Adaptive lineare Dimensionsreduktion in der Klassifikation

Dissertation

zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik
der Universität Dortmund

vorgelegt von
Karsten Lübke

Dortmund, Oktober 2006

1. Gutachter: Prof. Dr. Claus Weihs

2. Gutachter: Prof. Dr. Walter Krämer

Datum der mündlichen Prüfung: 28.11.2006

Worte des Dankes

Es ist gute Sitte sich zu Beginn bei den Menschen zu bedanken, die maßgeblich zum Gelingen eines Werkes beigetragen haben:
Da ist zunächst einmal mein Betreuer Herr Weihs zu nennen, der mich schon zu Zeiten meiner Diplomarbeit motiviert hat zu promovieren und mich im positiven Sinne gefordert und gefördert hat. Nicht nur zu Beginn meiner Zeit als Mitarbeiter am Lehrstuhl für Computergestützte Statistik haben mir sowohl Winni Theis als auch Uta Häsel viel über das Promovieren und Arbeiten an der Uni beigebracht. Diese Unijahre haben sehr viel Spass gemacht: dies ist Anja Busse, Nils Raabe, Gero Szepannek und vielen weiteren zu verdanken. Während der ganzen Zeit – und auch danach – war Uwe Ligges stets eine große Hilfe: sei es durch Kaffee, Humor oder Computerkenntnisse.
Weiterhin möchte ich Irina Czogiel danken, die diese Arbeit mit Sach- und Sprachverstand Korrektur gelesen hat. Herr Krämer als Zweitgutachter, Herr Kunert und Herr Knapp waren sofort bereit die Promotionskommision zu vervollständigen. Auch hierfür vielen Dank.
Ohne meinen – leider zu früh verstorbenen - Vater hätte ich nie Statistik studiert und diese Arbeit wäre nie geschrieben worden. Ich möchte mich bei meiner Mutter und meinen Bruder sowie meinen Freundinnen und Freunden für die Unterstützung und das Verständnis für den manchmal geistig abwesenden Wissenschaftler bedanken.
Zuletzt gilt noch ein besonderer Dank meiner Frau Christina Hauke. In all den Jahren hat sie mich immer motiviert diese Arbeit zu schreiben: „Du schaffst das schon...“ Ich habe es geschafft!

Inhaltsverzeichnis

Kapitel 1

Einleitung

In vielen wissenschaftlichen Gebieten erlangen statistische Verfahren eine wachsende Popularität und Bedeutung. Aufgrund der steigenden Leistungsfähigkeit der zur Verfügung stehenden Computer können dabei immer größere Datenmengen gesammelt und analysiert werden. Dies bringt zum einen den Vorteil, dass sich die statistische Analyse auf sehr detaillierte Informationen stützen kann, zum anderen kann eine hohe Datendimension aber auch von Nachteil sein. Zum Beispiel ist das Generieren von Hypothesen über die Daten mit Hilfe von visuellen Verfahren bei hochdimensionalen Daten nur bedingt möglich. Im Rahmen einer statistischen Analyse ist eine Dimensionsreduktion bei hochdimensionalen Daten vielfach von großem Nutzen. Mit ihrer Hilfe können die Daten oft auf zwei oder drei Dimensionen zurückgeführt werden, so dass Muster, Auffälligkeiten oder Zusammenhänge visualisiert werden können. Auch bei der eigentlichen Modellierung können durch eine vorgeschaltete Dimensionsreduktion numerische Probleme wie z.B. (fast) singuläre Matrizen umgangen werden. Im Rahmen des Sonderforschungsbereiches 475 „Komplexitätsreduktion in multivariaten Datenstrukturen“ an der Universität Dortmund wird in mehreren Projekten erforscht, mit welchen Methoden und Kriterien in komplexen und vielschichtigen Problemen angemessene Modelle sowie gute Prognosen erreicht werden können. Um aus hochdimensionalen, komplexen Datenstrukturen die wesentlichen Informationen zu extrahieren, eignen sich häufig Linearkombinationen der Original-

variablen. Sie sind einfach zu berechnen und ermöglichen in bestimmten Fällen eine inhaltliche Interpretation der Ergebnisse.

Häufig sollen im Rahmen einer statistischen Analyse Zusammenhänge zwischen erklärenden Variablen und einer nominalen Zielvariable gefunden werden. Ein solches Vorgehen der Zuordnung von Kategorien zu den Objekten wird auch Klassifikationsregel genannt. Formal ist eine Klassifikation eine Abbildung von erklärenden Variablen auf eine endliche Anzahl von vorbestimmten Kategorien. Statistische Klassifikationsregeln werden in der Praxis häufig benötigt. Eine Anwendung findet sich im Projekt B3 des SFB 475, „Multivariate Bestimmung und Untersuchung von Konjunkturzyklen“. Ein Ziel dieses Projektes ist es, mit Hilfe von makroökonomischen Daten die jeweilige Konjunkturphase zu bestimmen. Diese Anwendung ist der praktische Hintergrund dieser Arbeit.

Im Laufe der Zeit wurden viele Verfahren zur Dimensionsreduktion und zur Klassifikation vorgeschlagen. Dabei ist nicht klar, welches Verfahren in welcher Situation angewendet werden soll. Häufig werden neue Verfahren vorgeschlagen, auf einen oder mehreren Datensätzen angewendet, und aufgrund der Ergebnisse wird eine Überlegenheit des neuen Verfahrens postuliert. Dabei wird übersehen, dass es nach dem No-Free-Lunch Theorem (Wolpert, 2001) kein universell bestes Verfahren geben kann. Die Güte des Verfahrens hängt vielmehr von den zugrunde liegenden Daten ab. Daher besteht die Gefahr, dass die in einer Publikation verwendeten Daten besonders gut zu den Verfahren passen, das Verfahren speziell auf diesen Daten also eine hohe Güte hat. Die verwendeten Daten werden somit häufig über das Verfahren bestimmt. Dieses ist konträr zum Vorgehen in der Praxis, denn dort sind die Daten vorgegeben und das Ziel ist es, ein adäquates Verfahren anzuwenden.

In der vorliegenden Arbeit wird untersucht, wie aufgrund der zu untersuchenden Daten das jeweils beste Verfahren gefunden werden kann. Als Ergebnis dieser Überlegungen wird ein adaptives Verfahren zur linearen Dimensionsreduktion und Klassifikation entwickelt. Bei einem adaptiven Verfahren wird aufgrund der gegebenen Daten mit Hilfe einer so genannten Selektorstatistik entschieden, welches

bekannte Verfahren zur Anwendung kommt. Dieses Vorgehen hat zwei Vorteile. Erstens können aufwändige Fehlversuche beim Ausprobieren verschiedener Verfahren vermieden werden, und zweitens ist es bei geringer Anzahl an Beobachtungen nicht immer möglich, die Verfahren überhaupt zu evaluieren – dies ist ein Schritt, der bei einem adaptiven Verfahren für verschiedene Datensituationen schon bei der Entwicklung unternommen wurde. Daher ist es für die Entwicklung eines solchen Verfahrens notwendig zu wissen, wie die Daten bei einem Klassifikationsproblem beschaffen sein können, und wie sich die vorliegenden Daten charakterisieren lassen.

Um die Komplexität der Aufgabe zu reduzieren, werden in dieser Arbeit nur lineare Klassifikatoren und lineare Dimensionsreduktionsverfahren untersucht. Sogar im statistisch einfachen Fall, in dem die Daten multivariat normalverteilt mit klassenspezifischen Erwartungswerten und gemeinsamer Kovarianz vorliegen, ist bei mehr als drei Klassen eine optimale Dimensionsreduktion im Sinne einer minimalen Fehlklassifikationswahrscheinlichkeit nicht bekannt (McLachlan, 1992, S. 96). Gleichzeitig produziert die Standardmethode nach Fisher (1936) unnötig viele Fehlklassifikationen. In einem Beispiel von Schervish (1984) hat das Standardverfahren etwa eine Fehlerrate von 0.344, wohingegen mit optimaler Dimensionsreduktion eine Rate von 0.20 erreicht werden kann (Schervish, 1984). Da bei der vorliegenden Anwendung, der Klassifikation von Konjunkturphasen, vier Klassen (Aufschwung, Oberer Wendepunkt, Abschwung und Unterer Wendepunkt) vorliegen, wird das adaptive Verfahren konkret für den Fall von vier Klassen entwickelt. Die theoretischen Überlegungen sind aber allgemein gehalten, so dass das Vorgehen leicht auf eine beliebige Anzahl von Klassen übertragen werden kann.

Im folgenden Kapitel werden die Grundlagen von Klassifikation und Dimensionsreduktion dargelegt. Außerdem werden die verwendeten Verfahren und deren statistischen Eigenschaften allgemein vorgestellt. In Kapitel 3 wird gezeigt, wie diese Verfahren in der Praxis angewendet werden können, wobei das Hauptaugenmerk auf der Schätzung der nötigen Parameter und Matrizen liegt. Kapitel 4 beschäftigt sich mit der Frage, wie die Grundgesamtheit für Klassifikationsprobleme adäquat beschrieben werden kann. Ein Verfahren zur Entwicklung eines adaptiven Verfah-

rens zur Klassifikation wird in Kapitel 5 beschrieben. In Kapitel 6 wird das daraus resultierende adaptive Verfahren für den Fall von vier Klassen konkret vorgestellt. Dieses Verfahren wird dann zur Klassifikation von Konjunkturphasen in Kapitel 7 angewendet.

Kapitel 2

Dimensionsreduktion und Klassifikation

Ein klassisches Verfahren zur Dimensionsreduktion, welches implizit auch bei der Klassifikation mit Hilfe der linearen Diskriminanzanalyse (LDA) verwendet wird, ist das Fisher-Kriterium (Fisher, 1936). Allerdings gibt es in der Literatur Beispiele, anhand derer gezeigt wird, dass das Fisher-Kriterium in bestimmten Datensituationen nicht zu optimalen Ergebnissen bezüglich der Fehlklassifikationsrate führt (Schervish, 1984). Aus diesem Grund wurde eine Reihe von Alternativen dazu vorgeschlagen.

Innerhalb dieses Kapitels werden die Grundlagen zum Verständnis von Klassifikation und Dimensionsreduktion beschrieben. Außerdem werden Verfahren zur Dimensionsreduktion vorgestellt. Das Fisher-Kriterium als Standardverfahren wird in Abschnitt 2.2 vorgestellt, mögliche Alternativen dazu in den Abschnitten 2.3, 2.4 und 2.5. Dabei werden die Verfahren bezüglich des statistischen Hintergrundes und ihrer theoretischen Eigenschaften untersucht. Aus diesen Verfahren wird dann in den Kapiteln 5 und 6 das in der jeweiligen Situation Beste adaptiv ausgewählt.

