(\left(\prod_{k=t_j}^{t_{\tau-1}} \lambda_k^{(c)} \right) \sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \right) + \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{N_{t_\tau}^{(c)}} \left(\lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} \sum_{i=t_0+1}^{t_1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T + \ldots + \right. \\
&\qquad \left. + \lambda_{t_{\tau-1}}^{(c)} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \right) + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \\
&= \frac{\lambda_{t_{\tau-1}}^{(c)}}{N_{t_\tau}^{(c)}} \left(\lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-2}}^{(c)} \sum_{i=t_0+1}^{t_1} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T + \ldots + \right. \\
&\qquad \left. + \lambda_{t_{\tau-2}}^{(c)} \sum_{i=t_{\tau-3}+1}^{t_{\tau-2}} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T + \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \right) \\
&\quad + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \\
&= \frac{\lambda_{t_{\tau-1}}^{(c)}}{N_{t_\tau}^{(c)}} \underbrace{\left(\sum_{j=1}^{\tau-2} \left(\left(\prod_{k=t_j}^{t_{\tau-2}} \lambda_k^{(c)} \right) \sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \right) + \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \right)}_{= N_{t_{\tau-1}}^{(c)} \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(c)}} \\
&\quad + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(c)}} \cdot \lambda_{t_{\tau-1}}^{(c)} N_{t_{\tau-1}}^{(c)} \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(c)} + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(c)}} \cdot \lambda_{t_{\tau-1}}^{(c)} \left(v_{t_1}^{(c)} n_{t_0:t_1}^{(c)} + \ldots + v_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(c)} \\
&\quad + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(c)}} \cdot \lambda_{t_{\tau-1}}^{(c)} \left(\lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-2}}^{(c)} n_{t_0:t_1}^{(c)} + \ldots + \lambda_{t_{\tau-2}}^{(c)} n_{t_{\tau-3}:t_{\tau-2}}^{(c)} + n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(c)} \\
&\quad + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(\lambda_{t_1}^{(c)} \cdots \lambda_{t_{\tau-1}}^{(c)} n_{t_0:t_1}^{(c)} + \ldots + \lambda_{t_{\tau-2}}^{(c)} \lambda_{t_{\tau-1}}^{(c)} n_{t_{\tau-3}:t_{\tau-2}}^{(c)} \right. \\
&\qquad \left. + \lambda_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(c)} + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(v_{t_1}^{(c)} n_{t_0:t_1}^{(c)} + \ldots + v_{t_{\tau-1}}^{(c)} n_{t_{\tau-2}:t_{\tau-1}}^{(c)} + n_{t_{\tau-1}:t_\tau}^{(c)} - n_{t_{\tau-1}:t_\tau}^{(c)} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(c)} \\
&\quad + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(c)}} \left(N_{t_\tau}^{(c)} - n_{t_{\tau-1}:t_\tau}^{(c)} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(c)} + \frac{1}{N_{t_\tau}^{(c)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}_{(n_i^{(c)})}^{(c)} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} \right)^T
\end{aligned}
$$

$$
\begin{aligned}
&= \left(1 - \frac{n^{(c)}_{t_{\tau-1}:t_\tau}}{N^{(c)}_{t_\tau}}\right) \tilde{\mathbf{\Pi}}^{(c)}_{t_{\tau-1}} + \frac{1}{N^{(c)}_{t_\tau}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}^{(c)}_{(n_i^{(c)})} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})}\right)^T, \\
\tilde{\mathbf{\Sigma}}^{(c)}_{t_\tau} &= \sum_{j=1}^{\tau} \frac{v^{(c)}_{t_j} \sum_{i=t_{j-1}+1}^{t_j} \boldsymbol{x}^{(c)}_{(n_i^{(c)})} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})}\right)^T}{N^{(c)}_{t_\tau}} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}}\right)^T \\
&= \left(1 - \frac{n^{(c)}_{t_{\tau-1}:t_\tau}}{N^{(c)}_{t_\tau}}\right) \tilde{\mathbf{\Pi}}^{(c)}_{t_{\tau-1}} + \frac{1}{N^{(c)}_{t_\tau}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \boldsymbol{x}^{(c)}_{(n_i^{(c)})} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})}\right)^T - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}}\right)^T \\
&= \tilde{\mathbf{\Pi}}^{(c)}_{t_\tau} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \left(\tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}}\right)^T. \qquad \text{(B.18)}
\end{aligned}
$$

Herleitung der rekursiven Variante (5.35) der Normierungskonstante (5.30):

$$
\begin{aligned}
N^{(c)}_{t_\tau} &= \sum_{j=1}^{\tau} v^{(c)}_{t_j} n^{(c)}_{t_{j-1}:t_j} = \sum_{j=1}^{\tau-1} \left(\left(\prod_{i=t_j}^{t_{\tau-1}} \lambda^{(c)}_i \right) n^{(c)}_{t_{j-1}:t_j} \right) + n^{(c)}_{t_{\tau-1}:t_\tau} \\
&= \lambda^{(c)}_{t_1} \cdots \lambda^{(c)}_{t_{\tau-1}} n^{(c)}_{t_0:t_1} + \lambda^{(c)}_{t_2} \cdots \lambda^{(c)}_{t_{\tau-1}} n^{(c)}_{t_1:t_2} + \ldots + \lambda^{(c)}_{t_{\tau-1}} n^{(c)}_{t_{\tau-2}:t_{\tau-1}} + n^{(c)}_{t_{\tau-1}:t_\tau} \\
&= \lambda^{(c)}_{t_{\tau-1}} \left(\lambda^{(c)}_{t_1} \cdots \lambda^{(c)}_{t_{\tau-2}} n^{(c)}_{t_0:t_1} + \lambda^{(c)}_{t_2} \cdots \lambda^{(c)}_{t_{\tau-2}} n^{(c)}_{t_1:t_2} + \ldots + \lambda^{(c)}_{t_{\tau-2}} n^{(c)}_{t_{\tau-3}:t_{\tau-2}} + n^{(c)}_{t_{\tau-2}:t_{\tau-1}} \right) \\
&\quad + n^{(c)}_{t_{\tau-1}:t_\tau} \\
&= \lambda^{(c)}_{t_{\tau-1}} \left(\sum_{j=1}^{\tau-2} \left(\left(\prod_{i=t_j}^{t_{\tau-2}} \lambda^{(c)}_i \right) n^{(c)}_{t_{j-1}:t_j} \right) + n^{(c)}_{t_{\tau-2}:t_{\tau-1}} \right) + n^{(c)}_{t_{\tau-1}:t_\tau} = \lambda^{(c)}_{t_{\tau-1}} N^{(c)}_{t_{\tau-1}} + n^{(c)}_{t_{\tau-1}:t_\tau}.
\end{aligned}
\qquad \text{(B.19)}
$$

Zusammenhang von (5.45) und der rekursiven Variante (5.44) des Schätzers für die gepoolte Kovarianzmatrix bei dem *LDA-AF* Algorithmus auf Chunks:

$$
\begin{aligned}
\tilde{\mathbf{\Pi}}^{(P)}_{t_\tau} &= \sum_{c=1}^{M} \sum_{j=1}^{\tau} \left(\frac{v^{(P)}_{t_j}}{N^{(P)}_{t_\tau}} \sum_{i=t_{j-1}+1}^{t_j} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_j}} \right) \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_j}} \right)^T \right) \\
&= \frac{1}{N^{(P)}_{t_\tau}} \sum_{c=1}^{M} \left(\sum_{j=1}^{\tau-1} \left(\left(\prod_{k=t_j}^{t_{\tau-1}} \lambda^{(P)}_k \right) \sum_{i=t_{j-1}+1}^{t_j} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_j}} \right) \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_j}} \right)^T \right) \right. \\
&\qquad \left. + \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right) \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right)^T \right) \\
&= \frac{1}{N^{(P)}_{t_\tau}} \sum_{c=1}^{M} \left(\lambda^{(P)}_{t_1} \cdots \lambda^{(P)}_{t_{\tau-1}} \sum_{i=t_0+1}^{t_1} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_1}} \right) \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_1}} \right)^T \right. \\
&\qquad \left. + \ldots + \lambda^{(P)}_{t_{\tau-1}} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_{\tau-1}}} \right) \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_{\tau-1}}} \right)^T \right) \\
&\quad + \frac{1}{N^{(P)}_{t_\tau}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right) \left(\boldsymbol{x}^{(c)}_{(n_i^{(c)})} - \tilde{\boldsymbol{m}}^{(c)}_{n^{(c)}_{t_\tau}} \right)^T
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\lambda_{t_{\tau-1}}^{(P)}}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \Bigg(\lambda_{t_1}^{(P)} \cdots \lambda_{t_{\tau-2}}^{(P)} \sum_{i=t_0+1}^{t_1} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_1}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_1}^{(c)}}^{(c)} \right)^T \\
&\qquad + \ldots + \lambda_{t_{\tau-2}}^{(P)} \sum_{i=t_{\tau-3}+1}^{t_{\tau-2}} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_{\tau-2}}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_{\tau-2}}^{(c)}}^{(c)} \right)^T \\
&\qquad + \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} \right)^T \Bigg) \\
&\quad + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T \\
&= \frac{\lambda_{t_{\tau-1}}^{(P)}}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \Bigg(\sum_{j=1}^{\tau-2} \Bigg(\left(\prod_{k=t_j}^{t_{\tau-2}} \lambda_k^{(P)} \right) \sum_{i=t_{j-1}+1}^{t_j} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_j}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_j}^{(c)}}^{(c)} \right)^T \Bigg) \\
&\qquad \underbrace{+ \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_{\tau-1}}^{(c)}}^{(c)} \right)^T \Bigg)}_{= N_{t_{\tau-1}}^{(P)} \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(P)}} \\
&\quad + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(P)}} \cdot \lambda_{t_{\tau-1}}^{(P)} N_{t_{\tau-1}}^{(P)} \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(P)} + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(P)}} \cdot \lambda_{t_{\tau-1}}^{(P)} \left(v_{t_1}^{(P)} n_{t_0:t_1} + \ldots + v_{t_{\tau-1}}^{(P)} n_{t_{\tau-2}:t_{\tau-1}} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(P)} \\
&\quad + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(P)}} \cdot \lambda_{t_{\tau-1}}^{(P)} \left(\lambda_{t_1}^{(P)} \cdots \lambda_{t_{\tau-2}}^{(P)} n_{t_0:t_1} + \ldots + \lambda_{t_{\tau-2}}^{(P)} n_{t_{\tau-3}:t_{\tau-2}} + n_{t_{\tau-2}:t_{\tau-1}} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(P)} \\
&\quad + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(P)}} \left(\lambda_{t_1}^{(P)} \cdots \lambda_{t_{\tau-1}}^{(P)} n_{t_0:t_1} + \ldots + \lambda_{t_{\tau-2}}^{(P)} \lambda_{t_{\tau-1}}^{(P)} n_{t_{\tau-3}:t_{\tau-2}} + \lambda_{t_{\tau-1}}^{(P)} n_{t_{\tau-2}:t_{\tau-1}} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(P)} \\
&\quad + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T \\
&= \frac{1}{N_{t_\tau}^{(P)}} \left(v_{t_1}^{(P)} n_{t_0:t_1} + \ldots + v_{t_{\tau-1}}^{(P)} n_{t_{\tau-2}:t_{\tau-1}} + n_{t_{\tau-1}:t_\tau} - n_{t_{\tau-1}:t_\tau} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(P)} \\
&\quad + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{N_{t_\tau}^{(P)}} \left(N_{t_\tau}^{(P)} - n_{t_{\tau-1}:t_\tau} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(P)} \\
&\quad + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T \\
&= \left(1 - \frac{n_{t_{\tau-1}:t_\tau}}{N_{t_\tau}^{(P)}} \right) \tilde{\boldsymbol{\Pi}}_{t_{\tau-1}}^{(P)} + \frac{1}{N_{t_\tau}^{(P)}} \sum_{c=1}^{M} \sum_{i=t_{\tau-1}+1}^{t_\tau} \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right) \left(\boldsymbol{x}_{(n_i^{(c)})}^{(c)} - \tilde{\boldsymbol{m}}_{n_{t_\tau}^{(c)}}^{(c)} \right)^T .
\end{aligned}
\tag{B.20}
$$

Herleitung der rekursiven Variante (5.49) des Schätzers (5.48) für die a-priori Wahrscheinlichkeit von Klasse c:

$$
\begin{aligned}
\tilde{P}_{t_\tau}^{(c)} &= \sum_{j=1}^{\tau} \frac{v_{t_j}^{(0)} \sum_{i=t_{j-1}+1}^{t_j} \mathbb{1}_{\{c_i=c\}}}{N_{t_\tau}^{(0)}} \\
&= \frac{1}{N_{t_\tau}^{(0)}} \left(\sum_{j=1}^{\tau-1} \left(\left(\prod_{k=t_j}^{t_{\tau-1}} \lambda_k^{(0)} \right) \sum_{i=t_{j-1}+1}^{t_j} \mathbb{1}_{\{c_i=c\}} \right) + \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \right) \\
&= \frac{1}{N_{t_\tau}^{(0)}} \left(\lambda_{t_1}^{(0)} \cdots \lambda_{t_{\tau-1}}^{(0)} \sum_{i=t_0+1}^{t_1} \mathbb{1}_{\{c_i=c\}} + \ldots + \lambda_{t_{\tau-1}}^{(0)} \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \mathbb{1}_{\{c_i=c\}} + \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \right) \\
&= \frac{1}{N_{t_\tau}^{(0)}} \left(\lambda_{t_{\tau-1}}^{(0)} \left(\lambda_{t_1}^{(0)} \cdots \lambda_{t_{\tau-2}}^{(0)} \sum_{i=t_0+1}^{t_1} \mathbb{1}_{\{c_i=c\}} + \ldots + \right. \right. \\
&\qquad \left. \left. + \lambda_{t_{\tau-2}}^{(0)} \sum_{i=t_{\tau-3}+1}^{t_{\tau-2}} \mathbb{1}_{\{c_i=c\}} + \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \mathbb{1}_{\{c_i=c\}} \right) + \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \right) \\
&= \frac{\lambda_{t_{\tau-1}}^{(0)}}{N_{t_\tau}^{(0)}} \underbrace{\left(\sum_{j=1}^{\tau-2} \left(\left(\prod_{k=t_j}^{t_{\tau-2}} \lambda_k^{(0)} \right) \sum_{i=t_{j-1}+1}^{t_j} \mathbb{1}_{\{c_i=c\}} \right) + \sum_{i=t_{\tau-2}+1}^{t_{\tau-1}} \mathbb{1}_{\{c_i=c\}} \right)}_{= N_{t_{\tau-1}}^{(0)} \tilde{P}_{t_{\tau-1}}^{(c)}} \\
&\quad + \frac{1}{N_{t_\tau}^{(0)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \\
&= \frac{1}{N_{t_\tau}^{(0)}} \left(\lambda_{t_{\tau-1}}^{(0)} N_{t_{\tau-1}}^{(0)} \tilde{P}_{t_{\tau-1}}^{(c)} + \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \right) \\
&= \frac{1}{N_{t_\tau}^{(0)}} \left(\lambda_{t_{\tau-1}}^{(0)} \left(v_{t_1}^{(0)} n_{t_0:t_1} + \ldots + v_{t_{\tau-1}}^{(0)} n_{t_{\tau-2}:t_{\tau-1}} \right) \tilde{P}_{t_{\tau-1}}^{(c)} + \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \right) \\
&= \frac{1}{N_{t_\tau}^{(0)}} \left(\lambda_{t_{\tau-1}}^{(0)} \left(\lambda_{t_1}^{(0)} \cdots \lambda_{t_{\tau-2}}^{(0)} n_{t_0:t_1} + \ldots + \lambda_{t_{\tau-2}}^{(0)} n_{t_{\tau-3}:t_{\tau-2}} + n_{t_{\tau-2}:t_{\tau-1}} \right) \tilde{P}_{t_{\tau-1}}^{(c)} \right. \\
&\qquad \left. + \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{N_{t_\tau}^{(0)}} \left(\lambda_{t_1}^{(0)} \cdots \lambda_{t_{\tau-1}}^{(0)} n_{t_0:t_1} + \ldots + \lambda_{t_{\tau-2}}^{(0)} \lambda_{t_{\tau-1}}^{(0)} n_{t_{\tau-3}:t_{\tau-2}} + \lambda_{t_{\tau-1}}^{(0)} n_{t_{\tau-2}:t_{\tau-1}} \right) \tilde{P}_{t_{\tau-1}}^{(c)} \\
&\quad + \frac{1}{N_{t_\tau}^{(0)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \\
&= \frac{1}{N_{t_\tau}^{(0)}} \left(v_{t_1}^{(0)} n_{t_0:t_1} + \ldots + v_{t_{\tau-1}}^{(0)} n_{t_{\tau-2}:t_{\tau-1}} + n_{t_{\tau-1}:t_\tau} - n_{t_{\tau-1}:t_\tau} \right) \tilde{P}_{t_{\tau-1}}^{(c)} \\
&\quad + \frac{1}{N_{t_\tau}^{(0)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \\
&= \frac{1}{N_{t_\tau}^{(0)}} \left(N_{t_\tau}^{(0)} - n_{t_{\tau-1}:t_\tau} \right) \tilde{P}_{t_{\tau-1}}^{(c)} + \frac{1}{N_{t_\tau}^{(0)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \\
&= \left(1 - \frac{n_{t_{\tau-1}:t_\tau}}{N_{t_\tau}^{(0)}} \right) \tilde{P}_{t_{\tau-1}}^{(c)} + \frac{1}{N_{t_\tau}^{(0)}} \sum_{i=t_{\tau-1}+1}^{t_\tau} \mathbb{1}_{\{c_i=c\}} \\
&= \left(1 - \frac{n_{t_{\tau-1}:t_\tau}}{N_{t_\tau}^{(0)}} \right) \tilde{P}_{t_{\tau-1}}^{(c)} + \frac{1}{N_{t_\tau}^{(0)}} \cdot n_{t_{\tau-1}:t_\tau}^{(c)}. \qquad \text{(B.21)}
\end{aligned}
$$

Herleitung der rekursiven Variante der Normierungskonstante (5.50):

$$
\begin{aligned}
N_{t_\tau}^{(0)} &= \sum_{j=1}^{\tau} v_{t_j}^{(0)} n_{t_{j-1}:t_j} = \sum_{j=1}^{\tau-1} \left(\left(\prod_{i=t_j}^{t_{\tau-1}} \lambda_i^{(0)} \right) n_{t_{j-1}:t_j} \right) + n_{t_{\tau-1}:t_\tau} \\
&= \lambda_{t_1}^{(0)} \cdots \lambda_{t_{\tau-1}}^{(0)} n_{t_0:t_1} + \lambda_{t_2}^{(0)} \cdots \lambda_{t_{\tau-1}}^{(0)} n_{t_1:t_2} + \ldots + \lambda_{t_{\tau-1}}^{(0)} n_{t_{\tau-2}:t_{\tau-1}} + n_{t_{\tau-1}:t_\tau} \\
&= \lambda_{t_{\tau-1}}^{(0)} \left(\lambda_{t_1}^{(0)} \cdots \lambda_{t_{\tau-2}}^{(0)} n_{t_0:t_1} + \lambda_{t_2}^{(0)} \cdots \lambda_{t_{\tau-2}}^{(0)} n_{t_1:t_2} + \ldots + \lambda_{t_{\tau-2}}^{(0)} n_{t_{\tau-3}:t_{\tau-2}} + n_{t_{\tau-2}:t_{\tau-1}} \right) \\
&\quad + n_{t_{\tau-1}:t_\tau} \\
&= \lambda_{t_{\tau-1}}^{(0)} \left(\sum_{j=1}^{\tau-2} \left(\left(\prod_{i=t_j}^{t_{\tau-2}} \lambda_i^{(0)} \right) n_{t_{j-1}:t_j} \right) + n_{t_{\tau-2}:t_{\tau-1}} \right) + n_{t_{\tau-1}:t_\tau} \\
&= \lambda_{t_{\tau-1}}^{(0)} N_{t_{\tau-1}}^{(0)} + n_{t_{\tau-1}:t_\tau}. \qquad \text{(B.22)}
\end{aligned}
$$

C Tabellen

C.1 Progosefehler

Tabelle C.1: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Kreuzen“ ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4954 (0.001)	0.0764 (0.001)	0.0689 (0.001)	0.1	0.5017 (0.000)	5	0.0724 (0.001)
						20	0.0688 (0.001)
						50	0.0709 (0.001)
				0.3	0.5012 (0.000)		
				0.5	0.4954 (0.001)	5	0.0713 (0.001)
						20	*0.0678* (0.001)
						50	0.0702 (0.001)
				0.7	0.2139 (0.001)		
				0.9	0.0740 (0.001)	5	0.0719 (0.001)
						20	0.0689 (0.001)
						50	0.0707 (0.001)
10	0.0968 (0.003)	0.0872 (0.002)	0.0833 (0.002)	0.1	0.0867 (0.002)	5	0.0872 (0.002)
						20	0.0836 (0.002)
						50	0.0829 (0.002)
				0.3	0.0981 (0.003)		
				0.5	0.0968 (0.003)	5	0.0849 (0.002)
						20	0.0810 (0.002)
						50	0.0805 (0.002)
				0.7	0.0860 (0.002)		
				0.9	*0.0743* (0.001)	5	0.0825 (0.002)
						20	0.0784 (0.002)
						50	0.0790 (0.002)
20	0.0808 (0.002)	0.0820 (0.002)	0.0763 (0.001)	0.1	0.0742 (0.001)	5	0.0790 (0.002)
						20	0.0760 (0.002)
						50	0.0758 (0.002)
				0.3	0.0817 (0.002)		
				0.5	0.0808 (0.002)	5	0.0768 (0.002)
						20	0.0734 (0.001)
						50	0.0739 (0.001)
				0.7	0.0735 (0.001)		
				0.9	*0.0660* (0.001)	5	0.0757 (0.001)
						20	0.0726 (0.001)
						50	0.0735 (0.001)
50	0.0705 (0.001)	0.0778 (0.001)	0.0711 (0.001)	0.1	0.0667 (0.001)	5	0.0720 (0.001)
						20	0.0709 (0.001)
						50	0.0729 (0.002)
				0.3	0.0711 (0.001)		
				0.5	0.0705 (0.001)	5	0.0702 (0.001)
						20	0.0685 (0.001)
						50	0.0715 (0.001)
				0.7	0.0659 (0.001)		
				0.9	*0.0615* (0.001)	5	0.0699 (0.001)
						20	0.0683 (0.001)
						50	0.0715 (0.001)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
100	0.0659 (0.001)	0.0751 (0.001)	0.0683 (0.001)	0.1	0.0635 (0.001)	5	0.0682 (0.001)
						20	0.0677 (0.001)
						50	0.0697 (0.001)
				0.3	0.0664 (0.001)		
				0.5	0.0659 (0.001)	5	0.0667 (0.001)
						20	0.0655 (0.001)
						50	0.0686 (0.001)
				0.7	0.0627 (0.001)		
				0.9	*0.0599* (0.000)	5	0.0665 (0.001)
						20	0.0654 (0.001)
						50	0.0686 (0.001)
200	0.0632 (0.001)	0.0735 (0.001)	0.0665 (0.001)	0.1	0.0617 (0.001)	5	0.0659 (0.001)
						20	0.0651 (0.001)
						50	0.0680 (0.001)
				0.3	0.0635 (0.001)		
				0.5	0.0632 (0.001)	5	0.0645 (0.001)
						20	0.0631 (0.001)
						50	0.0668 (0.001)
				0.7	0.0609 (0.000)		
				0.9	*0.0591* (0.000)	5	0.0646 (0.001)
						20	0.0632 (0.001)
						50	0.0670 (0.001)
300	**0.0622** (0.001)	**0.0728** (0.001)	**0.0657** (0.001)	0.1	0.0610 (0.000)	5	0.0651 (0.001)
						20	0.0641 (0.001)
						50	0.0676 (0.001)
				0.3	0.0624 (0.001)		
				0.5	0.0622 (0.001)	5	0.0636 (0.001)
						20	**0.0620** (0.001)
						50	0.0660 (0.001)
				0.7	0.0603 (0.000)		
				0.9	***0.0587*** (0.000)	5	0.0638 (0.001)
						20	0.0622 (0.001)
						50	0.0665 (0.001)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0564 (Standardabweichung 0.116)

Tabelle C.2: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.0390 (0.000)	0.0111 (0.000)	0.0070 (0.000)	0.1	0.2581 (0.009)	5	0.0070 (0.000)
						20	0.0133 (0.000)
						50	0.0106 (0.000)
				0.3	0.0722 (0.001)		
				0.5	0.0390 (0.000)	5	0.0055 (0.000)
						20	0.0083 (0.000)
						50	0.0066 (0.000)
				0.7	0.0203 (0.000)		
				0.9	0.0068 (0.000)	5	***0.0045*** (0.000)
						20	0.0046 (0.000)
						50	0.0046 (0.000)
10	0.0172 (0.000)	0.0146 (0.001)	0.0113 (0.001)	0.1	0.0143 (0.000)	5	0.0109 (0.000)
						20	0.0157 (0.000)
						50	0.0133 (0.000)
				0.3	0.0177 (0.000)		
				0.5	0.0172 (0.000)	5	0.0102 (0.000)
						20	0.0122 (0.000)
						50	0.0108 (0.000)
				0.7	0.0133 (0.000)		
				0.9	0.0076 (0.000)	5	0.0074 (0.000)
						20	*0.0071* (0.000)
						50	0.0075 (0.000)
20	0.0133 (0.000)	0.0123 (0.000)	0.0089 (0.000)	0.1	0.0110 (0.000)	5	0.0087 (0.000)
						20	0.0125 (0.000)
						50	0.0105 (0.000)
				0.3	0.0137 (0.000)		
				0.5	0.0133 (0.000)	5	0.0081 (0.000)
						20	0.0097 (0.000)
						50	0.0085 (0.000)
				0.7	0.0102 (0.000)		
				0.9	0.0058 (0.000)	5	0.0057 (0.000)
						20	*0.0055* (0.000)
						50	0.0057 (0.000)
50	0.0119 (0.000)	0.0108 (0.000)	0.0075 (0.000)	0.1	0.0098 (0.000)	5	0.0079 (0.000)
						20	0.0113 (0.000)
						50	0.0095 (0.000)
				0.3	0.0122 (0.000)		
				0.5	0.0119 (0.000)	5	0.0073 (0.000)
						20	0.0088 (0.000)
						50	0.0076 (0.000)
				0.7	0.0091 (0.000)		
				0.9	0.0052 (0.000)	5	0.0050 (0.000)
						20	*0.0049* (0.000)
						50	0.0051 (0.000)
100	0.0114 (0.000)	0.0098 (0.000)	0.0069 (0.000)	0.1	0.0095 (0.000)	5	0.0076 (0.000)
						20	0.0109 (0.000)
						50	0.0092 (0.000)
				0.3	0.0118 (0.000)		
				0.5	0.0114 (0.000)	5	0.0070 (0.000)
						20	0.0085 (0.000)
						50	0.0074 (0.000)
				0.7	0.0088 (0.000)		
				0.9	0.0050 (0.000)	5	*0.0047* (0.000)
						20	*0.0047* (0.000)
						50	0.0048 (0.000)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	**0.0112** (0.000)	0.0093 (0.000)	0.0065 (0.000)	0.1	0.0093 (0.000)	5	0.0074 (0.000)
						20	0.0107 (0.000)
						50	0.0091 (0.000)
				0.3	0.0116 (0.000)		
				0.5	0.0112 (0.000)	5	0.0068 (0.000)
						20	0.0083 (0.000)
						50	0.0072 (0.000)
				0.7	0.0086 (0.000)		
				0.9	0.0049 (0.000)	5	*0.0046* (0.000)
						20	*0.0046* (0.000)
						50	*0.0046* (0.000)
300	**0.0112** (0.000)	**0.0091** (0.000)	**0.0064** (0.000)	0.1	0.0093 (0.000)	5	0.0074 (0.000)
						20	0.0106 (0.000)
						50	0.0090 (0.000)
				0.3	0.0115 (0.000)		
				0.5	0.0112 (0.000)	5	0.0068 (0.000)
						20	0.0082 (0.000)
						50	0.0072 (0.000)
				0.7	0.0086 (0.000)		
				0.9	**0.0048** (0.000)	5	***0.0045*** (0.000)
						20	***0.0045*** (0.000)
						50	0.0046 (0.000)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0039 (Standardabweichung 0.010)

Tabelle C.3: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **Gradual Drift mit „Kreuzen“ ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4898 (0.003)	*0.0039* (0.000)	**0.2383** (0.009)	0.1	0.4975 (0.002)	5	0.2349 (0.013)
						20	0.2326 (0.012)
						50	**0.2281** (0.012)
				0.3	0.5006 (0.002)		
				0.5	0.4898 (0.003)	5	0.2345 (0.013)
						20	0.2327 (0.012)
						50	0.2291 (0.012)
				0.7	0.2970 (0.003)		
				0.9	0.2510 (0.002)	5	0.2342 (0.013)
						20	0.2328 (0.012)
						50	0.2303 (0.012)
10	0.2604 (0.011)	*0.0039* (0.000)	0.2524 (0.015)	0.1	0.2401 (0.012)	5	0.2441 (0.016)
						20	0.2465 (0.014)
						50	0.2541 (0.014)
				0.3	0.2518 (0.011)		
				0.5	0.2604 (0.011)	5	0.2440 (0.016)
						20	0.2463 (0.015)
						50	0.2538 (0.014)
				0.7	0.2635 (0.011)		
				0.9	0.2604 (0.011)	5	0.2440 (0.016)
						20	0.2459 (0.015)
						50	0.2541 (0.014)
20	0.2504 (0.007)	*0.0040* (0.000)	0.2480 (0.012)	0.1	**0.2391** (0.007)	5	0.2409 (0.013)
						20	0.2415 (0.012)
						50	0.2486 (0.011)
				0.3	0.2452 (0.007)		
				0.5	0.2504 (0.007)	5	0.2408 (0.013)
						20	0.2417 (0.012)
						50	0.2483 (0.011)
				0.7	0.2525 (0.007)		
				0.9	0.2510 (0.007)	5	0.2408 (0.014)
						20	0.2417 (0.013)
						50	0.2486 (0.012)
50	**0.2454** (0.004)	*0.0040* (0.000)	0.2426 (0.010)	0.1	0.2403 (0.004)	5	0.2379 (0.011)
						20	0.2351 (0.010)
						50	0.2489 (0.010)
				0.3	0.2430 (0.004)		
				0.5	0.2454 (0.004)	5	0.2377 (0.011)
						20	0.2355 (0.010)
						50	0.2485 (0.010)
				0.7	0.2465 (0.004)		
				0.9	0.2451 (0.004)	5	0.2375 (0.011)
						20	0.2357 (0.011)
						50	0.2470 (0.010)
100	0.2462 (0.003)	*0.0038* (0.000)	0.2403 (0.008)	0.1	0.2432 (0.003)	5	0.2352 (0.010)
						20	0.2332 (0.009)
						50	0.2399 (0.009)
				0.3	0.2450 (0.003)		
				0.5	0.2462 (0.003)	5	0.2349 (0.010)
						20	0.2339 (0.009)
						50	0.2402 (0.009)
				0.7	0.2466 (0.003)		
				0.9	0.2449 (0.003)	5	0.2346 (0.010)
						20	0.2338 (0.009)
						50	0.2385 (0.009)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.2474 (0.002)	***0.0036*** (0.000)	0.2389 (0.007)	0.1	0.2458 (0.002)	5	0.2332 (0.008)
						20	0.2323 (0.008)
						50	0.2350 (0.008)
				0.3	0.2467 (0.002)		
				0.5	0.2474 (0.002)	5	0.2329 (0.008)
						20	0.2328 (0.008)
						50	0.2344 (0.008)
				0.7	0.2474 (0.003)		
				0.9	0.2457 (0.003)	5	0.2324 (0.008)
						20	0.2325 (0.008)
						50	0.2334 (0.009)
300	0.2475 (0.002)	***0.0036*** (0.000)	0.2384 (0.007)	0.1	0.2466 (0.002)	5	0.2320 (0.008)
						20	0.2323 (0.007)
						50	0.2342 (0.008)
				0.3	0.2471 (0.002)		
				0.5	0.2475 (0.002)	5	0.2317 (0.008)
						20	0.2326 (0.007)
						50	0.2331 (0.008)
				0.7	0.2475 (0.002)		
				0.9	0.2460 (0.003)	5	0.2312 (0.008)
						20	0.2321 (0.008)
						50	0.2321 (0.008)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: < 0.0001 (Standardabw. < 0.001)

Tabelle C.4: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **Gradual Drift mit „Austausch“ ($p = 3$)** getrennt nach Methoden für Online DA und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4870 (0.003)	0.3608 (0.004)	0.3475 (0.004)	0.1	0.5020 (0.002)	5	0.3693 (0.006)
						20	0.3595 (0.006)
						50	0.3619 (0.006)
				0.3	0.5015 (0.002)		
				0.5	0.4870 (0.003)	5	0.3697 (0.006)
						20	0.3617 (0.006)
						50	0.3639 (0.006)
				0.7	0.3337 (0.002)		
				0.9	*0.3015* (0.002)	5	0.3713 (0.006)
						20	0.3670 (0.006)
						50	0.3706 (0.007)
10	0.3763 (0.007)	0.3936 (0.007)	0.3910 (0.008)	0.1	*0.3652* (0.008)	5	0.3955 (0.008)
						20	0.3871 (0.008)
						50	0.3846 (0.008)
				0.3	0.3716 (0.008)		
				0.5	0.3763 (0.007)	5	0.3962 (0.008)
						20	0.3888 (0.008)
						50	0.3866 (0.008)
				0.7	0.3779 (0.007)		
				0.9	0.3775 (0.007)	5	0.3983 (0.009)
						20	0.3942 (0.008)
						50	0.3926 (0.008)
20	0.3453 (0.005)	0.3815 (0.006)	0.3750 (0.007)	0.1	*0.3356* (0.005)	5	0.3811 (0.007)
						20	0.3715 (0.006)
						50	0.3686 (0.006)
				0.3	0.3405 (0.005)		
				0.5	0.3453 (0.005)	5	0.3816 (0.007)
						20	0.3732 (0.007)
						50	0.3709 (0.006)
				0.7	0.3474 (0.005)		
				0.9	0.3490 (0.005)	5	0.3835 (0.007)
						20	0.3787 (0.007)
						50	0.3765 (0.007)
50	0.3206 (0.003)	0.3694 (0.005)	0.3580 (0.005)	0.1	*0.3145* (0.003)	5	0.3619 (0.005)
						20	0.3566 (0.005)
						50	0.3614 (0.006)
				0.3	0.3168 (0.003)		
				0.5	0.3206 (0.003)	5	0.3621 (0.005)
						20	0.3575 (0.005)
						50	0.3636 (0.006)
				0.7	0.3230 (0.003)		
				0.9	0.3268 (0.003)	5	0.3633 (0.005)
						20	0.3610 (0.005)
						50	0.3658 (0.006)
100	0.3094 (0.002)	0.3615 (0.004)	0.3467 (0.004)	0.1	*0.3062* (0.002)	5	0.3477 (0.004)
						20	0.3454 (0.004)
						50	0.3489 (0.005)
				0.3	0.3065 (0.002)		
				0.5	0.3094 (0.002)	5	0.3478 (0.004)
						20	0.3457 (0.004)
						50	0.3500 (0.005)
				0.7	0.3118 (0.002)		
				0.9	0.3166 (0.003)	5	0.3491 (0.004)
						20	0.3483 (0.004)
						50	0.3522 (0.005)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.3024 (0.002)	0.3564 (0.004)	0.3385 (0.004)	0.1	0.3019 (0.002)	5	0.3380 (0.003)
						20	0.3342 (0.003)
						50	0.3401 (0.004)
				0.3	*0.3004* (0.002)		
				0.5	0.3024 (0.002)	5	0.3382 (0.003)
						20	0.3344 (0.003)
						50	0.3411 (0.004)
				0.7	0.3046 (0.002)		
				0.9	0.3098 (0.002)	5	0.3397 (0.003)
						20	0.3377 (0.004)
						50	0.3438 (0.004)
300	**0.2998** (0.002)	**0.3545** (0.004)	**0.3350** (0.004)	0.1	0.3008 (0.002)	5	0.3344 (0.003)
						20	**0.3287** (0.003)
						50	0.3350 (0.004)
				0.3	***0.2982*** (0.002)		
				0.5	0.2998 (0.002)	5	0.3347 (0.003)
						20	0.3292 (0.003)
						50	0.3356 (0.004)
				0.7	0.3020 (0.002)		
				0.9	0.3069 (0.002)	5	0.3363 (0.003)
						20	0.3332 (0.003)
						50	0.3393 (0.004)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.2893 (Standardabweichung 0.122)

Tabelle C.5: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **Sudden Drift ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4840 (0.003)	0.1288 (0.003)	**0.1175** (0.003)	0.1	0.5043 (0.008)	5	0.1271 (0.003)
						20	0.1128 (0.002)
						50	0.1097 (0.003)
				0.3	0.5001 (0.002)		
				0.5	0.4840 (0.003)	5	0.1274 (0.003)
						20	0.1122 (0.002)
						50	*0.1090* (0.003)
				0.7	0.3128 (0.002)		
				0.9	0.1594 (0.001)	5	0.1289 (0.003)
						20	0.1162 (0.003)
						50	0.1152 (0.003)
10	0.1358 (0.004)	0.1643 (0.006)	0.1586 (0.008)	0.1	*0.1255* (0.003)	5	0.1648 (0.008)
						20	0.1383 (0.005)
						50	0.1307 (0.004)
				0.3	0.1273 (0.004)		
				0.5	0.1358 (0.004)	5	0.1656 (0.008)
						20	0.1422 (0.006)
						50	0.1312 (0.004)
				0.7	0.1440 (0.005)		
				0.9	0.1400 (0.004)	5	0.1682 (0.009)
						20	0.1528 (0.007)
						50	0.1432 (0.006)
20	0.1067 (0.002)	0.1453 (0.004)	0.1345 (0.004)	0.1	0.1036 (0.001)	5	0.1375 (0.004)
						20	0.1164 (0.003)
						50	0.1114 (0.002)
				0.3	*0.1022* (0.001)		
				0.5	0.1067 (0.002)	5	0.1382 (0.004)
						20	0.1194 (0.003)
						50	0.1113 (0.003)
				0.7	0.1113 (0.002)		
				0.9	0.1103 (0.002)	5	0.1402 (0.004)
						20	0.1280 (0.004)
						50	0.1206 (0.003)
50	**0.0992** (0.001)	0.1330 (0.003)	0.1208 (0.003)	0.1	0.1013 (0.001)	5	0.1189 (0.002)
						20	0.1073 (0.002)
						50	0.1113 (0.003)
				0.3	***0.0976*** (0.001)		
				0.5	0.0992 (0.001)	5	0.1193 (0.002)
						20	0.1086 (0.002)
						50	0.1106 (0.003)
				0.7	0.1015 (0.001)		
				0.9	0.1020 (0.001)	5	0.1207 (0.002)
						20	0.1137 (0.002)
						50	0.1139 (0.003)
100	0.1058 (0.001)	**0.1287** (0.003)	0.1177 (0.002)	0.1	0.1107 (0.001)	5	0.1132 (0.002)
						20	***0.1047*** (0.002)
						50	0.1128 (0.003)
				0.3	0.1056 (0.001)		
				0.5	0.1058 (0.001)	5	0.1134 (0.002)
						20	0.1053 (0.002)
						50	0.1115 (0.003)
				0.7	0.1070 (0.001)		
				0.9	0.1064 (0.001)	5	0.1147 (0.002)
						20	0.1091 (0.002)
						50	0.1117 (0.003)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.1256 (0.001)	0.1309 (0.003)	0.1214 (0.003)	0.1	0.1336 (0.001)	5	0.1153 (0.002)
						20	*0.1069* (0.002)
						50	0.1130 (0.003)
				0.3	0.1267 (0.001)		
				0.5	0.1256 (0.001)	5	0.1155 (0.002)
						20	0.1073 (0.002)
						50	0.1113 (0.003)
				0.7	0.1251 (0.001)		
				0.9	0.1200 (0.001)	5	0.1169 (0.002)
						20	0.1109 (0.002)
						50	0.1129 (0.003)
300	0.1463 (0.001)	0.1359 (0.003)	0.1272 (0.003)	0.1	0.1571 (0.002)	5	0.1204 (0.002)
						20	*0.1117* (0.002)
						50	0.1153 (0.003)
				0.3	0.1489 (0.001)		
				0.5	0.1463 (0.001)	5	0.1206 (0.002)
						20	0.1124 (0.002)
						50	0.1139 (0.003)
				0.7	0.1438 (0.001)		
				0.9	0.1330 (0.001)	5	0.1221 (0.002)
						20	0.1162 (0.002)
						50	0.1168 (0.003)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0786 (Standardabweichung 0)

Tabelle C.6: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **ohne Drift ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	**0.0794** (0.001)	0.1187 (0.002)	0.1074 (0.002)	0.1	0.0939 (0.001)	5	0.1238 (0.003)
						20	0.1001 (0.002)
						50	**0.0872** (0.001)
				0.3	0.0808 (0.001)		
				0.5	0.0794 (0.001)	5	0.1242 (0.003)
						20	0.1038 (0.002)
						50	0.0900 (0.001)
				0.7	***0.0787*** (0.001)		
				0.9	0.0806 (0.001)	5	0.1258 (0.003)
						20	0.1112 (0.002)
						50	0.1059 (0.002)
10	0.1165 (0.003)	0.1594 (0.006)	0.1533 (0.007)	0.1	0.1241 (0.003)	5	0.1624 (0.008)
						20	0.1383 (0.005)
						50	0.1253 (0.004)
				0.3	0.1168 (0.003)		
				0.5	*0.1164* (0.003)	5	0.1632 (0.008)
						20	0.1427 (0.006)
						50	0.1284 (0.004)
				0.7	0.1171 (0.003)		
				0.9	0.1199 (0.003)	5	0.1655 (0.009)
						20	0.1528 (0.007)
						50	0.1448 (0.006)
20	*0.0937* (0.001)	0.1405 (0.004)	0.1293 (0.004)	0.1	0.1004 (0.001)	5	0.1349 (0.004)
						20	0.1150 (0.002)
						50	0.1035 (0.002)
				0.3	0.0940 (0.001)		
				0.5	*0.0937* (0.001)	5	0.1355 (0.004)
						20	0.1187 (0.003)
						50	0.1063 (0.002)
				0.7	0.0942 (0.001)		
				0.9	0.0969 (0.001)	5	0.1374 (0.004)
						20	0.1271 (0.003)
						50	0.1204 (0.003)
50	*0.0843* (0.001)	0.1261 (0.003)	0.1135 (0.002)	0.1	0.0910 (0.001)	5	0.1149 (0.002)
						20	0.1027 (0.001)
						50	0.0963 (0.002)
				0.3	0.0847 (0.001)		
				0.5	*0.0843* (0.001)	5	0.1151 (0.002)
						20	0.1049 (0.002)
						50	0.0986 (0.002)
				0.7	0.0848 (0.001)		
				0.9	0.0873 (0.001)	5	0.1163 (0.002)
						20	0.1100 (0.002)
						50	0.1077 (0.002)
100	*0.0813* (0.001)	0.1183 (0.002)	0.1067 (0.002)	0.1	0.0884 (0.001)	5	0.1061 (0.002)
						20	0.0970 (0.001)
						50	0.0932 (0.001)
				0.3	0.0818 (0.001)		
				0.5	*0.0813* (0.001)	5	0.1062 (0.002)
						20	0.0983 (0.001)
						50	0.0943 (0.001)
				0.7	0.0818 (0.001)		
				0.9	0.0842 (0.001)	5	0.1073 (0.002)
						20	0.1021 (0.001)
						50	0.1001 (0.001)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	*0.0800* (0.001)	0.1138 (0.002)	0.1029 (0.002)	0.1	0.0877 (0.001)	5	0.1015 (0.001)
						20	0.0932 (0.001)
						50	0.0904 (0.001)
				0.3	0.0806 (0.001)		
				0.5	*0.0800* (0.001)	5	0.1016 (0.001)
						20	0.0940 (0.001)
						50	0.0900 (0.001)
				0.7	0.0804 (0.001)		
				0.9	0.0825 (0.001)	5	0.1028 (0.001)
						20	0.0978 (0.001)
						50	0.0955 (0.001)
300	*0.0796* (0.001)	**0.1123** (0.002)	**0.1017** (0.002)	0.1	0.0878 (0.001)	5	0.0999 (0.001)
						20	0.0914 (0.001)
						50	0.0892 (0.001)
				0.3	0.0801 (0.001)		
				0.5	*0.0796* (0.001)	5	0.1001 (0.001)
						20	0.0922 (0.001)
						50	0.0883 (0.001)
				0.7	0.0799 (0.001)		
				0.9	0.0819 (0.001)	5	0.1014 (0.001)
						20	0.0962 (0.001)
						50	0.0938 (0.001)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0786 (Standardabweichung 0)