2.1 Grundlagen und Notation

Unter dem Begriff Klassifikationsverfahren wird eine Vielzahl von Methoden zusammengefasst. Er umfasst sowohl überwachtes als auch unüberwachtes Lernen. Die Ausführungen dieser Arbeit beschränken sich auf das überwachte Lernen. Weitere in der Literatur gebräuchliche Begriffe für Klassifikation sind Diskriminanzanalyse oder Mustererkennung.

Die Zielsetzung einer Dimensionsreduktion ist, eine Abbildung zu finden, die von einem höherdimensionalen Raum in einen niederdimensionalen Raum abbildet. Dabei werden in dieser Arbeit nur endlich dimensionale Vektorräume mit Skalarprodukt betrachtet. Bei statistischen Fragestellungen kann unter Dimensionsreduktion sowohl eine Reduzierung der Anzahl von Variablen als auch eine Verringerung der Anzahl von Beobachtungen verstanden werden. Hier bezieht sich der Begriff Dimensionsreduktion auf die Anzahl von Variablen.

Folgende Notation und Bezeichnungen werden im Weiteren in dieser Arbeit verwendet. Dabei sind zur Vereinfachung der Notion die Vektoren jeweils Zeilenvektoren:

- d Dimension des Merkmalsraumes
- r Dimension des Projektionsraumes
- K Anzahl der Klassen
- N Anzahl der Beobachtungen
- $\mathcal{X}$ d-dimensionaler Zufallsvektor mit Realisierungen im $\mathbb{R}^d$ (Zeilenvektor)
- $\mathcal{K}$ Zufallsvariable mit Werten in $\{1, 2, \ldots, K\}$
- $x \in \mathbb{R}^d$ Beobachtungsvektor. Realisierung von $\mathcal{X}$
- $k \in \{1, \ldots, K\}$ Klasse von x. Realisierung von $\mathcal{K}$
- $n_k, \quad k \in \{1, \ldots, K\}, \sum_{j=1}^{K} n_j = N$, Anzahl von Beobachtungen von Klasse k

- $\pi_k, \quad k \in \{1, \ldots, K\}, \sum_{j=1}^{K} \pi_j = 1$, a priori Wahrscheinlichkeit von Klasse k
- $\Pi = \begin{pmatrix} \pi_1 & 0 & \ldots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & \pi_K \end{pmatrix} \in \mathbb{R}^{K \times K}$ Matrix mit den a priori Wahrscheinlichkeit in der Diagonalen
- $X \in \mathbb{R}^{N \times d}$ vorliegende Datenmatrix. N (zeilenweise) Beobachtungen von $\mathcal{X}$
- $\mu = E(\mathcal{X}) \in \mathbb{R}^d$ gemeinsamer Erwartungswert aller Klassen
- $\mu_k = E(\mathcal{X}|\mathcal{K} = k) \in \mathbb{R}^d$ Erwartungswertvektor der k-ten Klasse
- $\Sigma_k = Cov(\mathcal{X}|\mathcal{K} = k) \in \mathbb{R}^{d \times d}$ Kovarianzmatrix innerhalb der k-ten Klasse
- $\Xi = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_K \end{pmatrix} \in \mathbb{R}^{K \times d}$ Matrix mit den Erwartungswerten in den Zeilen
- $\Sigma_B = (\Xi - 1'_K \mu)' \Pi (\Xi - 1'_K \mu) \in \mathbb{R}^{d \times d}$ Streuungsmatrix zwischen den Erwartungswerten
- $\Sigma_W = \sum_{i=1}^{K} \pi_i Cov(\mathcal{X}|\mathcal{K} = i) \in \mathbb{R}^{d \times d}$ gepoolte Kovarianzmatrix innerhalb der Klassen

Im weiteren Verlauf der Arbeit werden bestimmte Annahmen immer wieder verwendet. Daher werden sie an dieser Stelle eingeführt.

(A1) Die a priori Wahrscheinlichkeiten sind gleich:
$\pi_1 = \cdots = \pi_K = \frac{1}{K}$.

(A2) Die Kovarianzmatrizen innerhalb der Klassen sind gleich:
$\Sigma_1 = \ldots = \Sigma_K = \Sigma_W$.

(A3) Die gemeinsame Kovarianzmatrix Σ_W ist regulär, es existiert also Σ_W^{-1}.

2.1.1 Grundlagen: Klassifikation

Ein Klassifikationsproblem liegt vor, wenn ein Objekt genau einer Klasse oder Gruppe angehört. Bei der statistischen Klassifikation wird eine Zuordnungsvorschrift von Objekten zu Klassen gesucht. Dies kann formal wie folgt formuliert werden: Die Grundgesamtheit Ω bestehe aus $K \geq 2$ Klassen $\Omega_1, \ldots, \Omega_K$, wobei gilt

$$\Omega_i \cap \Omega_j = \emptyset, \quad \forall i, j = 1, \ldots, K, i \neq j$$

und

$$\bigcup_{i=1}^{K} \Omega_i = \Omega.$$

Die Zufallsvariable $\mathcal{K}$ sei die Abbildung, die jedem $\omega \in \Omega$ die zugehörige Klasse zuweist: $\mathcal{K}(\omega) = k \Leftrightarrow \omega \in \Omega_k$. Im Folgenden wird aus Gründen der Übersichtlichkeit direkt mit $k \in \{1, \ldots, K\}$ die einem Objekt zugrundeliegende Klasse bezeichnet. Sei $\mathcal{X}$ die Zufallsvariable, die jedem $\omega \in \Omega$ einen $d < \infty$-dimensionalen Beobachtungsvektor $\mathcal{X}(\omega) = x$ zuordnet. Dabei wird in dieser Arbeit davon ausgegangen, dass $x \in \mathbb{R}^d$ ist. Die Zufallsvariable $\mathcal{X}$ ist damit d-dimensional metrisch verteilt. Andere Messniveaus (nominal oder ordinal) aller oder einzelner Komponenten des Vektors sind möglich, werden hier aber nicht behandelt.

Die beiden betrachteten Zufallsvariablen $(\mathcal{X}, \mathcal{K})$ bilden einen Zufallsvektor im $\mathbb{R}^d \times \{1, \ldots, K\}$. Die Wahrscheinlichkeitsverteilung $\mathcal{F}$ dieses Vektors ist der statistische Hintergrund des Klassifikationsproblems (Devroye *et al.*, 1996, Seite 2).

Einer statistischen Klassifikation liegt die Annahme zugrunde, dass die Verteilung der Zufallsvariable $\mathcal{X}$ von der zugehörigen Klasse k abhängt. Sei dazu

$$F_i(x) := P(\mathcal{X}(\omega) \leq x | \omega \in \Omega_i)$$

die klassenspezifische (multivariate) Verteilungsfunktion der Zufallsvariable $\mathcal{X}$ bedingt die Klasse i, und $f_i(x)$ die dazugehörende Dichtefunktion über den $\mathbb{R}^d$. Ferner sei $\pi_i := P(\Omega_i) = P(\mathcal{K} = i)$ die a priori Wahrscheinlichkeit der Klasse i, wobei gilt

$$\sum_{i=1}^{K} \pi_i = 1, \pi_i \geq 0 \quad \forall i \in \{1, \ldots, K\}.$$

Dann lässt sich die unbedingte Wahrscheinlichkeit von $\mathcal{X} = x$ nach dem Satz über die vollständige Wahrscheinlichkeit schreiben als

$$f(x) = \sum_{i=1}^{K} \pi_i f_i(x).$$

Um aufgrund einer Beobachtung x eines Objektes ω statistische Aussagen über die Klassenzugehörigkeit machen zu können, verwendet man die bedingte Wahrscheinlichkeit der Klassen gegeben die Beobachtung. Diese Wahrscheinlichkeit wird auch a posteriori Wahrscheinlichkeit genannt. Nach dem Satz von Bayes ist diese Wahrscheinlichkeit gegeben durch

$$p(k|x) = \frac{\pi_k f_k(x)}{f(x)}. \tag{2.1}$$

Dabei ist zu beachten, dass der Nenner in (2.1) für alle Klassen $k = 1, \ldots, K$ konstant ist und nur zur Normierung dient.

Definition 1 *Eine Klassifikationsregel (kurz: Klassifikator)* $\mathbf{a}$ *ist eine Abbildung, die jedem Beobachtungsvektor* $x \in \mathbb{R}^d$ *eine Klasse zuordnet*

$$\mathbf{a} : \mathbf{a}(x) : \mathbb{R}^d \to \{1, \ldots, K\}.$$

Eine Klassifikationsregel wird auch Allokationsregel (kurz: Allokation) genannt.

Ein wichtiges Kriterium zur Beurteilung eines Klassifikators ist die Fehlklassifikationswahrscheinlichkeit. Sei $c_k(\mathbf{a}) := P(\mathbf{a}(\mathcal{X}) = k|\mathcal{K} = k)$ die Wahrscheinlichkeit, dass der Klassifikator $\mathbf{a}$ ein Objekt aus Klasse k richtig klassifiziert, dann gilt für die Fehlklassifikationswahrscheinlichkeit $e(\mathbf{a})$ von $\mathbf{a}$

$$\begin{aligned} e(\mathbf{a}) &= \sum_{k=1}^{K} P(\mathcal{K} = k) P(\mathbf{a}(\mathcal{X}) \neq k|\mathcal{K} = k) \\ &= 1 - \sum_{k=1}^{K} P(\mathcal{K} = k) P(\mathbf{a}(\mathcal{X}) = k|\mathcal{K} = k) \\ &= 1 - \sum_{i=k}^{K} \pi_k c_k(\mathbf{a}). \end{aligned} \tag{2.2}$$

Vergleiche Devroye *et al.* (1996, S. 2). Daraus lässt sich folgende Definition ableiten:

Definition 2 *Der optimale Klassifikator* $\mathbf{a^o}$ *ist derjenige Klassifikator, der die Fehlklassifikationswahrscheinlichkeit minimiert:*

$$\mathbf{a^o} = \arg\min_{\mathbf{a}} e(\mathbf{a}). \tag{2.3}$$

Damit hängt $\mathbf{a^o}$ von der Verteilung $\mathcal{F}$ von $(\mathcal{X}, \mathcal{K})$ ab. Sollte diese Verteilung bekannt sein, lässt sich $\mathbf{a^o}$ berechnen (siehe Satz 1). Meistens ist die Verteilung $\mathcal{F}$ von $(\mathcal{X}, \mathcal{K})$, und damit auch $\mathbf{a^o}$, allerdings zumindest teilweise unbekannt.