Tabelle C.7: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Kreuzen“** ($p = 10$) getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4946 (0.000)	**0.0589** (0.002)	*0.0323* (0.000)	0.1	0.5025 (0.000)	5	0.0393 (0.001)
						20	0.0380 (0.001)
						50	0.0411 (0.001)
				0.3	0.5024 (0.000)		
				0.5	0.4946 (0.000)	5	0.0364 (0.000)
						20	0.0353 (0.000)
						50	0.0376 (0.001)
				0.7	0.2142 (0.000)		
				0.9	0.0566 (0.000)	5	0.0337 (0.000)
						20	0.0325 (0.000)
						50	0.0342 (0.000)
10	0.1553 (0.004)	0.0674 (0.003)	0.0459 (0.001)	0.1	0.1220 (0.003)	5	0.0687 (0.001)
						20	0.0679 (0.001)
						50	0.0700 (0.001)
				0.3	0.1595 (0.004)		
				0.5	0.1553 (0.004)	5	0.0749 (0.001)
						20	0.0740 (0.001)
						50	0.0756 (0.002)
				0.7	0.1168 (0.003)		
				0.9	0.0573 (0.001)	5	0.0468 (0.001)
						20	*0.0458* (0.001)
						50	0.0469 (0.001)
20	0.1138 (0.002)	0.0651 (0.003)	*0.0384* (0.001)	0.1	0.0912 (0.001)	5	0.0599 (0.001)
						20	0.0580 (0.001)
						50	0.0609 (0.001)
				0.3	0.1171 (0.002)		
				0.5	0.1138 (0.002)	5	0.0641 (0.001)
						20	0.0622 (0.001)
						50	0.0641 (0.001)
				0.7	0.0867 (0.001)		
				0.9	0.0456 (0.001)	5	0.0414 (0.001)
						20	0.0398 (0.001)
						50	0.0413 (0.001)
50	0.0868 (0.001)	0.0635 (0.003)	*0.0341* (0.000)	0.1	0.0712 (0.001)	5	0.0536 (0.001)
						20	0.0522 (0.001)
						50	0.0564 (0.001)
				0.3	0.0892 (0.001)		
				0.5	0.0868 (0.001)	5	0.0559 (0.001)
						20	0.0545 (0.001)
						50	0.0574 (0.001)
				0.7	0.0678 (0.001)		
				0.9	0.0386 (0.000)	5	0.0380 (0.000)
						20	0.0369 (0.000)
						50	0.0392 (0.001)
100	0.0734 (0.001)	0.0615 (0.003)	*0.0318* (0.000)	0.1	0.0612 (0.001)	5	0.0496 (0.001)
						20	0.0491 (0.001)
						50	0.0530 (0.001)
				0.3	0.0752 (0.001)		
				0.5	0.0734 (0.001)	5	0.0507 (0.001)
						20	0.0502 (0.001)
						50	0.0528 (0.001)
				0.7	0.0584 (0.001)		
				0.9	0.0355 (0.000)	5	0.0356 (0.000)
						20	0.0351 (0.000)
						50	0.0370 (0.001)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.0635 (0.001)	0.0599 (0.002)	*0.0302* (0.000)	0.1	0.0539 (0.001)	5	0.0463 (0.001)
						20	0.0460 (0.001)
						50	0.0503 (0.001)
				0.3	0.0648 (0.001)		
				0.5	0.0635 (0.001)	5	0.0464 (0.000)
						20	0.0460 (0.000)
						50	0.0484 (0.001)
				0.7	0.0516 (0.001)		
				0.9	0.0333 (0.000)	5	0.0334 (0.000)
						20	0.0329 (0.000)
						50	0.0346 (0.001)
300	**0.0590** (0.001)	0.0594 (0.002)	***0.0294*** (0.000)	0.1	0.0507 (0.001)	5	0.0449 (0.001)
						20	0.0444 (0.001)
						50	0.0490 (0.002)
				0.3	0.0601 (0.001)		
				0.5	0.0590 (0.001)	5	0.0445 (0.000)
						20	0.0440 (0.000)
						50	0.0463 (0.001)
				0.7	0.0487 (0.000)		
				0.9	**0.0324** (0.000)	5	0.0325 (0.000)
						20	**0.0318** (0.000)
						50	0.0335 (0.001)

Tabelle C.8: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ ($p = 10$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.0004 (0.000)	0.0152 (0.002)	**0.0003** (0.000)	0.1	0.0091 (0.000)	5	0.0017 (0.000)
						20	0.0027 (0.000)
						50	0.0023 (0.000)
				0.3	0.0008 (0.000)		
				0.5	0.0004 (0.000)	5	0.0003 (0.000)
						20	0.0004 (0.000)
						50	0.0004 (0.000)
				0.7	0.0002 (0.000)		
				0.9	*0.0001* (0.000)	5	*0.0001* (0.000)
						20	*0.0001* (0.000)
						50	*0.0001* (0.000)
10	0.0007 (0.000)	0.0183 (0.002)	0.0010 (0.000)	0.1	0.0011 (0.000)	5	0.0013 (0.000)
						20	0.0012 (0.000)
						50	0.0012 (0.000)
				0.3	0.0008 (0.000)		
				0.5	0.0007 (0.000)	5	0.0007 (0.000)
						20	0.0007 (0.000)
						50	0.0007 (0.000)
				0.7	0.0005 (0.000)		
				0.9	*0.0002* (0.000)	5	*0.0002* (0.000)
						20	*0.0002* (0.000)
						50	*0.0002* (0.000)
20	0.0002 (0.000)	0.0183 (0.002)	0.0004 (0.000)	0.1	0.0004 (0.000)	5	0.0005 (0.000)
						20	0.0004 (0.000)
						50	0.0004 (0.000)
				0.3	0.0003 (0.000)		
				0.5	0.0002 (0.000)	5	0.0002 (0.000)
						20	0.0002 (0.000)
						50	0.0002 (0.000)
				0.7	0.0002 (0.000)		
				0.9	*0.0001* (0.000)	5	*0.0001* (0.000)
						20	*0.0001* (0.000)
						50	*0.0001* (0.000)
50	0.0002 (0.000)	0.0181 (0.002)	**0.0003** (0.000)	0.1	0.0002 (0.000)	5	0.0003 (0.000)
						20	0.0003 (0.000)
						50	0.0003 (0.000)
				0.3	0.0002 (0.000)		
				0.5	0.0002 (0.000)	5	0.0002 (0.000)
						20	0.0002 (0.000)
						50	0.0002 (0.000)
				0.7	*0.0001* (0.000)		
				0.9	*0.0001* (0.000)	5	*0.0001* (0.000)
						20	*0.0001* (0.000)
						50	*0.0001* (0.000)
100	***0.0001*** (0.000)	0.0167 (0.002)	**0.0003** (0.000)	0.1	0.0002 (0.000)	5	0.0003 (0.000)
						20	0.0003 (0.000)
						50	0.0003 (0.000)
				0.3	*0.0001* (0.000)		
				0.5	*0.0001* (0.000)	5	*0.0001* (0.000)
						20	*0.0001* (0.000)
						50	*0.0001* (0.000)
				0.7	*0.0001* (0.000)		
				0.9	*0.0001* (0.000)	5	*0.0001* (0.000)
						20	*0.0001* (0.000)
						50	*0.0001* (0.000)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	***0.0001*** (0.000)	0.0155 (0.002)	**0.0003** (0.000)	0.1	0.0002 (0.000)	5	0.0003 (0.000)
						20	0.0002 (0.000)
						50	0.0002 (0.000)
				0.3	*0.0001* (0.000)		
				0.5	*0.0001* (0.000)	5	*0.0001* (0.000)
						20	*0.0001* (0.000)
						50	*0.0001* (0.000)
				0.7	*0.0001* (0.000)		
				0.9	*0.0001* (0.000)	5	*0.0001* (0.000)
						20	*0.0001* (0.000)
						50	*0.0001* (0.000)
300	**0.0001** (0.000)	**0.0152** (0.002)	**0.0003** (0.000)	0.1	0.0002 (0.000)	5	0.0003 (0.000)
						20	0.0002 (0.000)
						50	0.0002 (0.000)
				0.3	0.0001 (0.000)		
				0.5	0.0001 (0.000)	5	0.0001 (0.000)
						20	0.0001 (0.000)
						50	0.0001 (0.000)
				0.7	0.0001 (0.000)		
				0.9	***0.0000*** (0.000)	5	***0.0000*** (0.000)
						20	***0.0000*** (0.000)
						50	***0.0000*** (0.000)

Tabelle C.9: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittl. Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **Gradual Drift mit „Kreuzen“ ($p = 10$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4826 (0.003)	***0.0018*** (0.000)	0.2863 (0.005)	0.1	0.5002 (0.002)	5	0.2868 (0.007)
						20	0.2821 (0.006)
						50	**0.2734** (0.007)
				0.3	0.5020 (0.002)		
				0.5	0.4826 (0.003)	5	0.2872 (0.007)
						20	0.2831 (0.007)
						50	0.2759 (0.008)
				0.7	0.2995 (0.002)		
				0.9	0.2560 (0.002)	5	0.2874 (0.007)
						20	0.2844 (0.007)
						50	0.2780 (0.008)
10	0.3414 (0.007)	***0.0018*** (0.000)	0.3486 (0.008)	0.1	0.2690 (0.011)	5	0.3085 (0.010)
						20	0.3092 (0.009)
						50	0.3246 (0.010)
				0.3	0.3183 (0.008)		
				0.5	0.3414 (0.007)	5	0.3093 (0.010)
						20	0.3103 (0.010)
						50	0.3265 (0.010)
				0.7	0.3483 (0.006)		
				0.9	0.3370 (0.007)	5	0.3100 (0.010)
						20	0.3109 (0.010)
						50	0.3255 (0.010)
20	0.3057 (0.005)	***0.0018*** (0.000)	0.3238 (0.006)	0.1	0.2511 (0.006)	5	0.3040 (0.009)
						20	0.3016 (0.009)
						50	0.3090 (0.008)
				0.3	0.2854 (0.005)		
				0.5	0.3057 (0.005)	5	0.3049 (0.009)
						20	0.3030 (0.009)
						50	0.3113 (0.009)
				0.7	0.3112 (0.005)		
				0.9	0.3039 (0.005)	5	0.3057 (0.010)
						20	0.3049 (0.009)
						50	0.3115 (0.009)
50	0.2755 (0.003)	***0.0018*** (0.000)	0.3018 (0.005)	0.1	**0.2452** (0.004)	5	0.2998 (0.008)
						20	0.2957 (0.007)
						50	0.3026 (0.008)
				0.3	0.2617 (0.003)		
				0.5	0.2755 (0.003)	5	0.3007 (0.008)
						20	0.2970 (0.008)
						50	0.3049 (0.008)
				0.7	0.2803 (0.003)		
				0.9	0.2797 (0.003)	5	0.3014 (0.008)
						20	0.2990 (0.008)
						50	0.3051 (0.008)
100	0.2626 (0.002)	***0.0018*** (0.000)	0.2874 (0.004)	0.1	0.2460 (0.003)	5	0.2908 (0.006)
						20	0.2879 (0.006)
						50	0.2916 (0.007)
				0.3	0.2541 (0.002)		
				0.5	0.2626 (0.002)	5	0.2915 (0.006)
						20	0.2890 (0.006)
						50	0.2939 (0.008)
				0.7	0.2670 (0.002)		
				0.9	0.2695 (0.003)	5	0.2918 (0.006)
						20	0.2902 (0.006)
						50	0.2943 (0.007)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.2562 (0.002)	***0.0018*** (0.000)	0.2765 (0.004)	0.1	0.2472 (0.002)	5	0.2810 (0.005)
						20	0.2788 (0.005)
						50	0.2856 (0.007)
				0.3	0.2515 (0.002)		
				0.5	0.2562 (0.002)	5	0.2814 (0.005)
						20	0.2797 (0.005)
						50	0.2880 (0.007)
				0.7	0.2597 (0.002)		
				0.9	0.2635 (0.002)	5	0.2815 (0.005)
						20	0.2802 (0.005)
						50	0.2877 (0.007)
300	**0.2542** (0.002)	***0.0018*** (0.000)	**0.2719** (0.003)	0.1	0.2477 (0.002)	5	0.2759 (0.004)
						20	0.2739 (0.004)
						50	0.2825 (0.006)
				0.3	0.2507 (0.002)		
				0.5	0.2542 (0.002)	5	0.2763 (0.004)
						20	0.2748 (0.004)
						50	0.2852 (0.007)
				0.7	0.2572 (0.002)		
				0.9	0.2611 (0.002)	5	0.2763 (0.004)
						20	0.2752 (0.004)
						50	0.2845 (0.006)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: < 0.0001 (Standardabw. < 0.001)

Tabelle C.10: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittl. Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **Gradual Drift mit „Austausch“ ($p = 10$)** getrennt nach Methoden für Online DA und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4792 (0.003)	**0.3598** (0.004)	0.3681 (0.004)	0.1	0.5037 (0.002)	5	0.3824 (0.004)
						20	0.3774 (0.004)
						50	0.3767 (0.004)
				0.3	0.5027 (0.002)		
				0.5	0.4792 (0.003)	5	0.3825 (0.004)
						20	0.3790 (0.004)
						50	0.3785 (0.004)
				0.7	0.3404 (0.002)		
				0.9	*0.3181* (0.002)	5	0.3853 (0.004)
						20	0.3833 (0.004)
						50	0.3810 (0.004)
10	0.4208 (0.005)	0.4211 (0.005)	0.4333 (0.006)	0.1	*0.4072* (0.006)	5	0.4157 (0.005)
						20	0.4165 (0.005)
						50	0.4240 (0.006)
				0.3	0.4152 (0.005)		
				0.5	0.4208 (0.005)	5	0.4156 (0.005)
						20	0.4162 (0.005)
						50	0.4243 (0.005)
				0.7	0.4224 (0.005)		
				0.9	0.4191 (0.005)	5	0.4171 (0.005)
						20	0.4177 (0.005)
						50	0.4240 (0.005)
20	0.3873 (0.004)	0.3969 (0.004)	0.4100 (0.005)	0.1	*0.3719* (0.004)	5	0.4083 (0.005)
						20	0.4061 (0.005)
						50	0.4108 (0.005)
				0.3	0.3801 (0.004)		
				0.5	0.3873 (0.004)	5	0.4084 (0.005)
						20	0.4070 (0.005)
						50	0.4122 (0.005)
				0.7	0.3900 (0.004)		
				0.9	0.3899 (0.004)	5	0.4114 (0.005)
						20	0.4110 (0.005)
						50	0.4142 (0.005)
50	0.3550 (0.003)	0.3796 (0.004)	0.3857 (0.004)	0.1	*0.3420* (0.003)	5	0.3981 (0.004)
						20	0.3954 (0.004)
						50	0.4039 (0.005)
				0.3	0.3476 (0.003)		
				0.5	0.3550 (0.003)	5	0.3983 (0.004)
						20	0.3966 (0.004)
						50	0.4060 (0.005)
				0.7	0.3589 (0.003)		
				0.9	0.3636 (0.003)	5	0.4013 (0.005)
						20	0.4006 (0.005)
						50	0.4063 (0.005)
100	0.3366 (0.003)	0.3717 (0.004)	0.3679 (0.004)	0.1	*0.3275* (0.003)	5	0.3839 (0.004)
						20	0.3832 (0.004)
						50	0.3935 (0.005)
				0.3	0.3301 (0.003)		
				0.5	0.3366 (0.003)	5	0.3840 (0.004)
						20	0.3838 (0.004)
						50	0.3945 (0.005)
				0.7	0.3411 (0.003)		
				0.9	0.3488 (0.003)	5	0.3866 (0.004)
						20	0.3867 (0.004)
						50	0.3942 (0.005)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.3230 (0.002)	0.3662 (0.004)	0.3525 (0.003)	0.1	0.3186 (0.002)	5	0.3695 (0.003)
						20	0.3693 (0.003)
						50	0.3862 (0.004)
				0.3	*0.3179* (0.002)		
				0.5	0.3230 (0.002)	5	0.3695 (0.003)
						20	0.3694 (0.003)
						50	0.3857 (0.004)
				0.7	0.3279 (0.002)		
				0.9	0.3376 (0.003)	5	0.3721 (0.003)
						20	0.3722 (0.004)
						50	0.3852 (0.004)
300	**0.3173** (0.002)	0.3639 (0.004)	**0.3456** (0.003)	0.1	0.3157 (0.002)	5	0.3627 (0.003)
						20	**0.3619** (0.003)
						50	0.3803 (0.004)
				0.3	***0.3130*** (0.002)		
				0.5	0.3173 (0.002)	5	0.3628 (0.003)
						20	0.3623 (0.003)
						50	0.3796 (0.004)
				0.7	0.3222 (0.002)		
				0.9	0.3324 (0.003)	5	0.3655 (0.003)
						20	0.3654 (0.003)
						50	0.3796 (0.004)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.2893 (Standardabweichung 0.122)

Tabelle C.11: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **Sudden Drift** ($p = 10$) getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	0.4782 (0.002)	0.1939 (0.005)	0.1254 (0.004)	0.1	0.5063 (0.005)	5	0.1439 (0.002)
						20	***0.1234*** (0.002)
						50	0.1265 (0.002)
				0.3	0.5028 (0.002)		
				0.5	0.4782 (0.002)	5	0.1437 (0.002)
						20	0.1248 (0.002)
						50	0.1253 (0.003)
				0.7	0.3184 (0.002)		
				0.9	0.1699 (0.001)	5	0.1466 (0.002)
						20	0.1315 (0.002)
						50	0.1327 (0.003)
10	0.2295 (0.007)	0.2217 (0.006)	0.2316 (0.010)	0.1	0.2085 (0.007)	5	0.2007 (0.006)
						20	0.1924 (0.006)
						50	0.2021 (0.007)
				0.3	0.2126 (0.007)		
				0.5	0.2295 (0.007)	5	0.2002 (0.006)
						20	*0.1913* (0.006)
						50	0.1992 (0.007)
				0.7	0.2420 (0.007)		
				0.9	0.2216 (0.007)	5	0.2019 (0.006)
						20	0.1938 (0.006)
						50	0.1989 (0.007)
20	0.1612 (0.003)	0.1949 (0.004)	0.1724 (0.005)	0.1	0.1504 (0.003)	5	0.1787 (0.004)
						20	0.1571 (0.003)
						50	0.1605 (0.004)
				0.3	*0.1480* (0.002)		
				0.5	0.1612 (0.003)	5	0.1784 (0.004)
						20	0.1581 (0.003)
						50	0.1594 (0.004)
				0.7	0.1722 (0.003)		
				0.9	0.1603 (0.003)	5	0.1817 (0.004)
						20	0.1652 (0.004)
						50	0.1674 (0.004)
50	0.1264 (0.001)	0.1855 (0.004)	0.1366 (0.002)	0.1	0.1282 (0.002)	5	0.1633 (0.003)
						20	0.1425 (0.003)
						50	0.1497 (0.004)
				0.3	*0.1197* (0.001)		
				0.5	0.1264 (0.001)	5	0.1632 (0.003)
						20	0.1439 (0.002)
						50	0.1490 (0.004)
				0.7	0.1330 (0.002)		
				0.9	0.1299 (0.002)	5	0.1666 (0.003)
						20	0.1514 (0.003)
						50	0.1558 (0.004)
100	**0.1220** (0.001)	**0.1831** (0.004)	0.1237 (0.002)	0.1	0.1307 (0.001)	5	0.1504 (0.002)
						20	0.1353 (0.002)
						50	0.1451 (0.004)
				0.3	***0.1193*** (0.001)		
				0.5	0.1220 (0.001)	5	0.1503 (0.002)
						20	0.1361 (0.002)
						50	0.1438 (0.004)
				0.7	0.1257 (0.001)		
				0.9	0.1251 (0.001)	5	0.1533 (0.003)
						20	0.1420 (0.002)
						50	0.1471 (0.003)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.1356 (0.001)	0.1850 (0.004)	***0.1221*** (0.002)	0.1	0.1498 (0.001)	5	0.1447 (0.002)
						20	0.1313 (0.002)
						50	0.1415 (0.003)
				0.3	0.1362 (0.001)		
				0.5	0.1356 (0.001)	5	0.1445 (0.002)
						20	0.1315 (0.002)
						50	0.1390 (0.003)
				0.7	0.1362 (0.001)		
				0.9	0.1328 (0.001)	5	0.1473 (0.002)
						20	0.1365 (0.002)
						50	0.1415 (0.003)
300	0.1542 (0.001)	0.1885 (0.004)	*0.1267* (0.002)	0.1	0.1719 (0.001)	5	0.1471 (0.002)
						20	0.1326 (0.002)
						50	0.1420 (0.003)
				0.3	0.1569 (0.001)		
				0.5	0.1542 (0.001)	5	0.1468 (0.002)
						20	0.1331 (0.002)
						50	0.1395 (0.003)
				0.7	0.1524 (0.001)		
				0.9	0.1436 (0.001)	5	0.1497 (0.002)
						20	0.1384 (0.002)
						50	0.1422 (0.003)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0786 (Standardabweichung 0)