Bemerkung: Man kann eine Klassifikationsregel und deren Eigenschaften auch unter entscheidungstheoretischen Gesichtspunkten untersuchen. Dabei bietet es sich an, andere Eigenschaften eines Klassifikators zu betrachten (z.B. Risiko). Dies wird hier aber nicht behandelt. Eine solche Untersuchung findet sich z.B. in den Kapiteln 1–4 von Garczarek (2002).

Viele Klassifikationsregeln lassen sich als arg max Klassifikatoren darstellen. Dabei wird zunächst für jede Klasse ein Zugehörigkeitswert mittels einer klassenweisen Diskriminanzfunktion $\mathbf{a}_k(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ berechnet. $\mathbf{a}_k(x)$ wird auch Zugehörigkeit von x zur Klasse k genannt. Mit Hilfe der K Zugehörigkeiten erfolgt die endgültige Klassifikation dann mittels

$$\mathbf{a}(x) = \arg\max_{k \in \{1,\ldots,K\}} (\mathbf{a}_1(x), \ldots, \mathbf{a}_K(x)).$$

Ein Beispiel für einen arg max Klassifikator ist die Bayes Regel:

Definition 3 *Ein Bayes Klassifikator ordnet* x *derjenigen Klasse* k *zu, deren a posteriori Wahrscheinlichkeit an der Stelle* x *maximal ist:*

$$\mathbf{a^{Bayes}} := \arg\max_{k \in \{1,\ldots,K\}} (p(1|x), \ldots, p(K|x)).$$

Die Bayes Regel ist im Falle gleicher a priori Wahrscheinlichkeiten (A1: $\pi_i = \frac{1}{K} \forall i$) identisch mit der Maximum-Likelihood Klassifikation:

Definition 4 *Die Maximum Likelihood Klassifikationsregel ordnet jede Beobachtung x derjenigen Klasse k zu, deren Dichte an der Stelle x maximal ist:*

$$\mathbf{a^{ML}} := \arg \max_{k \in \{1,\ldots,K\}} (f_1(x), \ldots, f_K(x)),$$

Der Maximum-Likelihood Klassifikator ist somit auch ein arg max Klassifikator. Trotz der formalen Einfachheit hat ein Klassifikator nach Bayes attraktive Eigenschaften:

Satz 1 *Bei bekannten Verteilungen f_k und a priori Wahrscheinlichkeiten π_k ist die Bayes Regel die optimale Regel:*

$$\mathbf{a^{Bayes}} = \mathbf{a^o}.$$

Der Beweis ist standard (siehe z.B. Mardia *et al.* (1979, Seite 306f.)). Wegen der großen Bedeutung wird er hier aber rekapituliert.

Beweis: Sei $\alpha_k(x) \geq 0$ mit $\sum_{k=1}^{K} \alpha_k(x) = 1 \forall x$ die Wahrscheinlichkeit, dass eine Klassifikationsregel $\mathbf{a}$ eine Beobachtung x der Klasse k zuordnet, dann gilt für die Wahrscheinlichkeit, dass $\mathbf{a}$ ein Objekt aus Klasse k richtig klassifiziert:

$$c_k(\mathbf{a}) = \int_{R^d} \alpha_k(x) f_k(x) dx. \tag{2.4}$$

Für die Bayes Regel haben die $\alpha_k(x)$ dabei folgende Form:

$$\alpha_k^{\text{Bayes}}(x) = \begin{cases} 1 & \text{wenn} \quad \pi_k f_k(x) = \max_{i \in \{1,\ldots,K\}} \pi_i f_i(x) \\ 0 & \text{sonst.} \end{cases} \tag{2.5}$$

Sei $\mathbf{a}$ ein Klassifikator mit $e(\mathbf{a}) < e(\mathbf{a}^{\mathbf{Bayes}})$, aber wegen (2.4) und (2.5) gilt

$$\begin{aligned}
\sum_{i=1}^{K} \pi_i c_i(\mathbf{a}) &= \sum_{i=1}^{K} \int \pi_i \alpha_i(x) f_i(x) dx \\
&\leq \sum_{i=1}^{K} \int \alpha_i(x) \max_j \pi_j f_j(x) dx \\
&= \int (\sum_{i=1}^{K} \alpha_i(x)) \max_j \pi_j f_j(x) dx \\
&= \int \max_j \pi_j f_j(x) dx \\
&= \int \sum_{i=1}^{K} \alpha_i^{\text{Bayes}}(x) \pi_i f_i(x) dx \\
&= \sum_{i=1}^{K} \left(\pi_i \int \alpha_i^{\text{Bayes}}(x) f_i(x) dx \right) \\
&= \sum_{i=1}^{K} \pi_i c_i(\mathbf{a}^{\mathbf{Bayes}}).
\end{aligned}$$

Gleichzeitig folgt aufgrund von (2.2) ein Widerspruch zu $e(\mathbf{a}) < e(\mathbf{a}^{\mathbf{Bayes}})$ und damit ist die Bayes Regel optimal. □

Ein weiteres Kriterium zur Beurteilung eines Klassifikators ist die Komplexität der durch $\mathbf{a}$ hervorgerufenen Partition des Datenraumes:

Definition 5 *Die Klassengrenze eines Klassifikators* $\mathbf{a}$ *zwischen zwei Klassen* i *und* j *ist definiert als die Menge aller Punkte* $x \in \mathbb{R}^d$*, für die gilt*

$$\{x : \mathbf{a}_i(x) = \mathbf{a}_j(x), i, j \in \{1, \ldots, K\}, i \neq j\}.$$

Für mehr als 2 Klassen gilt für die Klassengrenzen eines arg max Klassifikators:

$$\begin{aligned}
\{x \quad &: \quad \exists i, j \in \{1, \ldots, K\}, i \neq j, \text{mit:} \\
\mathbf{a}_i \quad &= \quad \mathbf{a}_j = \max_{k \in \{1,\ldots,K\}} (\mathbf{a}_1(x), \ldots, \mathbf{a}_K(x))\}.
\end{aligned}$$

Definition 6 *Ein Klassifikator wird linear genannt, falls die Klassengrenzen linear in x sind.*

Aus der Definition folgt, dass ein arg max Klassifikator insbesondere dann linear ist, wenn alle paarweisen Klassengrenzen linear sind. Dies ist dann gegeben, wenn es eine monotone Transformation $\mathbf{t} : I\!R \to I\!R$ der klassenweisen Diskriminanzfunktion $\mathbf{a}_k(x) : I\!R^d \to I\!R$ gibt, so dass $\mathbf{t} \circ \mathbf{a}_k$ linear in x ist. Wenn durch eine monotone Transformation die Linearität in x von $\mathbf{t} \circ \mathbf{a}_k(x))$ gegeben ist, folgt die Linearität mit Definition 5 auch für die Klassengrenze. Praktisch bedeutet dies zum Beispiel für einen Bayes Klassifikator, dass sich die a posteriori Wahrscheinlichkeiten so umformen lassen, dass sich lineare Funktionen für die Klassengrenzen ergeben.

Mit der Beschränkung auf lineare Klassifikatoren ist die Menge aller möglichen Klassifikatoren $\{\mathbf{a} : I\!R^d \to \{1, \ldots, K\}\}$ stark eingegrenzt. In der Praxis kann diese Einschränkung durch Basis-Erweiterungen gelockert werden: Dabei werden die gemessenen Beobachtungen x aus dem $I\!R^d$ durch eine Abbildung $h : I\!R^d \to I\!R^{d^*}$ in einem (eventuell höherdimensionalen) neuen Merkmalsraum transformiert. So können zum Beispiel im neuen Merkmalsraum Quadrate und Produkte von einzelnen Komponenten des (Original-) Vektors gebildet werden. Eine Einführung in Basis-Erweiterungen findet sich z.B. im Kapitel 5 von Hastie *et al.* (2001). Ein populäres Beispiel für die Anwendung von Basis-Erweiterungen ist die Stützvektor-Methode (SVM) bei der mit einem Kernel-Trick ein 2-Klassen Klassifikator mit großem Erfolg angewendet wird. In dieser Arbeit wird davon ausgegangen, dass x schon alle relevante Information über ω enthält, so dass ein linearer Klassifikator optimal sein kann. Daher erhöht eine Basis-Erweiterung in dieser Arbeit nur unnötig die Komplexität.

Eines der ältesten Klassifikationsverfahren in der Statistik ist die lineare Diskriminanzanalyse (LDA).

Definition 7 *Die Klassifikationsregel der linearen Diskriminanzanalyse lautet:*

$$\begin{aligned} \mathbf{a}_k^{LDA} &= \pi_k |2\pi\Sigma_W|^{-0.5} \exp(-0.5(x-\mu_k)\Sigma_W^{-1}(x-\mu_k)') \\ \wedge\mathbf{a}^{LDA} &= \arg \max_{k\in\{1,\ldots,K\}} \mathbf{a}_k^{LDA}. \end{aligned}$$

Satz 2 *Die Verteilung innerhalb der Klassen eines Klassifikationsproblems genüge den Annahmen (A2) und (A3) (siehe S. 7) sowie*

- *die Daten sind innerhalb der Klassen multivariat normalverteilt mit Erwartungswert μ_k und Kovarianzmatrix Σ_k.*

Dann ist $\mathbf{a}^{LDA}$ die Bayes Regel für das Klassifikationsproblem.

Auch der Beweis zu diesen Satz ist bekannt, wird aus Vollständigkeitsgründen aber angegeben.