Tabelle C.12: Durchschnittlicher mittlerer Prognosefehler über die Zeit (durchschnittliche Varianz des Prognosefehlers über die Zeit in Klammern) für die Datensituation **ohne Drift ($p = 10$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	**0.0832** (0.001)	**0.1143** (0.003)	0.1117 (0.002)	0.1	0.1088 (0.001)	5	0.1364 (0.002)
						20	0.1141 (0.002)
						50	**0.1068** (0.002)
				0.3	0.0859 (0.001)		
				0.5	***0.0831*** (0.001)	5	0.1362 (0.002)
						20	0.1175 (0.002)
						50	0.1122 (0.002)
				0.7	0.0843 (0.001)		
				0.9	0.0905 (0.001)	5	0.1395 (0.002)
						20	0.1249 (0.002)
						50	0.1256 (0.002)
10	*0.1779* (0.006)	0.1985 (0.007)	0.2242 (0.010)	0.1	0.1913 (0.006)	5	0.1956 (0.006)
						20	0.1868 (0.006)
						50	0.1885 (0.006)
				0.3	0.1789 (0.006)		
				0.5	*0.1779* (0.006)	5	0.1951 (0.006)
						20	0.1868 (0.006)
						50	0.1891 (0.006)
				0.7	*0.1779* (0.006)		
				0.9	0.1792 (0.006)	5	0.1971 (0.006)
						20	0.1898 (0.006)
						50	0.1922 (0.006)
20	*0.1218* (0.002)	0.1498 (0.004)	0.1649 (0.004)	0.1	0.1373 (0.002)	5	0.1727 (0.004)
						20	0.1499 (0.003)
						50	0.1443 (0.003)
				0.3	0.1230 (0.002)		
				0.5	*0.1218* (0.002)	5	0.1723 (0.004)
						20	0.1529 (0.003)
						50	0.1496 (0.003)
				0.7	0.1227 (0.002)		
				0.9	0.1279 (0.002)	5	0.1759 (0.004)
						20	0.1606 (0.003)
						50	0.1621 (0.004)
50	*0.0984* (0.001)	0.1293 (0.004)	0.1273 (0.002)	0.1	0.1136 (0.001)	5	0.1563 (0.003)
						20	0.1332 (0.002)
						50	0.1282 (0.003)
				0.3	0.0993 (0.001)		
				0.5	*0.0984* (0.001)	5	0.1562 (0.003)
						20	0.1371 (0.002)
						50	0.1346 (0.003)
				0.7	0.0997 (0.001)		
				0.9	0.1064 (0.001)	5	0.1597 (0.003)
						20	0.1454 (0.003)
						50	0.1471 (0.003)
100	*0.0916* (0.001)	0.1222 (0.004)	0.1108 (0.001)	0.1	0.1065 (0.001)	5	0.1413 (0.002)
						20	0.1242 (0.002)
						50	0.1212 (0.002)
				0.3	0.0924 (0.001)		
				0.5	*0.0916* (0.001)	5	0.1411 (0.002)
						20	0.1271 (0.002)
						50	0.1260 (0.002)
				0.7	0.0930 (0.001)		
				0.9	0.0994 (0.001)	5	0.1443 (0.002)
						20	0.1336 (0.002)
						50	0.1352 (0.002)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	*0.0884* (0.001)	0.1180 (0.003)	0.1017 (0.001)	0.1	0.1034 (0.001)	5	0.1305 (0.002)
						20	0.1163 (0.001)
						50	0.1152 (0.002)
				0.3	0.0890 (0.001)		
				0.5	*0.0884* (0.001)	5	0.1302 (0.002)
						20	0.1183 (0.002)
						50	0.1182 (0.002)
				0.7	0.0896 (0.001)		
				0.9	0.0955 (0.001)	5	0.1332 (0.002)
						20	0.1237 (0.002)
						50	0.1255 (0.002)
300	*0.0873* (0.001)	0.1163 (0.003)	**0.0987** (0.001)	0.1	0.1027 (0.001)	5	0.1266 (0.002)
						20	0.1131 (0.001)
						50	0.1119 (0.002)
				0.3	0.0880 (0.001)		
				0.5	*0.0873* (0.001)	5	0.1264 (0.002)
						20	0.1145 (0.001)
						50	0.1140 (0.002)
				0.7	0.0886 (0.001)		
				0.9	0.0941 (0.001)	5	0.1294 (0.002)
						20	0.1200 (0.002)
						50	0.1216 (0.002)

Durchschnittlicher Bayesfehler über gesamten Datenstrom: 0.0786 (Standardabweichung 0)

C.2 Euklidischer Abstand

Tabelle C.13: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Kreuzen“ ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	8.6821 (0.019)	0.7202 (3.345)	*0.2256* (0.147)	0.1	15.6553 (0.129)	5	2.6893 (0.510)
	8.6978 (0.019)	0.8223 (3.767)	*0.2263* (0.151)		15.6841 (0.155)		2.7011 (0.523)
						20	2.8076 (0.533)
							2.8176 (0.548)
						50	2.6466 (0.659)
							2.6581 (0.667)
				0.3	12.1631 (0.019)		
					12.1810 (0.020)		
				0.5	8.6821 (0.019)	5	1.4488 (0.188)
					8.6978 (0.019)		1.4554 (0.189)
						20	1.4885 (0.168)
							1.4964 (0.168)
						50	1.4401 (0.242)
							1.4478 (0.238)
				0.7	5.2142 (0.021)		
					5.2223 (0.021)		
				0.9	1.7504 (0.025)	5	0.3220 (0.128)
					1.7514 (0.024)		0.3262 (0.128)
						20	0.3272 (0.105)
							0.3311 (0.105)
						50	0.3499 (0.137)
							0.3533 (0.138)
10	0.1287 (0.400)	0.1637 (0.708)	0.1666 (0.746)	0.1	*0.1281* (0.396)	5	0.1586 (0.683)
	0.1291 (0.397)	0.1672 (0.721)	0.1687 (0.743)		*0.1284* (0.393)		0.1598 (0.690)
						20	0.1393 (0.503)
							0.1395 (0.498)
						50	0.1368 (0.508)
							0.1371 (0.495)
				0.3	0.1283 (0.397)		
					0.1286 (0.394)		
				0.5	0.1287 (0.400)	5	0.1591 (0.689)
					0.1291 (0.397)		0.1610 (0.697)
						20	0.1409 (0.517)
							0.1414 (0.513)
						50	0.1378 (0.505)
							0.1384 (0.498)
				0.7	0.1295 (0.406)		
					0.1301 (0.403)		
				0.9	0.1332 (0.438)	5	0.1637 (0.736)
					0.1342 (0.434)		0.1660 (0.740)
						20	0.1473 (0.574)
							0.1482 (0.572)
						50	0.1428 (0.561)
							0.1447 (0.553)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
20	0.0839 (0.139)	0.1180 (0.295)	0.1199 (0.309)	0.1	*0.0832* (0.136)	5	0.1129 (0.267)
	0.0844 (0.138)	0.1201 (0.304)	0.1215 (0.305)		*0.0836* (0.136)		0.1134 (0.267)
						20	0.0986 (0.242)
							0.0996 (0.238)
						50	0.0967 (0.244)
							0.0960 (0.246)
				0.3	0.0835 (0.137)		
					0.0839 (0.136)		
				0.5	0.0839 (0.139)	5	0.1135 (0.269)
					0.0844 (0.138)		0.1142 (0.269)
						20	0.1011 (0.246)
							0.1010 (0.238)
						50	0.0971 (0.244)
							0.0975 (0.243)
				0.7	0.0850 (0.142)		
					0.0855 (0.142)		
				0.9	0.0892 (0.158)	5	0.1178 (0.289)
					0.0899 (0.158)		0.1187 (0.289)
						20	0.1070 (0.262)
							0.1078 (0.258)
						50	0.1017 (0.259)
							0.1022 (0.255)
50	0.0534 (0.052)	0.0813 (0.135)	0.0831 (0.137)	0.1	*0.0528* (0.051)	5	0.0765 (0.114)
	0.0530 (0.053)	0.0810 (0.142)	0.0815 (0.137)		*0.0524* (0.051)		0.0756 (0.115)
						20	0.0695 (0.103)
							0.0708 (0.104)
						50	0.0793 (0.439)
							0.0786 (0.454)
				0.3	0.0530 (0.051)		
					0.0525 (0.051)		
				0.5	0.0534 (0.052)	5	0.0772 (0.113)
					0.0530 (0.053)		0.0758 (0.115)
						20	0.0724 (0.102)
							0.0704 (0.103)
						50	0.0798 (0.435)
							0.0776 (0.425)
				0.7	0.0544 (0.055)		
					0.0544 (0.055)		
				0.9	0.0593 (0.066)	5	0.0812 (0.125)
					0.0590 (0.066)		0.0801 (0.127)
						20	0.0771 (0.116)
							0.0759 (0.119)
						50	0.0834 (0.390)
							0.0795 (0.362)
100	0.0393 (0.028)	0.0603 (0.074)	0.0634 (0.072)	0.1	0.0393 (0.030)	5	0.0562 (0.062)
	0.0401 (0.029)	0.0593 (0.075)	0.0571 (0.073)		0.0405 (0.029)		0.0548 (0.062)
						20	0.0543 (0.062)
							0.0571 (0.062)
						50	0.0591 (0.239)
							0.0607 (0.241)
				0.3	*0.0390* (0.028)		
					0.0399 (0.028)		
				0.5	0.0393 (0.028)	5	0.0563 (0.060)
					0.0401 (0.029)		0.0539 (0.061)
						20	0.0568 (0.059)
							0.0548 (0.060)
						50	0.0601 (0.253)
							0.0600 (0.242)
				0.7	0.0405 (0.030)		
					0.0412 (0.031)		
				0.9	0.0459 (0.039)	5	0.0612 (0.067)
					0.0451 (0.039)		0.0563 (0.068)
						20	0.0616 (0.068)
							0.0571 (0.068)
						50	0.0660 (0.234)
							0.0639 (0.232)

Fortsetzung auf der nächsten Seite

	ILDA	QDA-AF	LDA-AF	OLDC fix		OLDC adaptive	
N_{trend}				λ,λ_{start}		L	
200	*0.0293* (0.016)	0.0439 (0.038)	0.0453 (0.037)	0.1	0.0317 (0.021)	5	0.0438 (0.037)
	0.0331 (0.016)	0.0459 (0.038)	0.0441 (0.037)		0.0363 (0.020)		0.0477 (0.036)
						20	0.0441 (0.042)
							0.0524 (0.040)
						50	0.0559 (0.244)
							0.0610 (0.266)
				0.3	0.0297 (0.017)		
					0.0334 (0.017)		
				0.5	*0.0293* (0.016)	5	0.0413 (0.032)
					0.0331 (0.016)		0.0442 (0.031)
						20	0.0436 (0.036)
							0.0484 (0.035)
						50	0.0513 (0.225)
							0.0585 (0.244)
				0.7	0.0304 (0.018)		
					0.0344 (0.018)		
				0.9	0.0355 (0.024)	5	0.0444 (0.036)
					0.0379 (0.024)		0.0453 (0.036)
						20	0.0464 (0.040)
							0.0486 (0.039)
						50	0.0554 (0.221)
							0.0571 (0.241)
300	***0.0254*** (0.012)	**0.0366** (0.026)	**0.0365** (0.025)	0.1	0.0301 (0.019)	5	0.0391 (0.031)
	0.0302 (0.012)	**0.0408** (0.025)	**0.0395** (0.024)		0.0365 (0.018)		0.0449 (0.029)
						20	0.0410 (0.035)
							0.0516 (0.033)
						50	0.0614 (0.338)
							0.0643 (0.387)
				0.3	0.0264 (0.013)		
					0.0309 (0.013)		
				0.5	***0.0254*** (0.012)	5	**0.0348** (0.022)
					0.0302 (0.012)		**0.0389** (0.022)
						20	0.0375 (0.026)
							0.0443 (0.025)
						50	0.0515 (0.301)
							0.0619 (0.270)
				0.7	0.0266 (0.014)		
					0.0321 (0.013)		
				0.9	0.0314 (0.018)	5	0.0374 (0.025)
					0.0361 (0.018)		0.0407 (0.024)
						20	0.0405 (0.028)
							0.0459 (0.027)
						50	0.0553 (0.269)
							0.0609 (0.295)

Tabelle C.14: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	8.6821 (0.019)	0.7202 (3.345)	*0.2294* (0.149)	0.1	15.6553 (0.129)	5	5.1743 (1.020)
	8.6943 (0.020)	0.7776 (3.688)	*0.2412* (0.167)		15.5952 (0.151)		5.1325 (1.008)
						20	7.3005 (1.471)
							7.2419 (1.412)
						50	7.4826 (1.945)
							7.4317 (1.812)
				0.3	12.1631 (0.019)		
					12.1565 (0.023)		
				0.5	8.6821 (0.019)	5	2.6524 (0.353)
					8.6943 (0.020)		2.6634 (0.365)
						20	4.0121 (0.393)
							4.0122 (0.399)
						50	4.5677 (0.447)
							4.5708 (0.428)
				0.7	5.2142 (0.021)		
					5.2257 (0.022)		
				0.9	1.7504 (0.025)	5	0.5958 (0.045)
					1.7590 (0.024)		0.6092 (0.046)
						20	1.0345 (0.059)
							1.0478 (0.059)
						50	1.2565 (0.059)
							1.2666 (0.059)
10	0.1287 (0.400)	0.1637 (0.708)	0.1666 (0.746)	0.1	*0.1281* (0.396)	5	0.1328 (0.426)
	0.1291 (0.397)	0.1682 (0.727)	0.1690 (0.742)		*0.1284* (0.393)		0.1329 (0.424)
						20	0.1289 (0.399)
							0.1287 (0.396)
						50	0.1288 (0.399)
							0.1286 (0.396)
				0.3	0.1283 (0.397)		
					0.1286 (0.394)		
				0.5	0.1287 (0.400)	5	0.1342 (0.437)
					0.1291 (0.397)		0.1346 (0.437)
						20	0.1293 (0.404)
							0.1295 (0.401)
						50	0.1294 (0.404)
							0.1297 (0.402)
				0.7	0.1295 (0.406)		
					0.1301 (0.403)		
				0.9	0.1332 (0.438)	5	0.1410 (0.495)
					0.1342 (0.434)		0.1421 (0.495)
						20	0.1347 (0.447)
							0.1355 (0.444)
						50	0.1348 (0.449)
							0.1358 (0.448)

Fortsetzung auf der nächsten Seite

	ILDA	QDA-AF	LDA-AF	OLDC fix		OLDC adaptive	
N_{trend}				λ,λ_{start}		L	
20	0.0839 (0.139)	0.1180 (0.295)	0.1199 (0.309)	0.1	*0.0832* (0.136)	5	0.0900 (0.159)
	0.0844 (0.138)	0.1208 (0.306)	0.1218 (0.306)		*0.0836* (0.136)		0.0894 (0.158)
						20	0.0851 (0.141)
							0.0844 (0.140)
						50	0.0842 (0.139)
							0.0841 (0.139)
				0.3	0.0835 (0.137)		
					0.0839 (0.136)		
				0.5	0.0839 (0.139)	5	0.0911 (0.165)
					0.0844 (0.138)		0.0911 (0.165)
						20	0.0850 (0.143)
							0.0852 (0.142)
						50	0.0850 (0.142)
							0.0854 (0.142)
				0.7	0.0850 (0.142)		
					0.0855 (0.142)		
				0.9	0.0892 (0.158)	5	0.0987 (0.195)
					0.0899 (0.158)		0.0988 (0.195)
						20	0.0910 (0.165)
							0.0917 (0.165)
						50	0.0913 (0.167)
							0.0919 (0.167)
50	0.0534 (0.052)	0.0813 (0.135)	0.0830 (0.137)	0.1	*0.0528* (0.051)	5	0.0606 (0.069)
	0.0530 (0.053)	0.0806 (0.141)	0.0815 (0.138)		*0.0524* (0.051)		0.0598 (0.070)
						20	0.0578 (0.059)
							0.0551 (0.058)
						50	0.0553 (0.057)
							0.0538 (0.057)
				0.3	0.0530 (0.051)		
					0.0525 (0.051)		
				0.5	0.0534 (0.052)	5	0.0625 (0.072)
					0.0530 (0.053)		0.0607 (0.074)
						20	0.0565 (0.058)
							0.0550 (0.058)
						50	0.0555 (0.058)
							0.0550 (0.059)
				0.7	0.0544 (0.055)		
					0.0544 (0.055)		
				0.9	0.0593 (0.066)	5	0.0701 (0.090)
					0.0590 (0.066)		0.0680 (0.092)
						20	0.0625 (0.073)
							0.0620 (0.074)
						50	0.0624 (0.076)
							0.0614 (0.076)
100	0.0393 (0.028)	0.0603 (0.074)	0.0633 (0.072)	0.1	0.0393 (0.030)	5	0.0467 (0.043)
	0.0401 (0.029)	0.0589 (0.075)	0.0571 (0.073)		0.0405 (0.029)		0.0467 (0.044)
						20	0.0446 (0.038)
							0.0445 (0.038)
						50	0.0444 (0.039)
							0.0432 (0.038)
				0.3	*0.0390* (0.028)		
					0.0399 (0.028)		
				0.5	0.0393 (0.028)	5	0.0479 (0.043)
					0.0401 (0.029)		0.0468 (0.044)
						20	0.0435 (0.034)
							0.0434 (0.035)
						50	0.0446 (0.038)
							0.0426 (0.038)
				0.7	0.0405 (0.030)		
					0.0412 (0.031)		
				0.9	0.0459 (0.039)	5	0.0545 (0.055)
					0.0451 (0.039)		0.0520 (0.056)
						20	0.0494 (0.045)
							0.0494 (0.046)
						50	0.0506 (0.048)
							0.0484 (0.048)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	*0.0293* (0.016)	0.0439 (0.038)	0.0453 (0.037)	0.1	0.0317 (0.021)	5	0.0378 (0.030)
	0.0331 (0.016)	0.0448 (0.038)	0.0441 (0.037)		0.0363 (0.020)		0.0431 (0.029)
						20	0.0371 (0.029)
							0.0408 (0.029)
						50	0.0385 (0.035)
							0.0394 (0.033)
				0.3	0.0297 (0.017)		
					0.0334 (0.017)		
				0.5	*0.0293* (0.016)	5	0.0360 (0.026)
					0.0331 (0.016)		0.0405 (0.026)
						20	0.0340 (0.022)
							0.0374 (0.023)
						50	0.0357 (0.025)
							0.0371 (0.025)
				0.7	0.0304 (0.018)		
					0.0344 (0.018)		
				0.9	0.0355 (0.024)	5	0.0413 (0.032)
					0.0379 (0.024)		0.0448 (0.033)
						20	0.0381 (0.029)
							0.0434 (0.029)
						50	0.0404 (0.030)
							0.0426 (0.029)
300	***0.0254*** (0.012)	**0.0366** (0.026)	**0.0365** (0.025)	0.1	0.0301 (0.019)	5	0.0354 (0.026)
	0.0302 (0.012)	**0.0403** (0.025)	**0.0395** (0.024)		0.0365 (0.018)		0.0427 (0.025)
						20	0.0360 (0.026)
							0.0409 (0.026)
						50	0.0364 (0.033)
							0.0413 (0.034)
				0.3	0.0264 (0.013)		
					0.0309 (0.013)		
				0.5	***0.0254*** (0.012)	5	0.0312 (0.019)
					0.0302 (0.012)		0.0364 (0.019)
						20	**0.0300** (0.017)
							0.0343 (0.018)
						50	0.0324 (0.022)
							0.0349 (0.021)
				0.7	0.0266 (0.014)		
					0.0321 (0.013)		
				0.9	0.0314 (0.018)	5	0.0357 (0.024)
					0.0361 (0.018)		0.0415 (0.023)
						20	0.0339 (0.022)
							0.0418 (0.022)
						50	0.0364 (0.024)
							0.0401 (0.022)