Beweis: Die Dichtefunktion der multivariaten Normalverteilung lautet (siehe z.B. Mardia *et al.* (1979, S. 37))

$$f_k(x) = |2\pi\Sigma_k|^{-0.5} \exp(-0.5(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)'),$$

die sich unter (A2) umschreiben lässt zu

$$f_k(x) = |2\pi\Sigma_W|^{-0.5} \exp(-0.5(x-\mu_k)\Sigma_W^{-1}(x-\mu_k)').$$

Damit gilt für die a posteriori Verteilungen

$$p(k|x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^K \pi_k f_k(x)} = \frac{\pi_k f_k(x)}{f(x)},$$

und da der Nenner nicht von der Klasse k abhängt, sind die Diskriminanzfunktionen nach Bayes in diesem Fall:

$$\begin{aligned} \mathbf{a}_k^{\text{Bayes}} &= \pi_k |2\pi\Sigma_W|^{-0.5} \exp(-0.5(x-\mu_k)\Sigma_W^{-1}(x-\mu_k)') \\ \wedge\mathbf{a}^{\text{Bayes}} &= \arg \max_{k\in\{1,\ldots,K\}} \mathbf{a}_k^{\text{Bayes}}. \quad \square \end{aligned}$$

Daraus folgt unmittelbar mit Hilfe von Satz 1:

Korollar 1 *Unter den Annahmen aus Satz 2 ist die LDA der optimale Klassifikator:*

$$\mathbf{a}^{LDA} = \mathbf{a}^{\mathbf{o}}. \tag{2.6}$$

Die Klassifikationsregel der LDA kann auf mehrere Arten ausgedrückt werden, da folgende klassenweisen Diskrimininanzfunktionen und Zuordnungsvorschriften äquivalent sind:

$$\begin{aligned}
\mathbf{a}_k^{\mathrm{LDA}} &= \pi_k |2\pi\Sigma_W|^{-0.5} \exp(-0.5(x-\mu_k)\Sigma_W^{-1}(x-\mu_k)') \\
\wedge \mathbf{a}^{\mathrm{LDA}} &= \arg \max_{k\in\{1,\ldots,K\}} \mathbf{a}_k^{\mathrm{LDA}} \\
\Leftrightarrow \mathbf{a}_k^{*\mathrm{LDA}} &= -\log(\pi_k) + (x-\mu_k)\Sigma_W^{-1}(x-\mu_k)' && (2.7) \\
\wedge \mathbf{a}^{*\mathrm{LDA}} &= \arg \min_{k\in\{1,\ldots,K\}} \mathbf{a}_k^{*\mathrm{LDA}} && (2.8) \\
\Leftrightarrow \mathbf{a}_k^{**\mathrm{LDA}} &= \log(\pi_k) + x\Sigma_W^{-1}\mu_k' - \frac{1}{2}\mu_k\Sigma_W^{-1}\mu_k' && (2.9) \\
\wedge \mathbf{a}^{**\mathrm{LDA}} &= \arg \max_{k\in\{1,\ldots,K\}} \mathbf{a}_k^{*\mathrm{LDA}}. && (2.10)
\end{aligned}$$

Die erste Äquivalenz gilt aufgrund der Monotonie der $\exp(\cdot)$ Funktion und aus der Tatsache, dass zur Bestimmung der Klassengrenzen konstante Terme über die Klassen weggelassen werden können. Die zweite Äquivalenz ist eine Umformung aus der folgt:

Korollar 2 *Die lineare Diskriminanzanalyse ist ein linearer Klassifikator.*

Beweis: Aufgrund von (2.9) gilt für die Klassengrenze von i und j mit $i \neq j$:

$$\begin{aligned}
&\{x : x \in I\!R^d \;:\; \mathbf{a}_i^{**\mathrm{LDA}} = \mathbf{a}_j^{**\mathrm{LDA}}\} \\
=\{&x : x \in I\!R^d \;:\; \log(\pi_i) + x\Sigma_W^{-1}\mu_i' - \frac{1}{2}\mu_i\Sigma_W^{-1}\mu_i' = \\
&\qquad\qquad \log(\pi_j) + x\Sigma_W^{-1}\mu_j' - \frac{1}{2}\mu_j\Sigma_W^{-1}\mu_j'\} \\
=\{&x : x \in I\!R^d \;:\; x\left(\Sigma_W^{-1}(\mu_i-\mu_j)'\right) = \\
&\qquad\qquad \frac{1}{2}(\mu_i+\mu_j)\Sigma_W^{-1}(\mu_i-\mu_j)' - \log(\frac{\pi_i}{\pi_j}\}.
\end{aligned}$$

Da die letzte Gleichung linear in x ist, folgt nach Definition 6 das Korollar. □

Der ursprüngliche Ansatz zur Diskriminanzanalyse von Fisher (1936) bezog sich auf den 2 Klassen Fall und ist verteilungsfrei. Dabei verband er auch Dimensionsreduktion und Klassifikation. Rao (1948) erweiterte den Ansatz auf mehrere Klassen ($K \geq 2$) und verwendete für die Bestimmung der Klassengrenzen die Dichtefunktionen innerhalb der Klassen. Sollte die zweite Annahme (gleiche Kovarianzmatrizen innerhalb der Klassen) nicht erfüllt sein, ist die Bayes Regel die quadratische Diskriminanzanalyse, da die Klassengrenzen sich dann mit Hilfe von in x quadratischen Termen bestimmen lassen.

2.1.2 Grundlagen: Lineare Dimensionsreduktion

Der Begriff Dimensionsreduktion bezieht sich in dieser Arbeit auf eine Abbildung aus einem d-dimensionalen Raum in einen r-dimensionalen Raum mit $r < d$. Wie in Abschnitt 2.1.1 sind die betrachteten Räume Kreuzprodukte von reellen Zahlen, es werden also Abbildungen von $\mathbb{R}^d \rightarrow \mathbb{R}^r$ untersucht. Verfahren zur Dimensionsreduktion werden in der Statistik vielfältig verwendet:

1. Zur Visualisierung und damit zum besseren Kennenlernen der Daten (explorative Datenanalyse). Gegebenenfalls können nach einer solchen Visualisierung neue Hypothesen über die Daten generiert werden.
2. Als Vorstufe einer Analyse, um zum Beispiel Überanpassung, Schätzer mit hoher Varianz oder Singularitäten zu vermeiden. Nicht selten haben Schätzer aus reduzierten Daten bessere Eigenschaften als diejenigen, die aus den Originaldaten gewonnen werden.
3. Um einfachere Modelle zu erzeugen.
4. Zum Einsparen von Speicherplatz.

Bei der Dimensionsreduktion als Vorverarbeitung der Daten zur Klassifikation spielen insbesondere die Punkte 1 und 2 eine wichtige Rolle. Bei der Visualisierung

in 2 oder 3 Dimensionen ist es möglich, die Relationen der Klassen zueinander zu finden. Außerdem können auffällige Beobachtungen (z.B. Ausreißer) oder Hinweise für die zugrundeliegende bedingte Verteilung gesammelt werden. Das Teilgebiet der Projection Pursuit Methoden (Friedman, 1987) beschäftigt sich damit, möglichst aussagekräftige Projektionen zu finden. Überdies beinhalten viele Klassifikationsregeln das Invertieren von Matrizen. Daher werden Dimensionsreduktionsmethoden auch zur Vermeidung von (Fast-) Singularitäten angewendet. Solche Probleme treten beispielsweise bei Anwendungen in der Chemometrie (Næs & Indahl, 1998) häufig auf.

Bei der linearen Dimensionsreduktion setzen sich die Komponenten des Bildvektors $z \in \mathbb{R}^r$ aus Linearkombinationen der Komponenten des Ursprungsvektors $x \in \mathbb{R}^d$ zusammen:

$$z_i = \sum_{j=1}^{d} x_j g_{j,i}, \quad i = 1, \ldots, r, \quad g_{j,i} \in \mathbb{R} \quad \forall j = 1, \ldots, d, i = 1, \ldots, r, \tag{2.11}$$

wobei mit z_i bzw. x_i die i-te bzw j-te Komponente des jeweiligen Vektors bezeichnet sei. Die Koeffizienten $g_{i,j}$ lassen sich zu einer $d \times r$ Matrix G zusammenfassen, das heißt $z = xG$. Bei N Beobachtungen, die zeilenweise die Datenmatrix X bilden, folgt für die lineare Dimensionsreduktion, dass die ursprünglichen Daten $X \in \mathbb{R}^{N \times d}$ aus dem d-dimensionalen Raum mit Hilfe einer Projektionsmatrix $G \in \mathbb{R}^{d \times r}$ in einem r-dimensionalen Raum zu $Z = XG$ transformiert werden. Die transformierten Daten $Z \in \mathbb{R}^{N \times r}$ werden oft auch latente Faktoren genannt. Bei vielen Verfahren zur Dimensionsreduktion (siehe unten) werden außerdem noch Nebenbedingungen entweder an die Projektionsmatrix G oder an die latenten Faktoren Z gestellt. Besondere Anforderungen werden an eine Dimensionsreduktion bei der Verwendung im Rahmen eines Klassifikationsproblems gestellt. So sollte möglichst viel für die Klassifikation relevante Information im reduzierten Raum enthalten sein. Daher wurden eine Reihe von Kriterien zur Beurteilung vorgeschlagen, von denen einige im Folgenden werden.

2.2 Fisher-Kriterium

Das Ziel des Fisher-Kriteriums zur Dimensionsreduktion ist es, die Streuung zwischen den Klassen relativ zur Varianz innerhalb der Klassen im Projektionsraum zu maximieren. Dabei wird die Streuung zwischen den Klassen mittels der Streuung der Klassenmitten gemessen. Sei die Projektionsmatrix $G = (g_{\cdot,1} \cdots g_{\cdot,r})$, dann ist das Fisher-Kriterium wie folgt definiert:

Definition 8 *Das Fisher-Kriterium zur Beurteilung einer Projektionsmatrix $G \in \mathbb{R}^{d \times r}$ ist:*

$$J^F(G) = spur\left((G'\Sigma_W G)^{-1}(G'\Sigma_B G)\right). \tag{2.12}$$

Die optimale Projektionsmatrix nach Fisher ist die Matrix G, die (2.12) maximiert.

Da die Spur nicht von der Skalierung der $g_{\cdot,i}$ abhängt, ist das Ergebnis einer Matrix G in (2.12) äquivalent zu

$$J^{*\mathrm{F}}(G^*) = \mathrm{spur}(G^{*\prime}\Sigma_B G^*) \tag{2.13}$$

unter der Nebenbedingung

$$G^{*\prime}\Sigma_W G^* = I_r. \tag{2.14}$$

Dies lässt sich wie folgt zeigen. Sei $B \in \mathbb{R}^{r \times r}$ so gewählt, dass $B'G'\Sigma_W GB = I_r$ und sei $G^* = GB$. Dann gilt:

$$\begin{aligned} J^{*\mathrm{F}}(GB) &= \mathrm{spur}(B'G'\Sigma_B GB) \\ &= \mathrm{spur}\left((B'G'\Sigma_W GB)^{-1}(B'G'\Sigma_B GB)\right) \\ &= \mathrm{spur}\left(B^{-1}(G'\Sigma_W G)^{-1}(B')^{-1}B'(G'\Sigma_B G)B\right) \\ &= \mathrm{spur}\left((G'\Sigma_W G)^{-1}(G'\Sigma_B G)BB^{-1}\right)) = J^F(G). \end{aligned}$$

Daher gibt es zu jeden G ein äquivalentes G^*, das die Nebenbedingung $G^{*\prime}\Sigma_W G^* = I_r$ erfüllt.