Tabelle C.15: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **Sudden Drift ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	1.6816 (0.004)	**0.1233** (0.140)	**0.1295** (0.156)	0.1	2.2612 (0.105)	5	0.1144 (0.190)
	1.6819 (0.004)	**0.1259** (0.141)	**0.1327** (0.160)		2.2703 (0.105)		0.1138 (0.192)
						20	0.1063 (0.150)
							0.1066 (0.149)
						50	0.1226 (0.157)
							0.1220 (0.157)
				0.3	1.9479 (0.009)		
					1.9436 (0.009)		
				0.5	1.6816 (0.004)	5	0.1144 (0.191)
					1.6819 (0.004)		0.1133 (0.192)
						20	***0.1019*** (0.154)
							0.1025 (0.151)
						50	0.1116 (0.176)
							0.1112 (0.174)
				0.7	1.4098 (0.007)		
					1.4128 (0.006)		
				0.9	0.6410 (0.017)	5	0.1160 (0.196)
					0.6461 (0.016)		0.1149 (0.198)
						20	0.1067 (0.168)
							0.1062 (0.166)
						50	0.1100 (0.217)
							0.1094 (0.216)
10	0.1413 (0.402)	0.1782 (0.730)	0.1797 (0.740)	0.1	*0.1405* (0.398)	5	0.2010 (0.970)
	0.1403 (0.403)	0.1780 (0.751)	0.1784 (0.787)		*0.1395* (0.399)		0.1995 (0.987)
						20	0.1664 (0.626)
							0.1669 (0.633)
						50	0.1514 (0.499)
							0.1507 (0.511)
				0.3	0.1408 (0.400)		
					0.1398 (0.400)		
				0.5	0.1413 (0.402)	5	0.2016 (0.981)
					0.1403 (0.403)		0.2004 (1.001)
						20	0.1731 (0.689)
							0.1725 (0.690)
						50	0.1533 (0.518)
							0.1526 (0.537)
				0.7	0.1426 (0.409)		
					0.1415 (0.410)		
				0.9	0.1477 (0.443)	5	0.2036 (1.016)
					0.1464 (0.447)		0.2022 (1.041)
						20	0.1855 (0.818)
							0.1856 (0.852)
						50	0.1720 (0.710)
							0.1707 (0.760)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
20	**0.1114** (0.141)	0.1462 (0.301)	0.1467 (0.307)	0.1	***0.1095*** (0.139)	5	0.1617 (0.375)
	0.1099 (0.140)	0.1443 (0.303)	0.1437 (0.313)		***0.1082*** (0.138)		0.1592 (0.380)
						20	0.1369 (0.309)
							0.1397 (0.308)
						50	0.1251 (0.255)
							0.1240 (0.262)
				0.3	0.1102 (0.140)		
					0.1087 (0.139)		
				0.5	0.1114 (0.141)	5	0.1623 (0.377)
					0.1099 (0.140)		0.1594 (0.382)
						20	0.1422 (0.324)
							0.1433 (0.326)
						50	0.1261 (0.262)
							0.1253 (0.270)
				0.7	0.1141 (0.145)		
					0.1126 (0.144)		
				0.9	0.1225 (0.161)	5	0.1623 (0.383)
					0.1209 (0.160)		0.1601 (0.389)
						20	0.1504 (0.356)
							0.1509 (0.359)
						50	0.1403 (0.323)
							0.1396 (0.326)
50	0.1348 (0.057)	0.1474 (0.146)	0.1458 (0.147)	0.1	*0.1281* (0.055)	5	0.1480 (0.164)
	0.1387 (0.056)	0.1501 (0.146)	0.1482 (0.149)		*0.1319* (0.054)		0.1492 (0.166)
						20	0.1505 (0.212)
							0.1603 (0.212)
						50	0.1458 (0.425)
							0.1506 (0.465)
				0.3	0.1304 (0.056)		
					0.1342 (0.054)		
				0.5	0.1348 (0.057)	5	0.1488 (0.165)
					0.1387 (0.056)		0.1493 (0.166)
						20	0.1527 (0.188)
							0.1606 (0.191)
						50	0.1457 (0.416)
							0.1515 (0.455)
				0.7	0.1446 (0.061)		
					0.1484 (0.059)		
				0.9	0.1609 (0.073)	5	0.1482 (0.166)
					0.1634 (0.071)		0.1491 (0.167)
						20	0.1508 (0.173)
							0.1568 (0.179)
						50	0.1508 (0.321)
							0.1525 (0.335)
100	0.2253 (0.035)	0.1803 (0.091)	0.1795 (0.089)	0.1	0.2111 (0.034)	5	*0.1651* (0.088)
	0.2294 (0.033)	0.1865 (0.086)	0.1843 (0.089)		0.2153 (0.032)		0.1721 (0.087)
						20	0.1973 (0.151)
							0.2103 (0.152)
						50	0.2003 (0.908)
							0.2006 (0.966)
				0.3	0.2159 (0.034)		
					0.2196 (0.032)		
				0.5	0.2253 (0.035)	5	0.1653 (0.087)
					0.2294 (0.033)		*0.1718* (0.087)
						20	0.1995 (0.129)
							0.2090 (0.129)
						50	0.2034 (0.796)
							0.2026 (0.905)
				0.7	0.2449 (0.038)		
					0.2502 (0.036)		
				0.9	0.2585 (0.049)	5	0.1652 (0.088)
					0.2638 (0.046)		0.1719 (0.087)
						20	0.1852 (0.105)
							0.1949 (0.107)
						50	0.1986 (0.435)
							0.1970 (0.452)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	0.4272 (0.025)	0.2601 (0.065)	0.2641 (0.063)	0.1	0.4039 (0.025)	5	0.2315 (0.050)
	0.4295 (0.023)	0.2687 (0.061)	0.2688 (0.064)		0.4058 (0.024)		0.2395 (0.047)
						20	0.2956 (0.127)
							0.3042 (0.124)
						50	0.3288 (0.932)
							0.3324 (0.902)
				0.3	0.4117 (0.024)		
					0.4127 (0.022)		
				0.5	0.4272 (0.025)	5	0.2311 (0.049)
					0.4295 (0.023)		*0.2387* (0.047)
						20	0.2882 (0.100)
							0.2951 (0.100)
						50	0.3364 (0.802)
							0.3378 (0.772)
				0.7	0.4574 (0.029)		
					0.4622 (0.027)		
				0.9	0.4494 (0.037)	5	*0.2308* (0.049)
					0.4577 (0.035)		*0.2387* (0.047)
						20	0.2603 (0.065)
							0.2678 (0.063)
						50	0.3350 (0.546)
							0.3350 (0.474)
300	0.6289 (0.021)	0.3535 (0.056)	0.3572 (0.055)	0.1	0.6017 (0.023)	5	0.3139 (0.035)
	0.6304 (0.020)	0.3608 (0.055)	0.3620 (0.058)		0.6045 (0.023)		0.3211 (0.034)
						20	0.3948 (0.122)
							0.4028 (0.120)
						50	0.4987 (0.701)
							0.4973 (0.724)
				0.3	0.6108 (0.020)		
					0.6114 (0.019)		
				0.5	0.6289 (0.021)	5	0.3132 (0.035)
					0.6304 (0.020)		0.3207 (0.034)
						20	0.3771 (0.085)
							0.3857 (0.084)
						50	0.5111 (0.639)
							0.5092 (0.652)
				0.7	0.6631 (0.025)		
					0.6675 (0.024)		
				0.9	0.6291 (0.032)	5	*0.3129* (0.035)
					0.6384 (0.030)		*0.3204* (0.034)
						20	0.3429 (0.045)
							0.3515 (0.046)
						50	0.4845 (0.560)
							0.4855 (0.506)

Tabelle C.16: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **ohne Drift ($p = 3$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	**0.0083** (0.003)	0.0631 (0.120)	0.0649 (0.131)	0.1	0.0542 (0.071)	5	0.0988 (0.184)
	0.0077 (0.003)	0.0651 (0.118)	0.0677 (0.130)		0.0409 (0.123)		0.0981 (0.186)
						20	0.0630 (0.115)
							0.0663 (0.118)
						50	0.0564 (0.069)
							0.0474 (0.074)
				0.3	***0.0054*** (0.007)		
					0.0087 (0.010)		
				0.5	0.0083 (0.003)	5	0.0980 (0.185)
					0.0077 (0.003)		0.1001 (0.186)
						20	0.0736 (0.130)
							0.0758 (0.129)
						50	0.0562 (0.099)
							0.0560 (0.100)
				0.7	0.0104 (0.004)		
					0.0105 (0.004)		
				0.9	0.0190 (0.012)	5	0.1000 (0.190)
					0.0191 (0.012)		0.1016 (0.191)
						20	0.0843 (0.156)
							0.0860 (0.156)
						50	0.0782 (0.181)
							0.0785 (0.184)
10	0.1290 (0.400)	0.1654 (0.725)	0.1664 (0.752)	0.1	*0.1286* (0.397)	5	0.1891 (0.973)
	0.1289 (0.399)	0.1645 (0.724)	0.1664 (0.751)		*0.1284* (0.395)		0.1871 (0.973)
						20	0.1593 (0.650)
							0.1582 (0.649)
						50	0.1402 (0.493)
							0.1406 (0.492)
				0.3	0.1287 (0.398)		
					0.1285 (0.396)		
				0.5	0.1290 (0.400)	5	0.1902 (0.988)
					0.1289 (0.399)		0.1881 (0.991)
						20	0.1658 (0.723)
							0.1647 (0.725)
						50	0.1465 (0.551)
							0.1461 (0.549)
				0.7	0.1299 (0.407)		
					0.1298 (0.406)		
				0.9	0.1337 (0.436)	5	0.1917 (1.009)
					0.1338 (0.439)		0.1900 (1.021)
						20	0.1780 (0.850)
							0.1763 (0.854)
						50	0.1665 (0.768)
							0.1670 (0.784)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
20	0.0811 (0.137)	0.1148 (0.295)	0.1145 (0.303)	0.1	*0.0808* (0.135)	5	0.1322 (0.368)
	0.0851 (0.137)	0.1193 (0.298)	0.1214 (0.307)		*0.0846* (0.135)		0.1371 (0.372)
						20	0.1140 (0.293)
							0.1192 (0.292)
						50	0.0968 (0.220)
							0.1021 (0.218)
				0.3	*0.0808* (0.136)		
					0.0847 (0.136)		
				0.5	0.0811 (0.137)	5	0.1323 (0.370)
					0.0851 (0.137)		0.1377 (0.375)
						20	0.1179 (0.318)
							0.1235 (0.318)
						50	0.1032 (0.255)
							0.1081 (0.251)
				0.7	0.0821 (0.141)		
					0.0860 (0.141)		
				0.9	0.0863 (0.158)	5	0.1325 (0.377)
					0.0904 (0.157)		0.1381 (0.382)
						20	0.1263 (0.357)
							0.1309 (0.355)
						50	0.1178 (0.331)
							0.1242 (0.331)
50	0.0509 (0.052)	0.0805 (0.135)	0.0797 (0.136)	0.1	0.0510 (0.050)	5	0.0893 (0.157)
	0.0537 (0.052)	0.0846 (0.135)	0.0865 (0.138)		*0.0534* (0.051)		0.0952 (0.161)
						20	0.0901 (0.154)
							0.0974 (0.155)
						50	0.0826 (0.203)
							0.0860 (0.194)
				0.3	*0.0507* (0.051)		
					0.0534 (0.051)		
				0.5	0.0509 (0.052)	5	0.0892 (0.155)
					0.0537 (0.052)		0.0951 (0.159)
						20	0.0917 (0.160)
							0.0975 (0.161)
						50	0.0876 (0.240)
							0.0901 (0.232)
				0.7	0.0521 (0.054)		
					0.0550 (0.054)		
				0.9	0.0572 (0.065)	5	0.0888 (0.156)
					0.0603 (0.065)		0.0949 (0.160)
						20	0.0919 (0.161)
							0.0969 (0.163)
						50	0.0915 (0.216)
							0.0944 (0.215)
100	0.0366 (0.028)	0.0594 (0.073)	0.0601 (0.073)	0.1	0.0374 (0.028)	5	0.0638 (0.080)
	0.0370 (0.028)	0.0601 (0.071)	0.0610 (0.073)		0.0371 (0.029)		0.0654 (0.080)
						20	0.0722 (0.095)
							0.0744 (0.099)
						50	0.0829 (0.182)
							0.0850 (0.177)
				0.3	*0.0365* (0.027)		
					0.0367 (0.027)		
				0.5	0.0366 (0.028)	5	0.0635 (0.080)
					0.0370 (0.028)		0.0658 (0.079)
						20	0.0722 (0.093)
							0.0708 (0.093)
						50	0.0814 (0.194)
							0.0846 (0.191)
				0.7	0.0379 (0.031)		
					0.0383 (0.030)		
				0.9	0.0429 (0.039)	5	0.0632 (0.080)
					0.0436 (0.038)		0.0652 (0.079)
						20	0.0692 (0.086)
							0.0687 (0.084)
						50	0.0720 (0.116)
							0.0723 (0.120)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.0272 (0.017)	0.0407 (0.037)	0.0416 (0.037)	0.1	0.0286 (0.019)	5	0.0450 (0.039)
	0.0284 (0.016)	0.0439 (0.036)	0.0431 (0.036)		0.0290 (0.022)		0.0465 (0.039)
						20	0.0532 (0.055)
							0.0588 (0.060)
						50	0.0686 (0.094)
							0.0713 (0.101)
				0.3	*0.0268* (0.016)		
					0.0282 (0.017)		
				0.5	0.0272 (0.017)	5	0.0446 (0.039)
					0.0284 (0.016)		0.0462 (0.038)
						20	0.0519 (0.048)
							0.0541 (0.047)
						50	0.0643 (0.077)
							0.0665 (0.080)
				0.7	0.0285 (0.018)		
					0.0299 (0.017)		
				0.9	0.0328 (0.024)	5	0.0443 (0.039)
					0.0351 (0.023)		0.0458 (0.038)
						20	0.0484 (0.042)
							0.0517 (0.040)
						50	0.0519 (0.051)
							0.0569 (0.053)
300	0.0231 (0.012)	**0.0316** (0.025)	**0.0328** (0.025)	0.1	0.0252 (0.017)	5	0.0374 (0.026)
	0.0244 (0.012)	**0.0365** (0.023)	**0.0360** (0.023)		0.0253 (0.022)		0.0388 (0.025)
						20	0.0433 (0.039)
							0.0516 (0.046)
						50	0.0598 (0.076)
							0.0643 (0.075)
				0.3	*0.0223* (0.012)		
					0.0242 (0.013)		
				0.5	0.0231 (0.012)	5	0.0368 (0.026)
					0.0244 (0.012)		0.0388 (0.025)
						20	0.0429 (0.031)
							0.0466 (0.030)
						50	0.0560 (0.055)
							0.0586 (0.060)
				0.7	0.0244 (0.014)		
					0.0261 (0.013)		
				0.9	0.0279 (0.018)	5	**0.0364** (0.026)
					0.0310 (0.017)		**0.0384** (0.025)
						20	0.0395 (0.028)
							0.0442 (0.026)
						50	0.0433 (0.034)
							0.0462 (0.034)

Tabelle C.17: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Kreuzen“ ($p = 10$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	15.8902 (0.045)	6.6411 (37.355)	2.2606 (0.302)	0.1	28.5564 (0.086)	5	6.4667 (1.363)
	15.9003 (0.047)	6.5102 (36.727)	2.2065 (0.299)		28.4893 (0.085)		6.4361 (1.358)
						20	7.7186 (4.463)
							7.6792 (4.458)
						50	12.6416 (1.674)
							12.6010 (1.735)
				0.3	22.2368 (0.033)		
					22.2298 (0.036)		
				0.5	15.8902 (0.045)	5	3.6712 (0.243)
					15.9003 (0.047)		3.6769 (0.229)
						20	4.5176 (1.920)
							4.5220 (1.930)
						50	9.7514 (1.332)
							9.7412 (1.299)
				0.7	9.5469 (0.049)		
					9.5519 (0.050)		
				0.9	3.2063 (0.038)	5	*0.8375* (0.070)
					3.2068 (0.038)		*0.8430* (0.070)
						20	1.2260 (0.267)
							1.2289 (0.258)
						50	5.2986 (10.255)
							5.3191 (10.519)
10	0.2490 (0.700)	1.9477 (1.248)	1.9564 (1.505)	0.1	*0.2480* (0.694)	5	0.2871 (1.004)
	0.2477 (0.696)	1.8974 (1.124)	1.9056 (1.390)		*0.2466* (0.690)		0.2845 (0.990)
						20	0.2584 (0.790)
							0.2572 (0.780)
						50	0.2578 (0.824)
							0.2558 (0.814)
				0.3	0.2484 (0.696)		
					0.2470 (0.692)		
				0.5	0.2490 (0.700)	5	0.2866 (0.994)
					0.2477 (0.696)		0.2838 (0.983)
						20	0.2605 (0.794)
							0.2589 (0.793)
						50	0.2580 (0.820)
							0.2567 (0.824)
				0.7	0.2506 (0.709)		
					0.2493 (0.705)		
				0.9	0.2579 (0.760)	5	0.2944 (1.052)
					0.2567 (0.758)		0.2919 (1.042)
						20	0.2717 (0.870)
							0.2705 (0.868)
						50	0.2663 (0.901)
							0.2653 (0.881)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
20	0.1607 (0.182)	1.9262 (0.431)	1.9301 (0.457)	0.1	*0.1595* (0.179)	5	0.2005 (0.301)
	0.1587 (0.179)	1.8775 (0.415)	1.8815 (0.462)		*0.1574* (0.176)		0.1959 (0.299)
						20	0.1740 (0.297)
							0.1719 (0.296)
						50	0.1731 (0.320)
							0.1710 (0.316)
				0.3	0.1599 (0.180)		
					0.1578 (0.177)		
				0.5	0.1607 (0.182)	5	0.2001 (0.297)
					0.1587 (0.179)		0.1952 (0.295)
						20	0.1767 (0.276)
							0.1727 (0.278)
						50	0.1721 (0.286)
							0.1713 (0.284)
				0.7	0.1626 (0.186)		
					0.1606 (0.184)		
				0.9	0.1704 (0.207)	5	0.2080 (0.318)
					0.1685 (0.206)		0.2033 (0.318)
						20	0.1876 (0.285)
							0.1842 (0.288)
						50	0.1789 (0.283)
							0.1778 (0.278)
50	0.0996 (0.062)	1.9140 (0.209)	1.9171 (0.188)	0.1	*0.0983* (0.059)	5	0.1369 (0.124)
	0.1013 (0.059)	1.8675 (0.202)	1.8699 (0.185)		*0.0997* (0.057)		0.1356 (0.121)
						20	0.1153 (0.108)
							0.1157 (0.106)
						50	0.1286 (0.741)
							0.1296 (0.738)
				0.3	0.0987 (0.060)		
					0.1002 (0.057)		
				0.5	0.0996 (0.062)	5	0.1366 (0.122)
					0.1013 (0.059)		0.1353 (0.118)
						20	0.1167 (0.104)
							0.1188 (0.104)
						50	0.1228 (0.508)
							0.1240 (0.487)
				0.7	0.1020 (0.065)		
					0.1037 (0.062)		
				0.9	0.1110 (0.079)	5	0.1450 (0.135)
					0.1128 (0.075)		0.1436 (0.131)
						20	0.1305 (0.118)
							0.1300 (0.117)
						50	0.1248 (0.290)
							0.1276 (0.290)
100	0.0719 (0.032)	1.9092 (0.115)	1.9115 (0.091)	0.1	0.0716 (0.031)	5	0.1008 (0.067)
	0.0748 (0.031)	1.8636 (0.111)	1.8657 (0.089)		*0.0733* (0.030)		0.1016 (0.065)
						20	0.0907 (0.067)
							0.0913 (0.064)
						50	0.1057 (0.883)
							0.1063 (0.890)
				0.3	*0.0712* (0.031)		
					0.0736 (0.030)		
				0.5	0.0719 (0.032)	5	0.0995 (0.065)
					0.0748 (0.031)		0.1017 (0.063)
						20	0.0890 (0.061)
							0.0939 (0.061)
						50	0.0960 (0.570)
							0.1011 (0.532)
				0.7	0.0746 (0.035)		
					0.0777 (0.034)		
				0.9	0.0839 (0.045)	5	0.1070 (0.073)
					0.0880 (0.043)		0.1105 (0.071)
						20	0.1014 (0.070)
							0.1043 (0.069)
						50	0.1009 (0.293)
							0.1028 (0.290)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
200	*0.0550* (0.018)	1.9048 (0.058)	1.9059 (0.044)	0.1	0.0578 (0.018)	5	0.0793 (0.037)
	0.0556 (0.018)	1.8600 (0.055)	1.8606 (0.044)		0.0553 (0.018)		0.0763 (0.037)
						20	0.0786 (0.046)
							0.0735 (0.045)
						50	0.1044 (1.514)
							0.1014 (1.446)
				0.3	0.0555 (0.017)		
					0.0545 (0.017)		
				0.5	*0.0550* (0.018)	5	0.0744 (0.034)
					0.0556 (0.018)		0.0755 (0.034)
						20	0.0727 (0.040)
							0.0750 (0.039)
						50	0.0891 (0.711)
							0.0899 (0.782)
				0.7	0.0574 (0.020)		
					0.0589 (0.020)		
				0.9	0.0662 (0.027)	5	0.0796 (0.039)
					0.0693 (0.027)		0.0839 (0.038)
						20	0.0802 (0.041)
							0.0821 (0.040)
						50	0.0879 (0.379)
							0.0864 (0.335)
300	***0.0486*** (0.012)	**1.9030** (0.037)	**1.9032** (0.028)	0.1	0.0544 (0.015)	5	0.0720 (0.028)
	0.0474 (0.013)	**1.8565** (0.037)	**1.8573** (0.028)		0.0486 (0.015)		0.0650 (0.027)
						20	0.0756 (0.041)
							0.0660 (0.041)
						50	0.1216 (2.771)
							0.1128 (2.646)
				0.3	0.0501 (0.013)		
					0.0466 (0.013)		
				0.5	***0.0486*** (0.012)	5	**0.0637** (0.023)
					0.0474 (0.013)		**0.0624** (0.023)
						20	0.0660 (0.032)
							0.0639 (0.031)
						50	0.0989 (1.190)
							0.0939 (1.269)
				0.7	0.0507 (0.014)		
					0.0503 (0.014)		
				0.9	0.0588 (0.020)	5	0.0676 (0.026)
					0.0597 (0.019)		0.0686 (0.026)
						20	0.0700 (0.029)
							0.0688 (0.027)
						50	0.0899 (0.680)
							0.0827 (0.574)