Lemma 1 *Unter der Annahme einer regulären Matrix Σ_W folgt:*

$$rang(\Sigma_W^{-1}\Sigma_B) = rang(\Sigma_B) \leq \min(K-1, d). \tag{2.15}$$

Beweis: Folgt aus der Annahme, dass Σ_W regulär und Σ_B die Streuungsmatrix von K d-dimensionalen Vektoren ist.□
Sei nun

$$r^{\max} := \text{rang}(\Sigma_B). \tag{2.16}$$

Der erste Artikel zur Diskriminanzanalyse von Fisher (1936) bezog sich auf die Klassifikation von 2 Klassen. Dabei entwickelte Fisher auch eine Klassifikationsregel, die hier schon auf mehrere Klassen erweitert ist:

Definition 9 *Sei die Matrix $G \in I\!R^{d\times r}$ so konstruiert, dass die r Spalten das Fischer Kriterium (2.12) maximieren. Dabei sei G so normiert das gilt:*

$$G'\Sigma_W G = I_r. \tag{2.17}$$

Die Klassifikationsregel nach Fischer lautet dann:

$$\mathbf{a}_k^{Fisher} = (x-\mu_k)G((x-\mu_k)G)' = \sum_{j=1}^{r}(xg_{.,j} - \mu_k g_{.,j})^2 \tag{2.18}$$

$$\wedge\mathbf{a}^{Fisher} = \arg\min_{k\in\{1,\ldots,K\}} .\mathbf{a}_k^{Fisher} \tag{2.19}$$

Bei der Klassifikation nach Fischer werden die Beobachtungen in einen Unterraum projiziert. Danach werden die Beobachtungen derjenigen Klasse zugeordnet, deren Erwartungswert der Beobachtung im projizierten Raum nach der euklidischen Metrik am Nächsten liegt.

Satz 3 *Unter der Annahme (A1) sowie*

- $r = r^{max}$,

- *die Projektionsmatrix G sei so skaliert, dass gilt $G'\Sigma_W G = I_r$,*

gilt:

$$\mathbf{a}^{Fisher} = \mathbf{a}^{LDA} \tag{2.20}$$

Beweis siehe Kshirsagar & Arseven (1975). Aus Satz 3 folgt:

Korollar 3 *Unter den Annahmen (A1) und (A2) sowie*

- *die Daten sind innerhalb der Klassen multivariat normalverteilt mit Erwartungswert μ_k und Kovarianzmatrix Σ_k,*
- *die Projektionsmatrix G sei so skaliert, dass gilt $G'\Sigma_W G = I_r$,*

ist die Klassifikationsregel nach Fisher (Definition 9) optimal (nach Definition 2).

Beweis: Es gilt nach Satz 3 und (A1):

$$\mathbf{a}^{\text{Fisher}} = \mathbf{a}^{\text{LDA}} \tag{2.21}$$

zusammen mit der Normalverteilungsannahme und (A2) folgt mit Satz 2 und Korollar 1

$$\mathbf{a}^{\text{Fisher}} = \mathbf{a}^{o}. \quad \square \tag{2.22}$$

Korollar 4 *Es gilt*

$$\mathbf{a}^{LDA}(\mathcal{X}) = \mathbf{a}^{LDA}(\mathcal{X}G^{Fisher}) \tag{2.23}$$

und insbesondere unter den Annahmen (A1), (A2) und

- *die Daten sind innerhalb der Klassen multivariat normalverteilt mit Erwartungswert μ_k und Kovarianzmatrix Σ_k,*

gilt:

$$\mathbf{a}^{\mathbf{o}}(\mathcal{X}) = \mathbf{a}^{LDA}(\mathcal{X}G^{Fisher}) \tag{2.24}$$

Beweis: Der Beweis folgt aus dem Beweis zu Satz 3, da sich die Klassifikation nicht ändert, wenn anstelle von x die projizierte Beobachtung xG verwendet wird. □

Eine wichtige Konsequenz ist, dass ein Klassifikationsproblem in d Dimensionen ohne Informationsverlust für eine Lineare Diskriminanzanalyse in ein r dimensionales Klassifikationsproblem transformiert werden kann:

Satz 4 *Sei $r = rang(\Sigma_B)$, wobei nach Konstruktion von Σ_B gilt*

$$r \leq \min(K-1, d).$$

Ferner sei Σ_W regulär. Dann spannen die Mittelwertsvektoren $\mu_1, \cdots \mu_K$ einen r dimensionalen Raum auf. Zur Klassifikation mit Hilfe der linearen Diskriminanzanalyse ist nur die Projektion auf diesen Raum relevant.

2.3 Minimaler-Fehler-Klassifikator

Dieser Ansatz von Röhl *et al.* (2002) verfolgt das Ziel, direkt die Fehlklassifikationswahrscheinlichkeit im projizierten Raum zu minimieren.

Definition 10 *Das Minimaler-Fehler-Kriterium zur Beurteilung einer Projektionsmatrix $G \in \mathbb{R}^{d\times r}$ ist:*

$$J^{MFK}(G) = e(\mathbf{a}, G) = \sum_{k=1}^{K} P(\mathcal{K}=k)P(\mathbf{a}(\mathcal{X}G) \neq k|\mathcal{K}=k). \tag{2.25}$$

Die optimale Projektionsmatrix G nach dem Minimaler-Fehler-Kriterium ist diejenige Matrix, welche (2.25) minimiert.

Da der funktionale Zusammenhang zwischen Fehlerrate und Projektionsmatrix bei mehr als 2 Klassen (auch bei homoskedastischen, normalverteilten Klassen) mathematisch sehr komplex ist (Guseman *et al.*, 1975), ist die Bestimmung des optimalen G^{MFK} nach 2.25 sehr schwierig bis unmöglich. Einen Fall, bei dem die Bestimmung von G^{MFK} möglich ist, zeigt das folgende Lemma.

Lemma 2 *Unter den Annahmen (A1), (A2) sowie*

- *die Daten sind innerhalb der Klassen multivariat normalverteilt mit Erwartungswert μ_k und Kovarianzmatrix Σ_k,*
- *r ist die Anzahl der Eigenwerte λ_i von $\Sigma_W^{-1}\Sigma_B$ mit $\lambda_i > 0$*

gilt:

$$G^{MFK} = G^{Fisher}. \tag{2.26}$$

Außerdem ist

$$\mathbf{a}^{LDA}(\mathcal{X}G) = \mathbf{a}^{o}(\mathcal{X}). \tag{2.27}$$

Beweis: Folgt aus Satz 3 und Korrolar 4. □ Aufgrund dieser Überlegungen wird im folgenden davon ausgegangen, dass die Allokation nach der Projektion aufgrund des Minimaler-Fehler-Kriterium durch die Linearen Diskriminanzanalyse (siehe Definition 7) erfolgt.

2.4 Optimale-Separation-Projektion

Das in Luebke & Weihs (2005a) neu vorgestellte Verfahren der Optimalen-Separation-Projektion (OSP) ist eine Weiterentwicklung der Maximum-Minimum-Separations-Projektion (MMSP), welches in Luebke & Weihs (2004b) vorgeschlagen wurde.

Bei dem MMSP Verfahren wurde als Kriterium der minimale paarweise Abstand zwischen Erwartungswertvektoren maximiert, also

$$G^{\text{MMSP}} = \arg\max_{G} \text{mindist}(G), \tag{2.28}$$

wobei der minimale Abstand (mindist) definiert ist als

$$\text{mindist}(G) := \min_{i,j=1,\ldots,K, i\neq j} ||\mu_i G - \mu_j G||^2. \tag{2.29}$$

Zu beachten ist, dass unter der Nebenbedingung $G'\Sigma_W G = I_r$ der euklidische Abstand in (2.29) identisch zum Mahalanobis-Abstand im projizierten Raum ist. Dabei lehnt sich die Idee von MMSP und OSP an die Idee der Stützvektormethode (SVM, siehe z.B. Burges (1998)) an. Bei der SVM zur Klassifikation von zwei Klassen wird eine Hyperebene gesucht, so dass der Abstand der nächsten Punkte der beiden Klassen zu der Hyperebene maximal wird (maximaler Seitenrand (engl: margin)). Bei normalverteilten Daten ist der Abstand zwischen zwei Klassen durch den Mahalanobisabstand der Mittelwerte definiert. Da die Gefahr einer Fehlklassifikationen zwischen einander nahen Klassen größer ist (Hand, 1997, S. 7), wird bei MMSP der minimale Abstand maximiert.

Um das Kriterium der Optimalen Separations Projektion herzuleiten, sei zunächst folgendes Lemma betrachtet.

Lemma 3 *Unter den Annahmen (A1) bis (A3) und*

- *die Daten sind innerhalb der Klassen multivariat normalverteilt mit Erwartungswert μ_k und Kovarianzmatrix Σ_k,*

gilt

$$P(\mathbf{a}^{LDA}(\mathcal{X}) = i|\mathcal{K} = j) \leq \Phi\left(-\frac{1}{2}((\mu_i - \mu_j)\Sigma_W^{-1}(\mu_i - \mu_j)')^{\frac{1}{2}}\right), \quad i \neq j \tag{2.30}$$

wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist.

Beweis: Sei $\alpha_k(x) \geq 0$ mit $\sum_{i=1}^{K} \alpha_i(x) = 1 \forall x$ die Wahrscheinlichkeit, dass eine Klassifikationsregel eine Beobachtung x der Klasse i zuordnet, dann gilt unter Annahme (A1) (vergleiche Beweis zu Satz 1):

$$\alpha_i^{\mathrm{LDA}}(x) = \begin{cases} 1 & \text{wenn} \quad f_i(x) = \max_{j \in \{1,\ldots,K\}} f_j(x) \\ 0 & \text{sonst} \end{cases} .$$

Weiter sei

$$\alpha_{i,j}^{\mathrm{LDA}}(x) = \begin{cases} 1 & \text{wenn} \quad f_i(x) > f_j(x) \\ 0 & \text{sonst} \end{cases} ,$$

dann gilt

$$\alpha_i^{\mathrm{LDA}}(x) \leq \alpha_{i,j}^{\mathrm{LDA}}(x). \tag{2.31}$$

Und damit gilt

$$P(\mathbf{a}^{\mathrm{LDA}}(\mathcal{X}) = i | \mathcal{K} = j) = \int \alpha_i^{\mathrm{LDA}}(x) f_j(x) dx \leq \int \alpha_{i,j}^{\mathrm{LDA}}(x) f_j(x) dx.$$

$\int \alpha_{i,j}^{\mathrm{LDA}}(x) f_j(x) dx$ ist aber die Fehlklassifikationswahrscheinlichkeit einer Klasse im Zwei-Klassen Fall, welche gegeben ist durch (siehe z.B. McLachlan (1992, Seite 61))

$$\Phi\left(-\frac{1}{2}\left((\mu_1 - \mu_2)\Sigma_W^{-1}(\mu_1 - \mu_2)'\right)^{\frac{1}{2}}\right). \quad \square \tag{2.32}$$

Mit diesem Lemma lässt sich folgende Abschätzung nach Bonferroni für den Fehlklassifikationsfehler einer LDA zeigen.