Tabelle C.18: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **„Vorbeilaufen“ ($p = 10$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	15.8902 (0.045)	6.8466 (38.147)	*2.2235* (0.302)	0.1	28.5564 (0.086)	5	26.1481 (4.435)
	15.8971 (0.046)	6.9340 (39.023)	*2.1763* (0.322)		28.5135 (0.085)		26.1025 (4.390)
						20	27.5835 (0.902)
							27.5362 (0.838)
						50	27.3948 (0.954)
							27.3249 (1.110)
				0.3	22.2368 (0.033)		
					22.2321 (0.034)		
				0.5	15.8902 (0.045)	5	15.2793 (1.317)
					15.8971 (0.046)		15.2790 (1.371)
						20	15.6599 (0.192)
							15.6577 (0.219)
						50	15.6041 (0.276)
							15.6071 (0.301)
				0.7	9.5469 (0.049)		
					9.5513 (0.049)		
				0.9	3.2063 (0.038)	5	3.1520 (0.090)
					3.2080 (0.037)		3.1541 (0.087)
						20	3.1887 (0.043)
							3.1906 (0.042)
						50	3.1915 (0.044)
							3.1936 (0.043)
10	0.2490 (0.700)	1.9481 (1.261)	1.9562 (1.500)	0.1	*0.2480* (0.694)	5	0.2482 (0.695)
	0.2477 (0.696)	1.9258 (1.212)	1.9353 (1.495)		*0.2466* (0.690)		0.2469 (0.691)
						20	0.2482 (0.695)
							0.2467 (0.690)
						50	*0.2480* (0.694)
							0.2468 (0.690)
				0.3	0.2484 (0.696)		
					0.2470 (0.692)		
				0.5	0.2490 (0.700)	5	0.2491 (0.700)
					0.2477 (0.696)		0.2477 (0.696)
						20	0.2490 (0.700)
							0.2477 (0.696)
						50	0.2491 (0.700)
							0.2477 (0.696)
				0.7	0.2506 (0.709)		
					0.2493 (0.705)		
				0.9	0.2579 (0.760)	5	0.2580 (0.761)
					0.2567 (0.758)		0.2568 (0.758)
						20	0.2580 (0.760)
							0.2568 (0.758)
						50	0.2580 (0.760)
							0.2568 (0.758)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
20	0.1607 (0.182)	1.9259 (0.420)	1.9299 (0.463)	0.1	*0.1595* (0.179)	5	0.1598 (0.183)
	0.1587 (0.179)	1.9036 (0.407)	1.9085 (0.475)		*0.1574* (0.176)		0.1583 (0.180)
						20	0.1601 (0.181)
							0.1580 (0.178)
						50	0.1597 (0.179)
							0.1578 (0.177)
				0.3	0.1599 (0.180)		
					0.1578 (0.177)		
				0.5	0.1607 (0.182)	5	0.1608 (0.182)
					0.1587 (0.179)		0.1589 (0.180)
						20	0.1607 (0.182)
							0.1587 (0.179)
						50	0.1607 (0.182)
							0.1586 (0.179)
				0.7	0.1626 (0.186)		
					0.1606 (0.184)		
				0.9	0.1704 (0.207)	5	0.1705 (0.207)
					0.1685 (0.206)		0.1686 (0.206)
						20	0.1704 (0.207)
							0.1686 (0.206)
						50	0.1704 (0.207)
							0.1685 (0.206)
50	0.0996 (0.062)	1.9143 (0.211)	1.9169 (0.189)	0.1	*0.0983* (0.059)	5	0.0988 (0.063)
	0.1013 (0.059)	1.8923 (0.201)	1.8946 (0.184)		*0.0997* (0.057)		0.1004 (0.060)
						20	0.1005 (0.064)
							0.1013 (0.061)
						50	0.0991 (0.061)
							0.1012 (0.058)
				0.3	0.0987 (0.060)		
					0.1002 (0.057)		
				0.5	0.0996 (0.062)	5	0.0998 (0.062)
					0.1013 (0.059)		0.1014 (0.059)
						20	0.0996 (0.062)
							0.1014 (0.059)
						50	0.0998 (0.062)
							0.1014 (0.059)
				0.7	0.1020 (0.065)		
					0.1037 (0.062)		
				0.9	0.1110 (0.079)	5	0.1111 (0.079)
					0.1128 (0.075)		0.1128 (0.076)
						20	0.1111 (0.079)
							0.1128 (0.075)
						50	0.1110 (0.079)
							0.1127 (0.075)
100	0.0719 (0.032)	1.9090 (0.115)	1.9113 (0.092)	0.1	0.0716 (0.031)	5	0.0724 (0.033)
	0.0748 (0.031)	1.8867 (0.110)	1.8887 (0.091)		*0.0733* (0.030)		0.0744 (0.032)
						20	0.0751 (0.038)
							0.0763 (0.037)
						50	0.0738 (0.034)
							0.0763 (0.033)
				0.3	*0.0712* (0.031)		
					0.0736 (0.030)		
				0.5	0.0719 (0.032)	5	0.0723 (0.032)
					0.0748 (0.031)		0.0749 (0.031)
						20	0.0722 (0.032)
							0.0751 (0.031)
						50	0.0723 (0.032)
							0.0751 (0.031)
				0.7	0.0746 (0.035)		
					0.0777 (0.034)		
				0.9	0.0839 (0.045)	5	0.0840 (0.045)
					0.0880 (0.043)		0.0880 (0.043)
						20	0.0840 (0.045)
							0.0879 (0.043)
						50	0.0840 (0.045)
							0.0880 (0.043)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	*0.0550* (0.018)	1.9048 (0.058)	1.9059 (0.044)	0.1	0.0578 (0.018)	5	0.0584 (0.020)
	0.0556 (0.018)	1.8831 (0.056)	1.8845 (0.043)		0.0553 (0.018)		0.0569 (0.019)
						20	0.0624 (0.025)
							0.0593 (0.025)
						50	0.0617 (0.024)
							0.0612 (0.024)
				0.3	0.0555 (0.017)		
					0.0545 (0.017)		
				0.5	*0.0550* (0.018)	5	0.0554 (0.018)
					0.0556 (0.018)		0.0558 (0.018)
						20	0.0557 (0.018)
							0.0560 (0.018)
						50	0.0558 (0.018)
							0.0562 (0.018)
				0.7	0.0574 (0.020)		
					0.0589 (0.020)		
				0.9	0.0662 (0.027)	5	0.0663 (0.027)
					0.0693 (0.027)		0.0693 (0.027)
						20	0.0663 (0.027)
							0.0692 (0.027)
						50	0.0663 (0.027)
							0.0693 (0.027)
300	***0.0486*** (0.012)	**1.9025** (0.037)	**1.9036** (0.028)	0.1	0.0544 (0.015)	5	0.0546 (0.016)
	0.0474 (0.013)	**1.8822** (0.036)	**1.8837** (0.028)		0.0486 (0.015)		0.0504 (0.015)
						20	0.0588 (0.021)
							0.0529 (0.021)
						50	0.0592 (0.022)
							0.0557 (0.021)
				0.3	0.0501 (0.013)		
					0.0466 (0.013)		
				0.5	***0.0486*** (0.012)	5	**0.0489** (0.013)
					0.0474 (0.013)		**0.0477** (0.013)
						20	0.0495 (0.013)
							0.0479 (0.013)
						50	0.0495 (0.013)
							0.0481 (0.014)
				0.7	0.0507 (0.014)		
					0.0503 (0.014)		
				0.9	0.0588 (0.020)	5	0.0589 (0.020)
					0.0597 (0.019)		0.0598 (0.019)
						20	0.0590 (0.020)
							0.0597 (0.019)
						50	0.0590 (0.020)
							0.0597 (0.019)

Tabelle C.19: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **Sudden Drift ($p = 10$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	1.6896 (0.004)	0.3048 (0.096)	**0.2446** (0.256)	0.1	2.2676 (0.066)	5	0.1508 (0.092)
	1.6909 (0.004)	0.3056 (0.087)	**0.2393** (0.260)		2.2810 (0.075)		0.1484 (0.093)
						20	***0.1304*** (0.093)
							0.1290 (0.096)
						50	0.1498 (0.129)
							0.1509 (0.135)
				0.3	1.9540 (0.008)		
					1.9523 (0.008)		
				0.5	1.6896 (0.004)	5	0.1515 (0.093)
					1.6909 (0.004)		0.1487 (0.093)
						20	0.1347 (0.097)
							0.1337 (0.100)
						50	0.1470 (0.164)
							0.1490 (0.163)
				0.7	1.4180 (0.007)		
					1.4209 (0.006)		
				0.9	0.6540 (0.016)	5	0.1550 (0.096)
					0.6539 (0.016)		0.1526 (0.097)
						20	0.1410 (0.103)
							0.1393 (0.104)
						50	0.1505 (0.179)
							0.1502 (0.184)
10	0.2623 (0.693)	0.3418 (0.932)	0.3807 (1.453)	0.1	*0.2611* (0.687)	5	0.3307 (1.237)
	0.2636 (0.701)	0.3415 (0.992)	0.3848 (1.535)		*0.2623* (0.694)		0.3326 (1.253)
						20	0.2995 (0.979)
							0.3017 (0.999)
						50	0.2871 (0.973)
							0.2880 (0.999)
				0.3	0.2615 (0.689)		
					0.2628 (0.696)		
				0.5	0.2623 (0.693)	5	0.3312 (1.240)
					0.2636 (0.701)		0.3323 (1.258)
						20	0.3053 (1.020)
							0.3065 (1.036)
						50	0.2927 (1.072)
							0.2938 (1.074)
				0.7	0.2642 (0.703)		
					0.2654 (0.713)		
				0.9	0.2723 (0.757)	5	0.3345 (1.276)
					0.2735 (0.783)		0.3359 (1.313)
						20	0.3150 (1.099)
							0.3165 (1.146)
						50	0.3112 (1.177)
							0.3136 (1.252)

Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
20	0.1898 (0.182)	0.2728 (0.285)	0.3044 (0.471)	0.1	*0.1877* (0.179)	5	0.2594 (0.400)
	0.1884 (0.184)	0.2719 (0.281)	0.3047 (0.487)		*0.1863* (0.180)		0.2588 (0.401)
						20	0.2338 (0.392)
							0.2347 (0.405)
						50	0.2239 (0.474)
							0.2253 (0.482)
				0.3	0.1884 (0.180)		
					0.1871 (0.182)		
				0.5	0.1898 (0.182)	5	0.2589 (0.397)
					0.1884 (0.184)		0.2585 (0.398)
						20	0.2378 (0.377)
							0.2374 (0.383)
						50	0.2278 (0.478)
							0.2287 (0.487)
				0.7	0.1929 (0.187)		
					0.1915 (0.189)		
				0.9	0.2048 (0.209)	5	0.2621 (0.405)
					0.2026 (0.211)		0.2609 (0.408)
						20	0.2465 (0.383)
							0.2451 (0.391)
						50	0.2422 (0.447)
							0.2433 (0.457)
50	**0.1792** (0.063)	**0.2620** (0.138)	0.2731 (0.194)	0.1	***0.1727*** (0.061)	5	0.2297 (0.177)
	0.1787 (0.063)	**0.2642** (0.135)	0.2732 (0.197)		***0.1720*** (0.061)		0.2289 (0.178)
						20	0.2208 (0.256)
							0.2230 (0.266)
						50	0.2350 (1.136)
							0.2400 (1.158)
				0.3	0.1751 (0.062)		
					0.1744 (0.062)		
				0.5	0.1792 (0.063)	5	0.2295 (0.174)
					0.1787 (0.063)		0.2297 (0.175)
						20	0.2239 (0.209)
							0.2250 (0.209)
						50	0.2376 (1.016)
							0.2408 (1.079)
				0.7	0.1878 (0.067)		
					0.1875 (0.067)		
				0.9	0.2079 (0.081)	5	0.2309 (0.177)
					0.2075 (0.080)		0.2306 (0.177)
						20	0.2254 (0.185)
							0.2264 (0.188)
						50	0.2329 (0.488)
							0.2356 (0.494)
100	0.2473 (0.035)	0.2932 (0.083)	0.2865 (0.102)	0.1	*0.2322* (0.034)	5	0.2343 (0.094)
	0.2471 (0.035)	0.2976 (0.079)	0.2853 (0.100)		0.2320 (0.034)		*0.2317* (0.094)
						20	0.2483 (0.154)
							0.2496 (0.153)
						50	0.2651 (1.289)
							0.2688 (1.301)
				0.3	0.2378 (0.034)		
					0.2374 (0.034)		
				0.5	0.2473 (0.035)	5	0.2348 (0.094)
					0.2471 (0.035)		0.2323 (0.092)
						20	0.2529 (0.129)
							0.2536 (0.130)
						50	0.2626 (1.067)
							0.2709 (1.122)
				0.7	0.2663 (0.038)		
					0.2664 (0.038)		
				0.9	0.2869 (0.049)	5	0.2352 (0.095)
					0.2870 (0.049)		0.2331 (0.094)
						20	0.2485 (0.111)
							0.2491 (0.110)
						50	0.2554 (0.378)
							0.2537 (0.394)

Fortsetzung auf der nächsten Seite

	ILDA	QDA-AF	LDA-AF	OLDC fix		OLDC adaptive	
N_{trend}				λ,λ_{start}		L	
200	0.4341 (0.023)	0.3791 (0.058)	0.3477 (0.062)	0.1	0.4072 (0.023)	5	0.2859 (0.050)
	0.4371 (0.023)	0.3866 (0.057)	0.3494 (0.062)		0.4113 (0.023)		*0.2831* (0.051)
						20	0.3240 (0.101)
							0.3214 (0.101)
						50	0.3552 (0.485)
							0.3576 (0.554)
				0.3	0.4167 (0.022)		
					0.4202 (0.022)		
				0.5	0.4341 (0.023)	5	0.2863 (0.050)
					0.4371 (0.023)		0.2833 (0.051)
						20	0.3248 (0.080)
							0.3242 (0.077)
						50	0.3634 (0.415)
							0.3656 (0.411)
				0.7	0.4661 (0.025)		
					0.4684 (0.026)		
				0.9	0.4657 (0.034)	5	*0.2858* (0.050)
					0.4665 (0.035)		0.2836 (0.051)
						20	0.3118 (0.061)
							0.3104 (0.062)
						50	0.3492 (0.171)
							0.3486 (0.182)
300	0.6331 (0.019)	0.4791 (0.052)	0.4447 (0.054)	0.1	0.6019 (0.020)	5	0.3598 (0.034)
	0.6375 (0.020)	0.4839 (0.055)	0.4428 (0.048)		0.6066 (0.022)		*0.3599* (0.035)
						20	0.4092 (0.078)
							0.4068 (0.079)
						50	0.4781 (0.369)
							0.4722 (0.363)
				0.3	0.6129 (0.019)		
					0.6173 (0.019)		
				0.5	0.6331 (0.019)	5	*0.3597* (0.034)
					0.6375 (0.020)		0.3604 (0.035)
						20	0.4037 (0.056)
							0.4063 (0.054)
						50	0.4821 (0.336)
							0.4795 (0.301)
				0.7	0.6693 (0.022)		
					0.6733 (0.022)		
				0.9	0.6425 (0.028)	5	0.3598 (0.034)
					0.6437 (0.029)		0.3604 (0.035)
						20	0.3894 (0.041)
							0.3895 (0.042)
						50	0.4445 (0.154)
							0.4447 (0.151)

Tabelle C.20: Durchschnittlicher euklidischer Abstand zwischen wahren und mittleren geschätzten bzw. mittleren prognostizierten Erwartungswertvektoren nach (9.2) bzw. (9.3) über die Zeit (durchschnittliche Varianz der euklidischen Abstände zwischen geschätzten bzw. prognostizierten und wahren Erwartungswertvektoren über die Zeit in Klammern) für die Datensituation **ohne Drift ($p = 10$)** getrennt nach Methoden für Online Diskriminanzanalyse und ihrer Erweiterungen durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend}:
grau kursiv: minimaler Wert pro „Zeile“ (pro N_{trend});
grau fettgedruckt: minimaler Wert pro „Spalte“ (pro Methode);
schwarz kursiv und fettgedruckt: minimaler Wert insgesamt.
Zwei aufeinanderfolgende Zeilen enthalten jeweils die Ergebnisse für Klasse 1 und Klasse 2 ($M = 2$ Klassen) für eine Parameterkombination.

	ILDA	**QDA-AF**	**LDA-AF**	**OLDC fix**		**OLDC adaptive**	
N_{trend}				λ,λ_{start}		L	
ohne	***0.0158*** (0.003)	**0.0422** (0.052)	0.1214 (0.285)	0.1	0.0840 (0.048)	5	0.1177 (0.088)
	0.0207 (0.003)	**0.0447** (0.054)	0.1224 (0.295)		0.0621 (0.047)		0.1184 (0.087)
						20	0.0896 (0.086)
							0.0897 (0.085)
						50	0.0769 (0.126)
							0.0795 (0.125)
				0.3	0.0325 (0.006)		
					0.0309 (0.006)		
				0.5	***0.0158*** (0.003)	5	0.1176 (0.087)
					0.0207 (0.003)		0.1197 (0.087)
						20	0.0994 (0.093)
							0.0972 (0.092)
						50	0.0902 (0.159)
							0.0912 (0.159)
				0.7	0.0193 (0.004)		
					0.0231 (0.004)		
				0.9	0.0378 (0.012)	5	0.1210 (0.091)
					0.0403 (0.012)		0.1228 (0.091)
						20	0.1068 (0.096)
							0.1059 (0.095)
						50	0.1061 (0.157)
							0.1062 (0.160)
10	0.2477 (0.693)	0.2588 (0.763)	0.3192 (1.445)	0.1	*0.2466* (0.687)	5	0.3112 (1.208)
	0.2511 (0.698)	0.2625 (0.775)	0.3181 (1.402)		*0.2501* (0.692)		0.3163 (1.224)
						20	0.2824 (0.952)
							0.2867 (0.954)
						50	0.2683 (0.888)
							0.2724 (0.880)
				0.3	0.2470 (0.689)		
					0.2504 (0.694)		
				0.5	0.2477 (0.693)	5	0.3114 (1.209)
					0.2511 (0.698)		0.3162 (1.224)
						20	0.2891 (1.004)
							0.2930 (1.012)
						50	0.2795 (1.000)
							0.2831 (1.001)
				0.7	0.2494 (0.704)		
					0.2526 (0.708)		
				0.9	0.2569 (0.763)	5	0.3150 (1.254)
					0.2598 (0.763)		0.3194 (1.263)
						20	0.2985 (1.097)
							0.3020 (1.091)
						50	0.2990 (1.190)
							0.3010 (1.151)

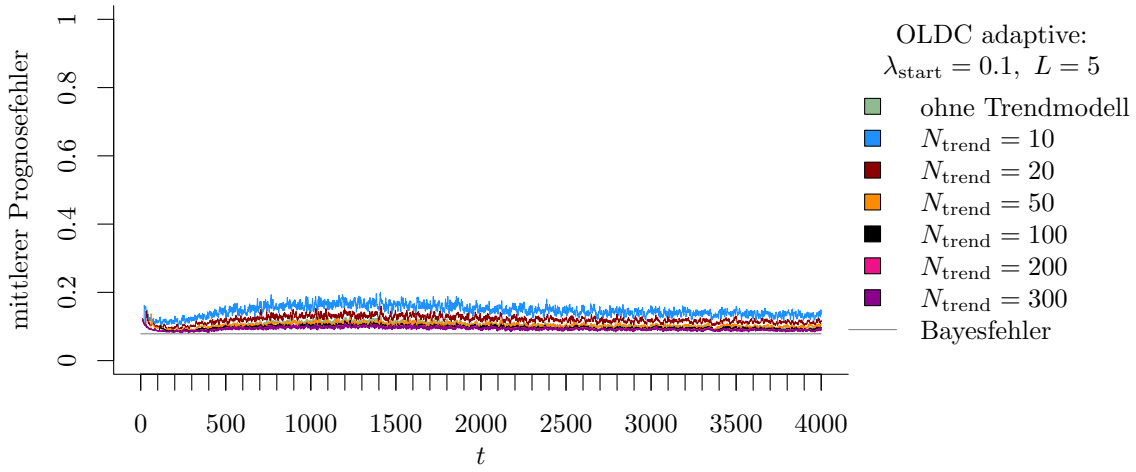
Fortsetzung auf der nächsten Seite

	ILDA	QDA-AF	LDA-AF	OLDC fix		OLDC adaptive	
N_{trend}				λ,λ_{start}		L	
20	0.1593 (0.181)	0.1707 (0.216)	0.2230 (0.450)	0.1	*0.1579* (0.178)	5	0.2249 (0.388)
	0.1621 (0.180)	0.1736 (0.217)	0.2252 (0.467)		*0.1609* (0.177)		0.2294 (0.390)
						20	0.2038 (0.342)
							0.2064 (0.339)
						50	0.1894 (0.356)
							0.1921 (0.351)
				0.3	0.1584 (0.179)		
					0.1613 (0.178)		
				0.5	0.1593 (0.181)	5	0.2248 (0.385)
					0.1621 (0.180)		0.2292 (0.385)
						20	0.2095 (0.353)
							0.2113 (0.351)
						50	0.1990 (0.394)
							0.2012 (0.394)
				0.7	0.1613 (0.186)		
					0.1640 (0.185)		
				0.9	0.1696 (0.208)	5	0.2272 (0.394)
					0.1720 (0.206)		0.2314 (0.394)
						20	0.2157 (0.371)
							0.2181 (0.367)
						50	0.2147 (0.407)
							0.2152 (0.407)
50	0.1011 (0.060)	0.1113 (0.078)	0.1621 (0.194)	0.1	*0.0992* (0.058)	5	0.1622 (0.168)
	0.1022 (0.060)	0.1127 (0.080)	0.1624 (0.205)		*0.1008* (0.058)		0.1652 (0.170)
						20	0.1586 (0.181)
							0.1557 (0.178)
						50	0.1601 (0.490)
							0.1576 (0.482)
				0.3	0.0999 (0.059)		
					0.1012 (0.059)		
				0.5	0.1011 (0.060)	5	0.1627 (0.166)
					0.1022 (0.060)		0.1648 (0.166)
						20	0.1618 (0.177)
							0.1574 (0.176)
						50	0.1653 (0.455)
							0.1639 (0.453)
				0.7	0.1037 (0.064)		
					0.1048 (0.063)		
				0.9	0.1134 (0.078)	5	0.1637 (0.169)
					0.1145 (0.077)		0.1660 (0.170)
						20	0.1620 (0.173)
							0.1599 (0.172)
						50	0.1622 (0.252)
							0.1630 (0.254)
100	0.0742 (0.031)	0.0850 (0.047)	0.1179 (0.099)	0.1	*0.0719* (0.030)	5	0.1211 (0.087)
	0.0745 (0.031)	0.0869 (0.049)	0.1238 (0.094)		*0.0729* (0.030)		0.1226 (0.087)
						20	0.1284 (0.118)
							0.1257 (0.120)
						50	0.1384 (0.328)
							0.1344 (0.347)
				0.3	0.0727 (0.030)		
					0.0732 (0.030)		
				0.5	0.0742 (0.031)	5	0.1213 (0.087)
					0.0745 (0.031)		0.1226 (0.086)
						20	0.1300 (0.107)
							0.1244 (0.109)
						50	0.1372 (0.262)
							0.1327 (0.252)
				0.7	0.0773 (0.034)		
					0.0778 (0.034)		
				0.9	0.0873 (0.044)	5	0.1220 (0.087)
					0.0885 (0.044)		0.1233 (0.087)
						20	0.1269 (0.095)
							0.1240 (0.095)
						50	0.1278 (0.127)
							0.1275 (0.125)

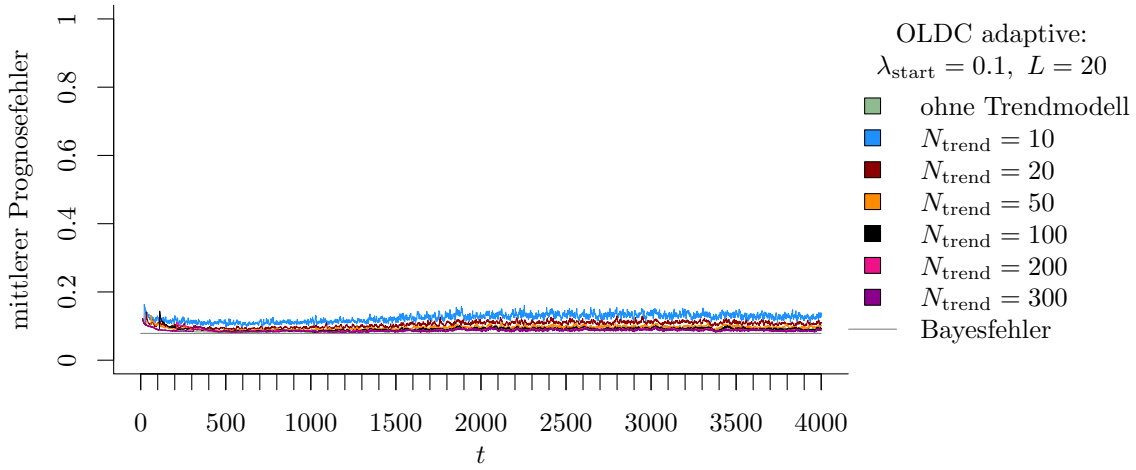
Fortsetzung auf der nächsten Seite

	ILDA	**QDA-AF**	**LDA-AF**		**OLDC fix**		**OLDC adaptive**
N_{trend}				λ,λ_{start}		L	
200	0.0561 (0.018)	0.0669 (0.027)	0.0891 (0.045)	0.1	*0.0542* (0.018)	5	0.0879 (0.043)
	0.0566 (0.017)	0.0668 (0.029)	0.0907 (0.044)		*0.0551* (0.019)		0.0879 (0.043)
						20	0.0956 (0.065)
							0.0992 (0.072)
						50	0.1099 (0.164)
							0.1130 (0.164)
				0.3	0.0546 (0.017)		
					0.0551 (0.017)		
				0.5	0.0561 (0.018)	5	0.0881 (0.043)
					0.0566 (0.017)		0.0880 (0.043)
						20	0.0981 (0.054)
							0.0955 (0.056)
						50	0.1072 (0.113)
							0.1057 (0.112)
				0.7	0.0591 (0.020)		
					0.0600 (0.019)		
				0.9	0.0681 (0.026)	5	0.0882 (0.043)
					0.0703 (0.027)		0.0878 (0.043)
						20	0.0943 (0.047)
							0.0922 (0.048)
						50	0.0942 (0.058)
							0.0934 (0.060)
300	0.0486 (0.013)	0.0597 (0.019)	**0.0724** (0.029)	0.1	*0.0471* (0.016)	5	**0.0711** (0.028)
	0.0496 (0.013)	0.0611 (0.021)	**0.0745** (0.031)		0.0483 (0.016)		0.0738 (0.028)
						20	0.0803 (0.047)
							0.0846 (0.051)
						50	0.0951 (0.111)
							0.1004 (0.114)
				0.3	*0.0471* (0.013)		
					0.0480 (0.013)		
				0.5	0.0486 (0.013)	5	0.0716 (0.028)
					0.0496 (0.013)		**0.0735** (0.028)
						20	0.0809 (0.034)
							0.0808 (0.035)
						50	0.0890 (0.073)
							0.0928 (0.071)
				0.7	0.0514 (0.015)		
					0.0529 (0.014)		
				0.9	0.0591 (0.020)	5	0.0713 (0.028)
					0.0627 (0.020)		**0.0735** (0.028)
						20	0.0766 (0.030)
							0.0776 (0.031)
						50	0.0759 (0.036)
							0.0795 (0.038)

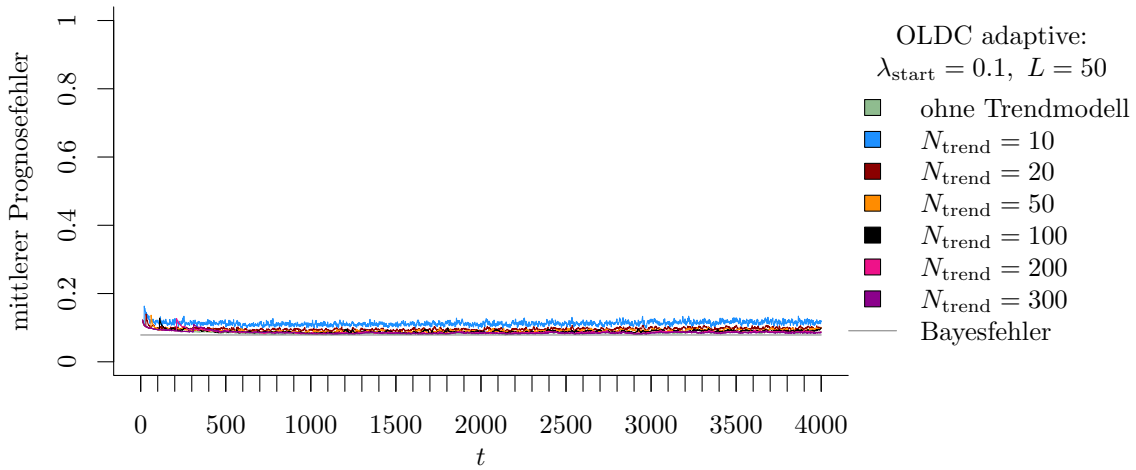
D Grafiken Simulationsergebnisse



(a) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

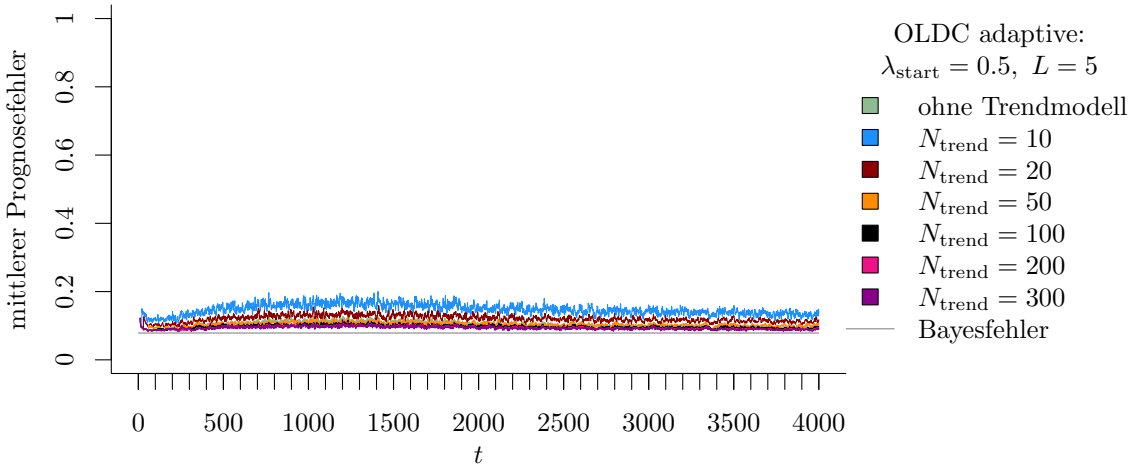


(b) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

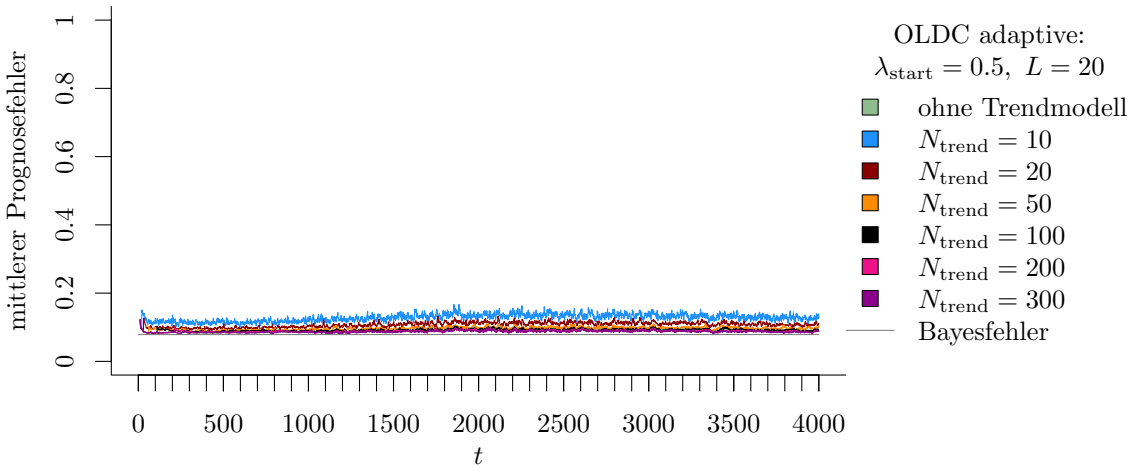


(c) **OLDC** mit $\lambda_{\text{start}} = 0.1$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

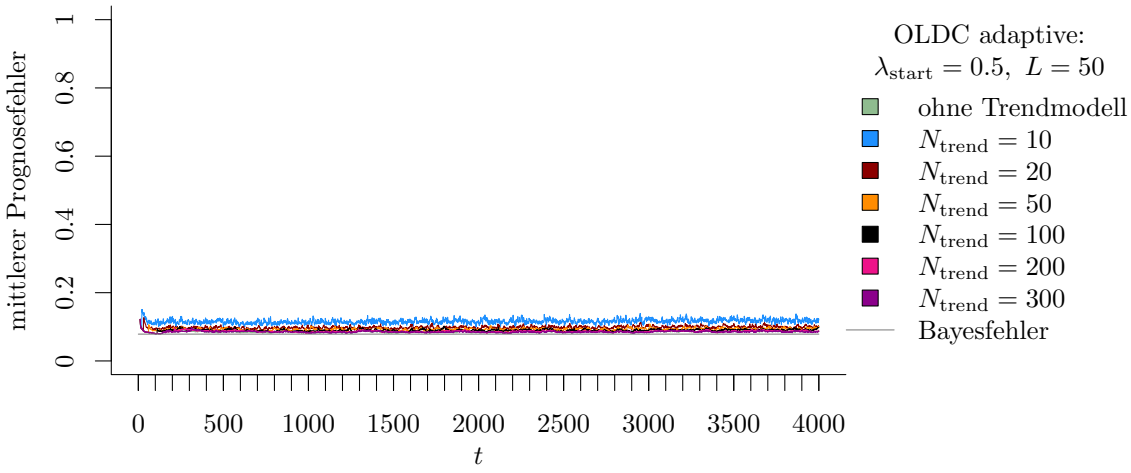
Abbildung D.1: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **ohne Drift** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{start} = 0.5$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.

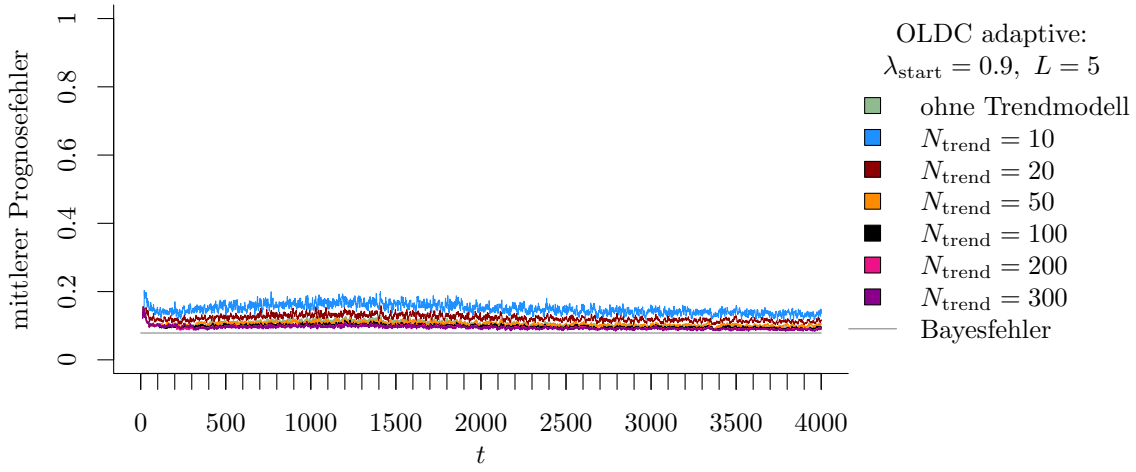


(b) **OLDC** mit $\lambda_{start} = 0.5$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.

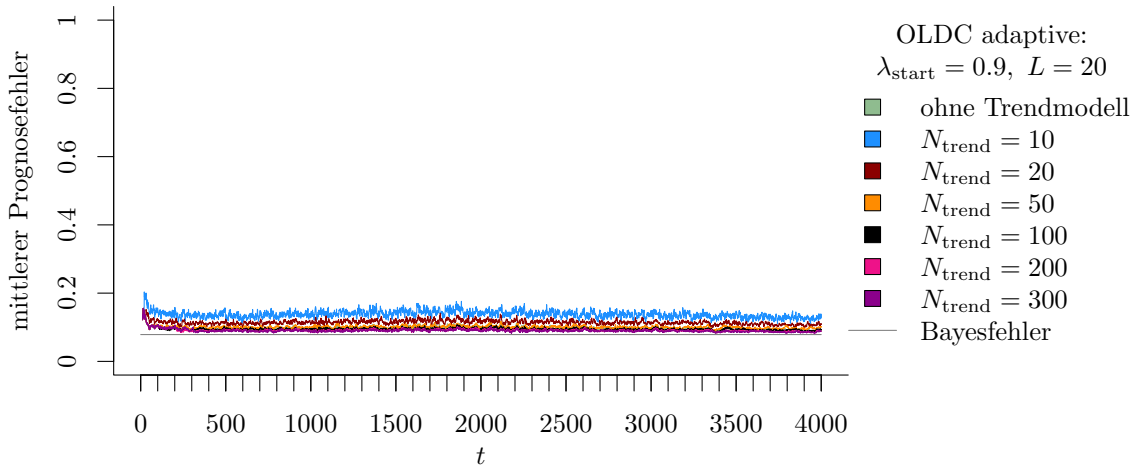


(c) **OLDC** mit $\lambda_{start} = 0.5$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

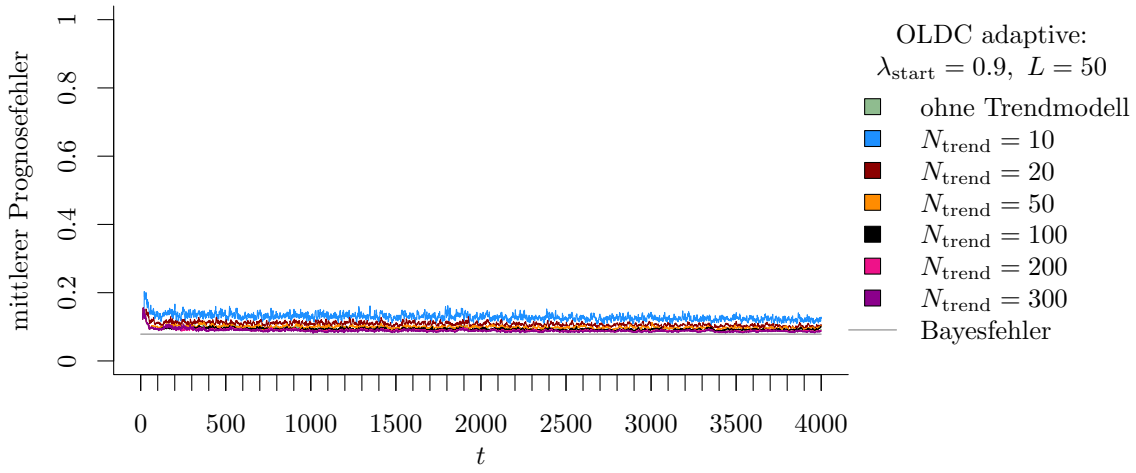
Abbildung D.2: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **ohne Drift** im zweidimensionalen Raum.



(a) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 5$ und Erweiterung durch lokale lineare Regressionsmodelle.



(b) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 20$ und Erweiterung durch lokale lineare Regressionsmodelle.



(c) **OLDC** mit $\lambda_{\text{start}} = 0.9$ und $L = 50$ und Erweiterung durch lokale lineare Regressionsmodelle.

Abbildung D.3: Mittlerer Prognosefehler über die Zeit für *OLDC* mit adaptiver Lernrate mit verschiedenen Parameterkombinationen und für die Erweiterung durch lokale lineare Regressionsmodelle für verschiedene Werte von N_{trend} auf der Datensituation **ohne Drift** im zweidimensionalen Raum.

Literaturverzeichnis

Aggarwal, Charu C., Hrsg. (2007). *Data Streams. Models and Algorithms*. Bd. 31. Advances in Database Systems. New, York: Springer. ISBN: 978-0-387-28759-1. DOI: 10.1007/978-0-387-47534-9.

Alaiz-Rodríguez, Rocío und Nathalie Japkowicz (2008). „Assessing the Impact of Changing Environments on Classifier Performance“. In: *Advances in Artificial Intelligence*. Hrsg. von Sabine Bergler. Bd. 5032. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, S. 13–24. ISBN: 978-3-540-68825-9. DOI: 10.1007/978-3-540-68825-9.

Anagnostopoulos, Christoforos, Dimitris K. Tasoulis, Niall M. Adams, Nicos G. Pavlidis und David J. Hand (2012). „Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification“. In: *Statistical Analysis and Data Mining* 5(2), S. 139–166. ISSN: 1932-1872. DOI: 10.1002/sam.10151.

Anderson, Ed, Zhao-jun Bai, Chris Bischof, Susan Blackford, Jim Demmel, Jack Dongarra, Jeremy Du Croz, Anne Greenbaum, Sven Hammarling, Alan McKenney und Danny Sorensen (1999). *LAPACK Users' Guide*. 3. Aufl. Philadelphia: Society for Industrial und Applied Mathematics. ISBN: 978-0-89871-447-0. DOI: 10.1137/1.9780898719604.

Aschersleben, Philipp (2016). „Online-Klassifikation bei Concept Drift mit Naive Bayes“. Masterarbeit. Technische Universität Dortmund.

Babcock, Brian, Shivnath Babu, Mayur Datar, Rajeev Motwani und Jennifer Widom (2002). „Models and Issues in Data Stream Systems“. In: *Proceedings of the Twenty-first ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. PODS '02. New York, NY, USA: ACM, S. 1–16. ISBN: 978-1-58113-507-7. DOI: 10.1145/543613.543615.

Bickel, Steffen, Michael Brückner und Tobias Scheffer (2009). „Discriminative Learning Under Covariate Shift“. In: *Journal of Machine Learning Research* 10, S. 2137–2155. ISSN: 1532-4435.

Bischl, Bernd, Michel Lang, Olaf Mersmann, Jörg Rahnenführer und Claus Weihs (2015). „BatchJobs and BatchExperiments: Abstraction Mechanisms for Using R in Batch Environments“. In: *Journal of Statistical Software* 64(11), S. 1–25. ISSN: 1548-7660. DOI: 10.18637/jss.v064.i11.

Bryan, Joseph G. (1951). „The generalized discriminant function: mathematical foundation and computational routine“. In: *Harvard Educational Review* 21(2), S. 90–95. ISSN: 0017-8055.

Cieslak, David A. und Nitesh V. Chawla (2009). „A framework for monitoring classifiers' performance: when and why failure occurs?“ In: *Knowledge and Information Systems* 18(1), S. 83–108. ISSN: 0219-3116. DOI: 10.1007/s10115-008-0139-1.

Delany, Sarah Jane, Pádraig Cunningham, Alexey Tsymbal und Lorcan Coyle (2005). „A case-based technique for tracking concept drift in spam filtering“. In: *Knowledge-Based Systems* 18(4), S. 187–195. ISSN: 0950-7051. DOI: 10.1016/j.knosys.2004.10.002.

Duda, Richard O., Peter E. Hart und David G. Stork (2001). *Pattern Classification*. 2. Aufl. New York: John Wiley & Sons, Inc. ISBN: 978-0-471-05669-0.

Fahrmeir, Ludwig, Walter Häußler und Gerhard Tutz (1996a). „Diskriminanzanalyse“. In: *Multivariate statistische Verfahren*. Hrsg. von Ludwig Fahrmeir, Alfred Hamerle und Gerhard Tutz. Berlin, New York: de Gruyter, S. 357–436. ISBN: 978-3-11-081602-0.