Satz 5 *Es gelten die Annahmen (A1) bis (A3) und*

- *die Daten sind innerhalb der Klassen multivariat normalverteilt mit Erwartungswert μ_k und Kovarianzmatrix Σ_k.*

Außerdem sei $\delta_M^2(i,j)$ der Mahalanobis-Abstand zwischen den Klassen i und j. Dann gilt

$$e(\mathbf{a}^{LDA}) \leq \frac{K-1}{K} \sum_{i=1}^{K} \Phi\left(-\frac{1}{2} \min_{j=1,\ldots,K, j \neq i} \delta_M(i,j)\right). \tag{2.33}$$

Beweis: Es gilt für jede beliebige Klasse k

$$\begin{aligned}
P(\mathbf{a}^{\mathrm{LDA}}(\mathcal{X}) \neq k|\mathcal{K}=k) &= P(\cup_{i\neq k}\{\mathbf{a}^{\mathrm{LDA}}(\mathcal{X})=i|\mathcal{K}=k\}) \\
&= \sum_{i\neq k} P(\mathbf{a}^{\mathrm{LDA}}(\mathcal{X})=i|\mathcal{K}=k) \\
&\leq \sum_{i\neq k} \max_{i\neq k} P(\mathbf{a}^{\mathrm{LDA}}(\mathcal{X})=i|\mathcal{K}=k) \\
&= (K-1)\max_{i\neq k} P(\mathbf{a}^{\mathrm{LDA}}(\mathcal{X})=i|\mathcal{K}=k) \\
&= (K-1)\max_{i\neq k} \Phi\left(-\frac{1}{2}((\mu_i-\mu_k)\Sigma_W^{-1}(\mu_i-\mu_k)')^{\frac{1}{2}}\right) \\
&= (K-1)\Phi\left(-\frac{1}{2}\min_{i\neq k}\delta_M(k,i)\right),
\end{aligned}$$

wobei die erste Ungleichung aufgrund der Abschätzung nach Bonferroni folgt und die dritte Ungleichung mit Hilfe von Lemma 3. Insgesamt ist mit

$$e(\mathbf{a}) = \sum_{i=1}^{K} \pi_i P(\mathbf{a}(\mathcal{X}) \neq i|\mathcal{K}=i)$$

die Behauptung gezeigt. □

Aus diesem Satz folgt, dass das Maximum-Minimum-Separations-Projektion Kriterium eine höhere obere Schranke für die LDA Fehlklassifikationswahrscheinlichkeit

$$e(\mathbf{a}^{\mathrm{LDA}}) \leq (K-1)\Phi\left(-\frac{1}{2}\min_{i,j=1,\ldots,K,i\neq j}\delta_M(i,j)\right)$$

minimiert, sofern $\delta_M(i,j)$ nach einer Projektion mit G berechnet wird. Dies liegt daran, dass in (2.29) nur der minimale Klassenabstand zwischen zwei Klassen verwendet wird und nicht für jede Klasse der Abstand zur jeweils nächsten. Allerdings wird diese Schranke eventuell unnötig hoch sein: ist zum Beispiel die größte einzelne Fehlklassifikationswahrscheinlichkeit zwischen zwei Klassen größer als $\frac{1}{K-1}$, so folgt, dass der Wert der Schranke größer als Eins ist, eine triviale obere Schranke für die Fehlklassifikationswahrscheinlichkeit. Diese Überlegungen führen zu einem neuen Kriterium:

Definition 11 *Sei*

$$\delta_M^2(i,j|G) := ((\mu_i - \mu_j)G)\,(G'\Sigma_W G)^{-1}\,((\mu_i - \mu_j)G)' \tag{2.34}$$

der Mahalanobis-Abstand zwischen den Klassen i und j im projizierten Raum, dann ist das Optimale-Separations-Projektionskriterium gegeben durch

$$J^{OSP}(G) = \frac{K-1}{K}\sum_{i=1}^{K}\Phi\left(-\frac{1}{2}\min_{j=1,\ldots,K,j\neq i}\delta_M(i,j|G)\right). \tag{2.35}$$

Die optimale Projektionsmatrix G nach dem Optimalen-Separations-Kriterium ist die Projektionsmatrix, die (2.35) minimiert.

Um die Berechnung des Mahalanobis-Abstandes zu vereinfachen, kann die Nebenbedingung

$$G'\Sigma_W G = I_r \tag{2.36}$$

verwendet werden. Die Nebenbedingung $G'\Sigma_W G = I_r$ impliziert keine Einschränkung der Allgemeinheit, da es für jedes G mit vollem Spaltenlang eine reguläre Matrix $B \in \mathbb{R}^{r\times r}$ gibt, so dass $(GB)'\Sigma_W(GB) = I_r$ gilt und aufgrund von

$$\begin{aligned}
\delta_M(i,j|G) &= ((\mu_i - \mu_j)G)\,(G'\Sigma_W G)^{-1}\,((\mu_i - \mu_j)G)' \\
&= ((\mu_i - \mu_j)G)\left(B'^{-1}B'G'\Sigma_W GBB^{-1}\right)^{-1}((\mu_i - \mu_j)G)' \\
&= ((\mu_i - \mu_j)G)\,(B^{-1})^{-1}I_r(B'^{-1})^{-1}\,((\mu_i - \mu_j)G)' \\
&= ((\mu_i - \mu_j)GB)\,((\mu_i - \mu_j)GB)' = \delta_M(i,j|GB) = \delta_E(i,j|GB)
\end{aligned}$$

ändert sich der Mahalanobis-Abstand unter nicht-singulären Transformationen nicht. Unter der Nebenbedingung ist der Mahalanobisabstand gleich dem Euklidischen-Abstand der Mittelwerte.

2.5 Vorhersageoptimale-Projektion

Bei der Vorhersageoptimalen-Projektion wird die enge Verbindung zwischen den Projektionen einer Klassifikation und einer Regression ausgenutzt. Aufgrund dieser Verbindung kann ein prognoseorientiertes Regressionsverfahren (siehe Luebke &

Weihs (2003) oder Luebke & Weihs (2004a)) für die Klassifikation übertragen werden. Die Verbindung einer Klassifikation nach Fisher und Regression ist für zwei Klassen sehr einfach (siehe z.B. Duda *et al.* (2001, Seite 240ff.)), aber bei mehr als zwei Klassen nicht mehr trivial. Für die Probleme bei der Übertragung siehe Hastie *et al.* (1994, Kapitel 5).

Die Beziehung der Projektionen von linearer Diskriminanzanalyse zu kanonischer Korrelationsanalyse (siehe zum Beispiel Mardia *et al.* (1979, Kapitel 10)) und optimal Scoring (siehe De Leeuw *et al.* (1976) oder Gifi (1990, Seite 243f.)) wurde nach der grundlegenden Arbeit von Breiman & Ihaka (1984) ausführlich in Artikeln von Hastie *et al.* (1994) und Hastie *et al.* (1995) dargestellt, siehe auch Ripley (1996). Über diese Beziehungen lassen sich Erkenntnisse, die bei Regressionsverfahren gewonnen werden, auch in der Klassifikation einsetzen. Leider ist der Beweis in Hastie *et al.* (1995) zur Verknüpfung von Klassifikation mit der kanonischen Korrelationsanalyse falsch, da sie die Quadratsummen der Klassenvariable nicht zentrieren. Ein fehlerfreier Beweis findet sich in Barker & Rayens (2003), der aber nur für Beobachtungen und nicht für Populationen geführt wurde.

2.5.1 Notation und algebraische Grundlagen

Um die Verbindung von Regression, Klassifikation und kanonischer Korrelationsanalyse für die zugrundeliegende Population zu verdeutlichen, sei der Zufallsvektor $\mathcal{Y}^* \in I\!R^{K-1}$ der Indexvektor der Klasse, also

$$\mathcal{Y}_k^*(\omega) = \begin{cases} 1 & \text{wenn } \mathcal{K}(\omega) = k, \quad k \in \{1, \ldots, K-1\} \\ 0 & \text{sonst} \end{cases} .$$

Der Indexvektor der Klasse wird nur bis zur Klasse $K-1$ gebildet, da ansonsten die Kovarianz singulär wäre. Das Ergebnis für die Verbindung zur Klassifikation ändert sich nicht, wenn anstelle dessen

$$\mathcal{Y}_k = \begin{cases} 1 & \text{wenn } \mathcal{K} = k, \quad k \in \{1, \ldots, K\} \\ 0 & \text{sonst} \end{cases}$$

verwendet und mit einer generalisierten Inversen gerechnet wird, siehe dazu Barker & Rayens (2003). Es gilt $E(\mathcal{Y}^*) = (\pi_1 \cdots \pi_{K-1})$ und wegen

$$\Sigma_{\mathcal{Y}^*} := Cov(\mathcal{Y}^*) = E(\mathcal{Y}^{*\prime}\mathcal{Y}^*) - E(\mathcal{Y}^*)'E(\mathcal{Y}^*)$$

lässt sich $\Sigma_{\mathcal{Y}^*}$ wie folgt berechnen: Bezeichne Π^* die Diagonalmatrix mit den $K-1$ a priori Wahrscheinlichkeiten. Dann gilt

$$E(\mathcal{Y}^{*\prime}\mathcal{Y}^*) = \sum_{i}^{K-1} \pi_i E(\mathcal{Y}^{*\prime}\mathcal{Y}^* | \mathcal{K} = i)$$

und

$$\Sigma_{\mathcal{Y}^*} = \Pi^* - \Pi^* 1'_{K-1} 1_{K-1} \Pi^*. \tag{2.37}$$

Daraus folgt

$$\Sigma_{\mathcal{Y}^*}^{-1} = \frac{1}{\pi_K} 1'_{K-1} 1_{K-1} + (\Pi^*)^{-1}, \tag{2.38}$$

da gilt

$$\begin{aligned}
\Sigma_{\mathcal{Y}^*}\Sigma_{\mathcal{Y}^*}^{-1} &= \left(\Pi^* - \Pi^* 1_{K-1} 1'_{K-1} \Pi^*\right) \left(\frac{1}{\pi_K} 1_{K-1} 1'_{K-1} + (\Pi^*)^{-1}\right) \\
&= \frac{1}{\pi_K} \Pi^* 1_{K-1} 1'_{K-1} - \frac{1}{\pi_K} \Pi^* 1_{K-1} 1'_{K-1} \Pi^* 1_{K-1} 1'_{K-1} \\
&\quad + \Pi^* (\Pi^*)^{-1} - \Pi^* 1_{K-1} 1'_{K-1} \Pi^* (\Pi^*)^{-1} \\
&= I_{K-1} + \frac{1}{\pi_K} \Pi^* 1_{K-1} 1'_{K-1} \\
&\quad - \frac{1}{\pi_K} \Pi^* 1_{K-1} \underbrace{1'_{K-1} \Pi^* 1_{K-1}}_{=1-\pi_K} 1'_{K-1} - \Pi^* 1_{K-1} 1'_{K-1} \\
&= I_{K-1} + \frac{1}{\pi_K} \Pi^* 1_{K-1} 1'_{K-1} \\
&\quad - \frac{1-\pi_K}{\pi_K} \Pi^* 1_{K-1} 1'_{K-1} - \Pi^* 1_{K-1} 1'_{K-1} \\
&= I_{K-1}.
\end{aligned}$$

Für die folgenden Betrachtungen sei o.B.d.A. $\mu = 0$ (ansonsten betrachte $\mathcal{X}^c := \mathcal{X} - E(\mathcal{X})$). Für

$$\Sigma_{\mathcal{X}} := Cov(\mathcal{X}) = E(\mathcal{X}'\mathcal{X})$$

gilt

$$\begin{aligned}
\Sigma_{\mathcal{X}} &= \sum_{i=1}^{K} \pi_i E(\mathcal{X}'\mathcal{X}|\mathcal{K}=i) \\
&= \sum_{i=1}^{K} \pi_i E((\mathcal{X}-\mu_i+\mu_i)'(\mathcal{X}-\mu_i+\mu_i)|\mathcal{K}=i) \\
&= \sum_{i=1}^{K} \pi_i \left[E((\mathcal{X}-\mu_i)'(\mathcal{X}-\mu_i)|\mathcal{K}=i) + E((\mathcal{X}-\mu_i)'\mu_i|\mathcal{K}=i)\right. \\
&\quad \left. +E(\mu_i'(\mathcal{X}-\mu_i)|\mathcal{K}=i) + \mu_i'\mu_i\right] \\
&= \sum_{i=1}^{K} \pi_i \left[E((\mathcal{X}-\mu_i)'(\mathcal{X}-\mu_i)|\mathcal{K}=i) + (E(\mathcal{X}|\mathcal{K}=i)-\mu_i)'\mu_i)\right. \\
&\quad \left. +\mu_i'(E(\mathcal{X}|\mathcal{K}=i)-\mu_i) + \mu_i'\mu_i\right] \\
&= \sum_{i=1}^{K} \pi_i \left[E((\mathcal{X}-\mu_i)'(\mathcal{X}-\mu_i)|\mathcal{K}=i) + \mu_i'\mu_i\right] \\
&= \Sigma_W + \Sigma_B =: \Sigma_T.
\end{aligned}$$

Mit $\mu = 0$ gilt:

$$\begin{aligned}
\Sigma_{\mathcal{X}'\mathcal{Y}^*}: &= Cov(\mathcal{X}'\mathcal{Y}^*) = E(\mathcal{X}'(\mathcal{Y}^* - E(\mathcal{Y}^*)) = E(\mathcal{X}'\mathcal{Y}^*) - E(\mathcal{X}')E(\mathcal{Y}^*) \\
&= E(\mathcal{X}'\mathcal{Y}^*) = \sum_{i}^{K-1} \pi_i E(\mathcal{X}'\mathcal{Y}^*|\mathcal{K}=i) \\
&= \Xi^{*\prime}\Pi^* = Cov(\mathcal{Y}^{*\prime}\mathcal{X})' =: \Sigma'_{\mathcal{Y}^{*\prime}\mathcal{X}}
\end{aligned}$$

mit $\Xi^* = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_{K-1} \end{pmatrix}$. Daraus folgt, dass

$$\begin{aligned}
\Sigma_{\mathcal{X}'\mathcal{Y}^*}\Sigma_{\mathcal{Y}^*}^{-1}\Sigma_{\mathcal{Y}^{*\prime}\mathcal{X}} &= (\Xi^{*\prime}\Pi^*)(\frac{1}{\pi_K}1'_{K-1}1_{K-1} + (\Pi^*)^{-1})(\Pi^*\Xi^*) \\
&= \Xi^{*\prime}\Pi^*\Xi^* + \frac{1}{\pi_K}\Xi^{*\prime}\Pi^*1'_{K-1}\underbrace{1_{K-1}\Pi^*\Xi^*}_{-\pi_K\mu_K} \\
&= \sum_{i=1}^{K-1} \pi_i\mu_i'\mu_i + \pi_K\mu_K'\mu_K \\
&= \Sigma_B.
\end{aligned}$$

Zusammenfassend gilt also:

$$\Sigma_{\mathcal{X}} = \Sigma_T = \Sigma_W + \Sigma_B \tag{2.39}$$

und

$$\Sigma_{\mathcal{X}'\mathcal{Y}^*}\Sigma_{\mathcal{Y}^*}^{-1}\Sigma_{\mathcal{Y}^{*\prime}\mathcal{X}} = \Sigma_B. \tag{2.40}$$

Um die Verbindung von Regression und Klassifikation zu verdeutlichen, wird zunächst der Zusammenhang von Klassifikation nach Fischer und einer kanonischen Korrelationsanalyse betrachtet.

2.5.2 Verbindung von Klassifikation nach Fisher und kanonischer Korrelationsanalyse

Ziel einer kanonischen Korrelationsanalyse ist es, Zusammenhänge zwischen zwei Merkmalsgruppen zu erkennen.

Definition 12 *Bei einer kanonischen Korrelationsanalyse werden nacheinander diejenigen Vektoren $g_{\cdot,i} \in \mathbb{R}^d$ und $h_{\cdot,i} \in \mathbb{R}^{K-1}$ gesucht, die die Korrelation zwischen $\mathcal{Y}^* h_{\cdot,i}$ und $\mathcal{X} g_{\cdot,i}$ maximieren. Dieses Problem ist äquivalent zu*

$$\begin{aligned} &c_i := \max_{g,h} g'_{\cdot,i}\Sigma_{\mathcal{X}'\mathcal{Y}^*}h_{\cdot,i} \quad \textit{unter der Nebenbedingung} \\ &g_{\cdot,i}\Sigma_{\mathcal{X}}g_{\cdot,i} = 1 \textit{ und } h'_{\cdot,i}\Sigma_{\mathcal{Y}^*}h_{\cdot,i} = 1. \end{aligned} \tag{2.41}$$

Weiterhin seien $g'_{\cdot,i}\Sigma_{\mathcal{X}}g_{\cdot,j} = 0$ sowie $h'_{\cdot,i}\Sigma_{\mathcal{Y}^*}h_{\cdot,j} = 0$ falls $i \neq j$ und $c_1 \geq c_2 \geq \cdots \geq c_r$.

Vergleiche Mardia *et al.* (1979, Seite 282f.).
Das folgende Korollar wird zur Übertragung von kanonischer Korrelationsanalyse und optimal Scoring benötigt:

Korollar 5 *Bei festem $h_{\cdot,i}$ ist die i-te Projektion von $\mathcal{X}$ einer kanonischen Korrelationsanalyse gegeben durch*

$$g_{\cdot,i} = \arg\max_{g} = g'\Sigma_{\mathcal{X}'\mathcal{Y}^*}h_{\cdot,i}. \tag{2.42}$$

Die Berechnung von g erfolgt mit Hilfe einer Eigenwertzerlegung:

Satz 6 *Die Vektoren $g_{\cdot,i}$, die (2.41) maximieren, sind die zum Eigenwert c_i gehörenden Eigenvektoren von $\Sigma_{\mathcal{X}}^{-1}\Sigma_{\mathcal{X}'\mathcal{Y}^*}\Sigma_{\mathcal{Y}^*}^{-1}\Sigma_{\mathcal{Y}^*\mathcal{X}}$.*

Beweis: Siehe Seber (1984, Satz A7.7, Seite 527).
Insgesamt gelten folgende Gleichungen für eine kanonische Korrelationsanalyse (Hastie *et al.*, 1995):

$$\Sigma_{\mathcal{Y}^*}^{-1}\Sigma_{\mathcal{Y}^*\mathcal{X}}\Sigma_{\mathcal{X}}^{-1} = H'CG \tag{2.43}$$

$$H'\Sigma_{\mathcal{Y}^*}H = I_r \tag{2.44}$$

$$G'\Sigma_{\mathcal{X}}G = I_r, \tag{2.45}$$

wobei C eine Diagonalmatrix mit den c_i aus (2.41) in der Diagonalen ist. Aus einer generalisierten Eigenwertzerlegung folgt:

$$\Sigma_{\mathcal{X}}^{-1}\Sigma_{\mathcal{X}'\mathcal{Y}}H = GC \tag{2.46}$$

$$G'\Sigma_{\mathcal{X}'\mathcal{Y}^*}\Sigma_{\mathcal{Y}^*}^{-1}\Sigma_{\mathcal{Y}^*\mathcal{X}}G = C^2. \tag{2.47}$$

Die Vektoren $g_{\cdot,i}$ die das Fisher-Kriterium (siehe Definition 8) maximieren, sind die Eigenvektoren von

$$\Sigma_W^{-1}\Sigma_B. \tag{2.48}$$

(siehe unten, Satz 10), die optimalen $g_{\cdot,i}$ der kanonischen Korrelationsanalyse sind nach Satz 6 und den Gleichungen (2.39) und (2.40) die Eigenvektoren von

$$\Sigma_{\mathcal{X}}^{-1}(\Sigma_{\mathcal{X}'\mathcal{Y}^*}\Sigma_{\mathcal{Y}^*}^{-1}\Sigma_{\mathcal{Y}^*\mathcal{X}}) = \Sigma_T^{-1}\Sigma_B. \tag{2.49}$$