Fahrmeir, Ludwig, Heinz Kaufmann und Christian Kredler (1996b). „Regressionsanalyse“. In: *Multivariate statistische Verfahren*. Hrsg. von Ludwig Fahrmeir, Alfred Hamerle und Gerhard Tutz. Berlin, New York: de Gruyter, S. 93–168. ISBN: 978-3-11-081602-0.

Fahrmeir, Ludwig, Thomas Kneib und Stefan Lang (2009). *Regression. Modelle, Methoden und Anwendungen*. 2. Aufl. Statistik und ihre Anwendungen. Berlin, Heidelberg: Springer-Verlag. ISBN: 978-3-642-01836-7. DOI: 10.1007/978-3-642-01837-4.

Filzmoser, Peter, Kristel Joossens und Christophe Croux (2006). „Multiple group linear discriminant analysis: robustness and error rate“. In: *Compstat 2006 – Proceedings in Computational Statistics*. Hrsg. von Alfredo Rizzi und Maurizio Vichi. Heidelberg, New York: Physica-Verlag HD, S. 521–532. ISBN: 978-3-7908-1708-9. DOI: 10.1007/978-3-7908-1709-6.

Fischer, Gerd (2014). *Lineare Algebra. Eine Einführung für Studienanfänger*. 18. Aufl. Wiesbaden: Springer Fachmedien. ISBN: 978-3-658-03945-5. DOI: 10.1007/978-3-658-03945-5.

Fisher, Ronald A. (1936). „The Use of Multiple Measurements in Taxonomic Problems“. In: *Annals of Eugenics* 7(2), S. 179–188. ISSN: 0003-4800. DOI: 10.1111/j.1469-1809.1936.tb02137.x.

Fukunaga, Keinosuke (1990). *Introduction to Statistical Pattern Recognition*. 2. Aufl. Computer Science and Scientific Computing. San Diego, CA, USA: Academic Press, Inc. ISBN: 978-0-12-269851-4.

Gaber, Mohamed Medhat, Arkady Zaslavsky und Shonali Krishnaswamy (2005). „Mining Data Streams: A Review“. In: *ACM SIGMOD Record* 34(2), S. 18–26. ISSN: 0163-5808. DOI: 10.1145/1083784.1083789.

Gama, João (2010). *Knowledge Discovery from Data Streams*. Data Mining and Knowledge Discovery Series. Boca Raton, FL: Chapman & Hall/CRC. ISBN: 978-1-4398-2611-9.

Gama, João, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy und Abdelhamid Bouchachia (2014). „A Survey on Concept Drift Adaptation“. In: *ACM Computing Surveys* 46(4). Nr. 44, S. 1–37. ISSN: 0360-0300. DOI: 10.1145/2523813.

Gao, Jing, Wei Fan, Jiawei Han und Philip S. Yu (2007). „A General Framework for Mining Concept-Drifting Data Streams with Skewed Distributions“. In: *Proceedings of the Seventh SIAM International Conference on Data Mining*. Hrsg. von Chid Apte, Bing Liu, Srinivasan Parthasarathy und David Skillicorn. Philadelphia: Society for Industrial und Applied Mathematics, S. 3–14. ISBN: 978-0-89871-630-6. DOI: 10.1137/1.9781611972771.1.

Garber, Fred D. und Abdelhamid Djouadi (1988). „Bounds on the Bayes classification error based on pairwise risk functions“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10(2), S. 281–288. ISSN: 0162-8828. DOI: 10.1109/34.3891.

Genz, Alan und Frank Bretz (2009). *Computation of Multivariate Normal and t Probabilities*. Bd. 195. Lecture Notes in Statistics. Berlin, Heidelberg: Springer-Verlag. ISBN: 978-3-642-01688-2. DOI: 10.1007/978-3-642-01689-9.

Genz, Alan, Frank Bretz, Tetsuhisa Miwa, Xuefei Mi, Friedrich Leisch, Fabian Scheipl und Torsten Hothorn (2019). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-10. URL: https://CRAN.R-project.org/package=mvtnorm.

Golub, Gene H. und Charles F. Van Loan (1996). *Matrix Computations*. 3. Aufl. Baltimore, London: The John Hopkins University Press. ISBN: 978-0-8018-5414-9.

Greene, William H., Hrsg. (2002). *Econometric Analysis*. Bd. 5. Upper Saddle River: Prentice Hall. ISBN: 978-0-13-066189-0.

Groß, Jürgen (2003). *Linear Regression*. Bd. 175. Lecture Notes in Statistics. Berlin, Heidelberg: Springer-Verlag. ISBN: 978-3-540-40178-0. DOI: 10.1007/978-3-642-55864-1.

Hand, David J. (2006). „Classifier Technology and the Illusion of Progress“. In: *Statistical Science* 21(1), S. 1–14. ISSN: 0883-4237. DOI: 10.1214/088342306000000060.

Hartung, Joachim und Bärbel Elpelt (1999). *Multivariate Statistik: Lehr- und Handbuch der angewandten Statistik*. 6. Aufl. München: Oldenbourg. ISBN: 978-3-486-25287-3.

Hartung, Joachim, Bärbel Elpelt und Karl-Heinz Klösener (1995). *Statistik: Lehr- und Handbuch der angewandten Statistik*. 10. Aufl. München: Oldenbourg. ISBN: 978-3-486-23387-2.

Harville, David A. (2008). *Matrix Algebra from a Statistician's Perspective*. Reprint der Ausgabe von 1997. New York: Springer. ISBN: 978-0-387-78356-7. DOI: 10.1007/b98818.

Hastie, Trevor, Robert Tibshirani und Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. Aufl. Series in Statistics. Corrected 12th printing. New York: Springer. ISBN: 978-0-387-84858-7. DOI: 10.1007/978-0-387-84858-7.

Hellinger, Ernst (1909). „Neue Begründung der Theorie quadratischer Formen von unendlich vielen Veränderlichen“. In: *Journal für die reine und angewandte Mathematik* 136, S. 210–271. ISSN: 1435-5345. DOI: 10.1515/crll.1909.136.210.

Hoens, T. Ryan, Nitesh V. Chawla und Polikar Robi (2011). „Heuristic Updatable Weighted Random Subspaces for Non-stationary Environments“. In: *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*. IEEE, S. 241–250. ISBN: 978-1-4577-2075-8. DOI: 10.1109/ICDM.2011.75.

Hoens, T. Ryan, Robi Polikar und Nitesh V. Chawla (2012). „Learning from streaming data with concept drift and imbalance: an overview“. In: *Progress in Artificial Intelligence* 1(1), S. 89–101. ISSN: 2192-6360. DOI: 10.1007/s13748-011-0008-0.

Huang, David T. J., Yun Sing Koh, Gillian Dobbie und Russel Pears (2013). „Tracking Drift Types in Changing Data Streams“. In: *Advanced Data Mining and Applications*. Hrsg. von Hiroshi Motoda, Zhaohui Wu, Longbing Cao, Osmar Zaiane, Min Yao und Wei Wang. Bd. 8346. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, S. 72–83. ISBN: 978-3-642-53914-5. DOI: 10.1007/978-3-642-53914-5_7.

Huang, Jiayuan, Alex J. Smola, Arthur Gretton, Karsten M. Borgwardt und Bernhard Schölkopf (2007). „Correcting Sample Selection Bias by Unlabeled Data“. In: *Advances in Neural Information Processing Systems 19*. Hrsg. von

Bernhard Schölkopf, John Platt und Thomas Hoffman. Cambridge, Massachusetts: The MIT Press, S. 601–608. ISBN: 978-0-262-19568-3. DOI: 10.7551/mitpress/7503.001.0001.

Huberty, Carl J. (1994). *Applied Discriminant Analysis*. 1. Aufl. Series in Probability and Mathematical Statistics. New York: John Wiley & Sons, Inc. ISBN: 978-0-471-31145-4.

Hulten, Geoff, Laurie Spencer und Pedro Domingos (2001). „Mining Time-changing Data Streams“. In: *KDD-2001: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Hrsg. von Foster Provost und Ramakrishnan Srikant. New York, NY, USA: ACM, S. 97–106. ISBN: 978-158113-391-2. DOI: 10.1145/502512.502529.

Johnson, Richard A. und Dean W. Wichern (2007). *Applied Multivariate Statistical Analysis*. 6. Aufl. Upper Saddle River, NJ, USA: Pearson Education, Inc. ISBN: 978-013-514350-6.

Kelly, Mark G., David J. Hand und Niall M. Adams (1999). „The Impact of Changing Populations on Classifier Performance“. In: *KDD-99: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Hrsg. von Surajit Chaudhuri und David Madigan. New York, NY, USA: ACM, S. 367–371. ISBN: 978-1-58113-143-7. DOI: 10.1145/312129.312285.

Knüsel, Leo (1993). *Equivalence of the maximum-likelihood method and the canonical variables method in discriminant analysis*. Manuskript. Universität München.

Kolajo, Taiwo, Olawande Daramola und Ayodele Adebiyi (2019). „Big data stream analysis: a systematic literature review“. In: *Journal of Big Data* 6(1), S. 47. ISSN: 2196-1115. DOI: 10.1186/s40537-019-0210-7.

Krzanowski, Wojtek J. und Francis H. C. Marriott (1995). *Multivariate Analysis Part 2: Classification, Covariance Structures and Repeated Measurements*. Bd. 2. Kendall's Library of Statistics. London: Arnold. ISBN: 978-0-340-59325-7.

Kullback, Solomon und Richard A. Leibler (1951). „On Information and Sufficiency“. In: *The Annals of Mathematical Statistics* 22(1), S. 79–86. ISSN: 0003-4851. DOI: 10.1214/aoms/1177729694.

Kuncheva, Ludmila I. (2004). „Classifier Ensembles for Changing Environments“. In: *Multiple Classifier Systems*. Hrsg. von Fabio Roli, Josef Kittler und Terry Windeatt. Bd. 3077. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, S. 1–15. ISBN: 978-3-540-25966-4. DOI: 10.1007/978-3-540-25966-4_1.

Kuncheva, Ludmila I. und Catrin O. Plumpton (2008). „Adaptive Learning Rate for Online Linear Discriminant Classifiers“. In: *Structural, Syntactic, and Statistical Pattern Recognition*. Hrsg. von Niels da Vitoria Lobo, Takis Kasparis, Fabio Roli, James T. Kwok, Michael Georgiopoulos, Georgios C. Anagnostopoulos und Marco Loog. Bd. 5342. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, S. 510–519. ISBN: 978-3-540-89689-0. DOI: 10.1007/978-3-540-89689-0_55.

Ligges, Uwe und Martin Mächler (2003). „Scatterplot3d - an R Package for Visualizing Multivariate Data“. In: *Journal of Statistical Software* 8(11), S. 1–20. ISSN: 1548-7660. DOI: 10.18637/jss.v008.i11.

Lumley, Thomas (2013). *biglm: bounded memory linear and generalized linear models*. R package version 0.9-1. URL: https://CRAN.R-project.org/package=biglm.

Mahalanobis, Prasanta C. (1936). „On the generalised distance in statistics“. In: *Proceedings of the National Institute of Sciences of India* 2(1), S. 49–55. ISSN: 0370-0046.

Mardia, Kanti V., John T. Kent und John M. Bibby (1979). *Multivariate Analysis*. Probability and Mathematical Statistics. London, San Diego: Academic Press Limited. ISBN: 978-0-12-471252-2.

McLachlan, Geoffrey J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN: 978-0-471-61531-6. DOI: 10.1002/0471725293.

van Meegen, Carmen (2015). „Ungleiche a priori Wahrscheinlichkeiten in linearen Diskriminanzanalyseverfahren“. Bachelorarbeit. Technische Universität Dortmund.

van Meegen, Carmen, Sarah Schnackenberg und Uwe Ligges (2019). „Unequal priors in linear discriminant analysis“. In: *Journal of Classification*. Online First. ISSN: 1432-1343. DOI: 10.1007/s00357-019-09336-2.

Mitchell, Ann F. S. und Wojtek J. Krzanowski (1985). „The Mahalanobis distance and elliptic distributions“. In: *Biometrika* 72(2), S. 464–467. ISSN: 0006-3444. DOI: 10.1093/biomet/72.2.464.

Mood, Alexander M., Franklin A. Graybill und Duane C. Boes (1974). *Introduction to the Theory of Statistics*. 3. Aufl. Series in Probability and Statistics. New York: McGraw-Hill. ISBN: 978-0-07-085465-9.

Moreno-Torres, Jose G., Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla und Francisco Herrera (2012). „A Unifying View on Dataset Shift in Classification“. In:

Pattern Recognition 45(1), S. 521–530. ISSN: 0031-3203. DOI: 10.1016/j.patcog.2011.06.019.

Moreno-Torres, Jose G., Xavier Llorà, David E. Goldberg und Rohit Bhargava (2013). „Repairing fractures between data using genetic programming-based feature extraction: A case study in cancer diagnosis“. In: *Information Sciences* 222, S. 805–823. ISSN: 0020-0255. DOI: 10.1016/j.ins.2010.09.018.

Mukhopadhyay, Parimal (2009). *Multivariate Statistical Analysis*. New Jersey, London, Singapur: World Scientific. ISBN: 978-981-279-175-7. DOI: 10.1142/6744.

Narasimhamurthy, Anand und Ludmila I. Kuncheva (2007). „A Framework for Generating Data to Simulate Changing Environments“. In: *AIAP'07: Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications*. Hrsg. von V. Devedži. Anaheim, CA, USA: ACTA Press, S. 384–389. ISBN: 978-0-88986-629-4.

Narasimhan, Balasubramanian, Steven G. Johnson, Thomas Hahn, Annie Bouvier und Kiên Kiêu (2018). *cubature: Adaptive Multivariate Integration over Hypercubes*. R package version 2.0.3. URL: https://CRAN.R-project.org/package=cubature.

Pang, Shaoning, Seiichi Ozawa und Nikola Kasabov (2005a). „Chunk Incremental LDA Computing on Data Streams“. In: *Advances in Neural Networks – ISNN 2005*. Hrsg. von Jun Wang, Xiaofeng Liao und Zhang Yi. Bd. 3497. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, S. 51–56. ISBN: 978-3-540-32067-8. DOI: 10.1007/11427445_9.

— (2005b). „Incremental Linear Discriminant Analysis for Classification of Data Streams“. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* 35(5), S. 905–914. ISSN: 1083-4419. DOI: 10.1109/TSMCB.2005.847744.

Petersen, Kaare B. und Michael S. Pedersen (2012). *The Matrix Cookbook*. Version: November 15, 2012, abgerufen am: 25.06.2019. URL: http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf.

Pires, Ana M. und João A. Branco (1996). „Generalization of Fisher's linear discriminant“. In: *RECPAD'96: Proceedings of the 8th Portuguese Conference on Pattern Recognition*. Guimarães, Portugal, S. 415–418. ISBN: 972-8063-06-7.

Quiñonero-Candela, Joaquin, Masashi Sugiyama, Anton Schwaighofer und Neil D. Lawrence, Hrsg. (2009). *Dataset Shift in Machine Learning*. Neural Information Processing series. Cambridge, MA, USA: The MIT Press. ISBN: 978-0-262-17005-5. DOI: 10.7551/mitpress/9780262170055.001.0001.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rao, C. Radhakrishna (1948). „The Utilization of Multiple Measurements in Problems of Biological Classification“. In: *Journal of the Royal Statistical Society, Series B* 10(2), S. 159–203. ISSN: 1369-7412. DOI: 10.1111/j.2517-6161.1948.tb00008.x.

— (1973). *Linear Statistical Inference and its Applications*. 2. Aufl. Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN: 978-0-471-21875-3. DOI: 10.1002/9780470316436.

Reinsel, David, John Gantz und John Rydning (2018). *IDC White paper: Data Age 2025: The Digitization of the World. From Edge to Core*. #US44413318. Gesponsert von SEAGATE. International Data Corporation (IDC). URL: www.DataAge2025.com.

Rencher, Alvin C. und William F. Christensen (2012). *Methods of Multivariate Analysis*. 3. Aufl. Series in Probability and Statistics. Hoboken, New Jersey: John Wiley & Sons, Inc. ISBN: 978-0-470-17896-6. DOI: 10.1002/9781118391686.

Riedmiller, Martin und Heinrich Braun (1993). „A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm“. In: *Proceedings of the IEEE International Conference on Neural Networks*. Bd. 1. Piscataway, NJ: IEEE, S. 586–591. ISBN: 0-7803-0999-5. DOI: 10.1109/ICNN.1993.298623.

Salganicoff, Marcos (1997). „Tolerating Concept and Sampling Shift in Lazy Learning Using Prediction Error Context Switching“. In: *Artificial Intelligence Review* 11(1), S. 133–155. ISSN: 1573-7462. DOI: 10.1023/A:1006515405170.

Schlimmer, Jeffrey C. und Richard H. Granger (1986). „Incremental learning from noisy data“. In: *Machine Learning* 1(3), S. 317–354. ISSN: 1573-0565. DOI: 10.1007/BF00116895.

Schmidt, Karsten und Götz Trenkler (2006). *Einführung in die Moderne Matrix-Algebra. Mit Anwendungen in der Statistik*. 2. Aufl. Berlin, Heidelberg: Springer-Verlag. ISBN: 978-3-540-33007-3. DOI: 10.1007/3-540-33008-9.

Schnackenberg, Sarah, Uwe Ligges und Claus Weihs (2018). „Online Linear Discriminant Analysis for Data Streams with Concept Drift“. In: *Archives of Data Science*. Series A 5(1). Online First, S. 1–20. ISSN: 2363-9881. DOI: 10.5445/KSP/1000087327/02.

Shaffer, Juliet P. (1991). „The Gauss-Markov Theorem and Random Regressors“. In: *The American Statistician* 45(4), S. 269–273. ISSN: 0003-1305. DOI: 10.1080/00031305.1991.10475819.

Sharpsteen, Charlie und Cameron Bracken (2018). *tikzDevice: R Graphics Output in LaTeX Format*. R package version 0.12. URL: https://CRAN.R-project.org/package=tikzDevice.

Shimodaira, Hidetoshi (2000). „Improving predictive inference under covariate shift by weighting the log-likelihood function". In: *Journal of Statistical Planning and Inference* 90(2), S. 227–244. ISSN: 0378-3758. DOI: 10.1016/S0378-3758(00)00115-4.

Storkey, Amos J. (2009). „When training and test sets are different: characterising learning transfer". In: *Dataset Shift in Machine Learning*. Hrsg. von Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer und Neil D. Lawrence. Neural Information Processing Series. Cambridge, MA, USA: The MIT Press, S. 3–28. ISBN: 978-0-262-17005-5. DOI: 10.7551/mitpress/9780262170055.001.0001.

The Numerical Algorithms Group Ltd (2012). *NAG Library Routine Document: F07AAF (DGESV)*. abgerufen am: 06.11.2019. URL: http://www.nag.com/numeric/fl/nagdoc_fl22/xhtml/F07/f07aaf.xml.

Toussaint, Godfried T. (1974). „Bibliography on estimation of misclassification". In: *IEEE Transactions on Information Theory* 20(4), S. 472–479. ISSN: 0018-9448. DOI: 10.1109/TIT.1974.1055260.

Toutenburg, Helge (2003). *Lineare Modelle. Theorie und Anwendungen*. 2. Aufl. Heidelberg: Physica-Verlag. ISBN: 978-3-7908-1519-1. DOI: 10.1007/978-3-642-57348-4.

Tsymbal, Alexey (2004). *The problem of concept drift: definitions and related work*. Techn. Ber. Trinity College Dublin. School of Computer Science & Statistics.

Tumer, Kagan und Joydeep Ghosh (1996). „Estimating the Bayes error rate through classifier combining". In: *ICPR '96: Proceedings of 13th International Conference on Pattern Recognition*. Bd. 2. Washington, DC, USA: IEEE Computer Society, S. 695–699. ISBN: 978-0-8186-7282-8. DOI: 10.1109/ICPR.1996.546912.

Venables, William N. und Bryan D. Ripley (2002). *Modern Applied Statistics with S*. 4. Aufl. Statistics and Computing. New York: Springer. ISBN: 978-0-387-21706-2. DOI: 10.1007/978-0-387-21706-2. URL: http://www.stats.ox.ac.uk/pub/MASS4.

Wares, Scott, John Isaacs und Eyad Elyan (2019). „Data stream mining: methods and challenges for handling concept drift". In: *SN Applied Sciences* 1. Nr. 1412. ISSN: 2523-3971. DOI: 10.1007/s42452-019-1433-0.

Warnes, Gregory R., Ben Bolker und Thomas Lumley (2018). *gtools: Various R Programming Tools*. R package version 3.8.1. URL: https://CRAN.R-project.org/package=gtools.

Webb, Geoffrey I., Roy Hyde, Hong Cao, Hai Long Nguyen und Francois Petitjean (2016). „Characterizing concept drift“. In: *Data Mining and Knowledge Discovery* 30(4), S. 964–994. ISSN: 1573-756X. DOI: 10.1007/s10618-015-0448-4.

Widmer, Gerhard und Miroslav Kubat (1993). „Effective learning in dynamic environments by explicit context tracking“. In: *Machine Learning: ECML-93*. Hrsg. von Pavel B. Brazdil. Bd. 667. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, S. 227–243. ISBN: 978-3-540-47597-2. DOI: 10.1007/3-540-56602-3_139.

— (1996). „Learning in the presence of concept drift and hidden contexts“. In: *Machine Learning* 23(1), S. 69–101. ISSN: 1573-0565. DOI: 10.1007/BF00116900.

Yamazaki, Keisuke, Motoaki Kawanabe, Sumio Watanabe, Masashi Sugiyama und Klaus-Robert Müller (2007). „Asymptotic Bayesian Generalization Error when Training and Test Distributions Are Different“. In: *ICML '07: Proceedings of the 24th International Conference on Machine Learning*. Hrsg. von Zoubin Ghahramani. New York, NY, USA: ACM, S. 1079–1086. ISBN: 978-1-59593-793-3. DOI: 10.1145/1273496.1273632.

Žliobaitė, Indrė (2010). *Learning under Concept Drift: an Overview*. arXiv: 1010.4784.