Daraus folgt:

Korollar 6 *Die Projektionsvektoren g für eine Klassifikation nach Fischer und einer kanonischen Korrelationsanalyse sind bis auf die Skalierung identisch.*

Beweis: Es gilt:

$$\begin{aligned}
& \Sigma_T^{-1}\Sigma_B g_{\cdot,i} = c_i g_{\cdot,i} \\
\Leftrightarrow\ & \Sigma_B g_{\cdot,i} = c_i \Sigma_T g_i = c_i \Sigma_B g_{\cdot,i} + c_i \Sigma_W g_{\cdot,i} \\
\Leftrightarrow\ & (1-c_i)\Sigma_B g_{\cdot,i} = c_i \Sigma_W g_{\cdot,i} \\
\Leftrightarrow\ & \Sigma_B g_{\cdot,i} = \frac{c_i}{1-c_i}\Sigma_W g_{\cdot,i} \\
\Leftrightarrow\ & \Sigma_W^{-1}\Sigma_B g_{\cdot,i} = \frac{c_i}{1-c_i} g_{\cdot,i}. \quad \square
\end{aligned}$$

Aufgrund von (2.47) und (2.40) gilt

$$G'\Sigma_B G = C^2, \tag{2.50}$$

woraus mit (2.39) und der Nebenbedingung $G'\Sigma_T G = I$ folgt, dass

$$G'\Sigma_W G = I - C^2. \tag{2.51}$$

Insgesamt gilt mit (2.51) zur Umrechnung der Projektionsmatrizen

$$G^{\text{Fisher}} = G(I - C^2)^{-\frac{1}{2}}. \tag{2.52}$$

Also kann der Unterschied zwischen einer Klassifikation nach Fischer und einer kanonischen Korrelationsanalyse durch die unterschiedlichen Nebenbedingungen, $G'\Sigma_T G = I_r$ bzw. $G'\Sigma_W G = I_r$ beschrieben werden. Dabei legen Σ_T bzw. Σ_W im r-dimensionalen Bildraum die Metrik fest, mit der die Abstände für eine Zuordnung berechnet werden.

2.5.3 Verbindung Optimal Scoring und Klassifikation

Optimal Scoring ist eine Erweiterung einer klassischen Regression, in der den endogenen Variablen Scores zugewiesen werden, welche dann auf die exogenen Variablen regressiert werden.

Definition 13 *Bei einem optimalen Scoring Problem wird der erwartete quadratische Fehler, welcher gegeben ist durch*

$$MSE(H, M) = E\left(||\mathcal{Y}^* H - \mathcal{X} M||^2\right), \tag{2.53}$$

wobei

- $H \in \mathbb{R}^{(K-1)\times r}$ *die Score Matrix der Klassen,*
- $M \in \mathbb{R}^{d\times r}$ *die Regressionsmatrix und*
- $\|\cdot\|$ *die Frobenius Norm für Matrizen ist*

minimiert. Um triviale Lösungen zu vermeiden, wird an die Score-Matrix H die Nebenbedingung gestellt, dass die resultierenden Scores $\mathcal{Y}^* H$ unkorreliert mit Varianz 1 *sind, also*

$$H' \Sigma_{\mathcal{Y}^*} H = I_{r-1}. \tag{2.54}$$

Vergleiche mit dem mittleren quadratischen Residuum von Hastie *et al.* (2001, Seite 392).
Bei fixem H gilt für das optimale unverzerrte M (siehe z.B. Stapleton (1995, Seite 53)):

$$M = \left(E(\mathcal{X}'\mathcal{X})\right)^{-1} E(\mathcal{X}'\mathcal{Y}^*) H = \Sigma_{\mathcal{X}}^{-1} \Sigma_{\mathcal{X}'\mathcal{Y}^*} H. \tag{2.55}$$

Zusammenfassend gilt daher:

Satz 7 *Die Projektionsmatrizen, die im Rahmen von*

- *Diskriminanzanalyse nach Fisher*
- *kanonischer Korrelationsanalyse*

- *optimal Scoring*

berechnen werden, sind bis auf Skalierungen identisch.

Beweis: Die Äquivalenz von Fisher und kanonischer Korrelationsanalyse wurde in Korollar 6 bewiesen. Aufgrund von (2.46) und (2.55) gilt:

$$M = GC \tag{2.56}$$

und mit (2.52) folgt insgesamt

$$G^{\text{Fisher}} = M \left(C^2(1-C^2)\right)^{-\frac{1}{2}}. \quad \square \tag{2.57}$$

Während beim optimalen Scoring keine Nebenbedingung an die Projektionsmatrix gestellt wird, sind die Nebenbedingungen bei der kanonischen Korrelationsanalyse und bei Fisher unterschiedlich:

- M^{OS} beliebig
- $G'^{\text{Fisher}} \Sigma_W G^{\text{Fisher}} = I$
- $G'^{\text{CCA}} \Sigma_T G^{\text{CCA}} = I$

Das im optimalen Scoring gefundene M kann nach (2.57) so umskaliert werden, dass es den Nebenbedingungen des Fisher-Kriteriums genügt. Daher kann die Klassifikation der Objekte aufgrund der Projektionen $\mathcal{X}M$ erfolgen.

2.5.4 Vorhersageorientierung

Der Vorteil des Umwegs über die Regression zur Klassifikation ist die dadurch gewonnene Flexibilität: Anstelle der normalen Regression (2.55) können andere Verfahren zur Schätzung von M verwendet werden. Es ist bekannt, dass die Vorhersagen einer solchen multiplen, multivariaten Regression schlecht sein können, wenn die erklärenden Variablen $\mathcal{X}$ kollinear sind, oder wenn die Anzahl der Beobachtungen n

nicht viel größer ist als die Anzahl der zu schätzenden Parameter $d \times r$ (Helland & Almøy, 1994). Dies kann zum Beispiel durch Überanpassung oder nicht stabile Schätzer verursacht werden. Eine Möglichkeit, diese Probleme zu lösen, ist eine Regression mit reduziertem Rang (engl.: Reduced Rank Regression, RRR, siehe zum Beispiel Reinsel & Velu (1998)). Bei einer Regression mit reduziertem Rang werden die erklärenden Variablen $\mathcal{X}$ auf wenige latente Faktoren projiziert, welche dann als Regressoren für die abhängigen Variablen verwendet werden. Also

$$M = GB,$$

wobei

- $M \in I\!R^{d \times r}$ die Regressionsmatrix,
- $G \in I\!R^{d \times s}$ mit $s \leq r$ die Projektionsmatrix,
- $B \in I\!R^{s \times r}$ die Regressionsmatrix auf die latenten Faktoren

ist und damit

$$(\mathcal{X}G)B = \mathcal{X}(GB) = \mathcal{X}M.$$

Es gibt viele Möglichkeiten, eine solche Projektion G zu finden (Schmidli, 1995), aber es ist nicht klar, welche Methode für die jeweils vorliegenden Daten verwendet werden soll. Vielen Methoden zur Bestimmung von G ist gemein, dass die Nebenbedingung

$$(\mathcal{X}G)'(\mathcal{X}G) = I \tag{2.58}$$

verwendet wird. Der Schätzer für B ist dann der gewöhnliche kleinste Quadrate Schätzer der Projektion von $\mathcal{Y}H$ auf $\mathcal{X}G$, nämlich

$$\hat{B} = ((\mathcal{X}G)'(\mathcal{X}G))^{-1}(\mathcal{X}G)'\mathcal{Y}H = (\mathcal{X}G)'\mathcal{Y}H. \tag{2.59}$$

Gleichzeitig entfällt aufgrund der Nebenbedingung (2.58) in (2.59) die Invertierung der Matrix $\mathcal{X}G$. So wird das Problem der Multikollinearität und dabei insbesondere

auch von numerischen Schwierigkeiten bei der Berechnung der Inversen in (2.55) umgangen.

Bei einer vorhersageorientierten Betrachtung wird die Performanz einer Schätzmethode bei zukünftigen Beobachtungen untersucht: Angenommen, die Projektionsmatrizen M und H werden aufgrund von n Realisierungen X, Y der Zufallsvariablen $\mathcal{X}, \mathcal{Y}$ geschätzt. (Anstelle von $\mathcal{Y}^*$ kann ohne Einschränkung der Gültigkeit $\mathcal{Y}$ verwendet werden, sofern $r < K$ gilt.) Die Güte der Schätzung kann dann mit Hilfe von n_0 zukünftigen Realisierungen X_0, Y_0 untersucht werden. Die Punktprognose für $Y_0 H$ ist

$$\widehat{Y_0 H} = X_0 \hat{M}_{X,Y} \hat{H}_{X,Y}. \tag{2.60}$$

Dann ist der Verlust gegeben durch

$$L = \frac{1}{n_0} \| Y_0 H - \widehat{Y_0 H} \|^2. \tag{2.61}$$

Der Verlust (2.61) ist eine Transformation des mittleren Bestimmtheitsmaßes (R^2) der Zielvariablen YH (Schmidli, 1995, S. 23), wobei das R^2 aufgrund von zukünftigen Beobachtungen bestimmt wird.

$$R^2_{\text{mittel}} = 1 - \frac{L}{r} \tag{2.62}$$

Beachte, dass die Varianz innerhalb der einzelnen Zielvariablen nach (2.54) eins ist. Üblicherweise ist man aber nicht nur an der Güte der Schätzer für einige Beobachtungen, sondern an der generellen bzw. erwarteten Güte für alle Beobachtungen interessiert. Daher wird häufig der mittlere erwartete Vorhersagefehler (Schmidli, 1995, S. 24) betrachtet (engl.: Mean Squared Error of Prediction):

$$\begin{aligned} MSEP &= \frac{1}{n_0} E_{Y|X} E_{Y_0|X_0} \| (Y_0 H - \widehat{Y_0 H}) \|^2 \\ &= \frac{1}{n_0} E_{Y|X} E_{Y_0|X_0} \| (Y_0 \hat{H}_{X,Y} - X_0 \hat{M}_{X,Y} \hat{H}_{X,Y}) \|^2 \\ &= \frac{1}{n_0} E_{Y|X} E_{Y_0|X_0} \| (Y_0 \hat{H}_{X,Y} - X_0 (\hat{G}_{X,Y} \hat{G}'_{X,Y} X' (Y \hat{H}_{X,Y}))) \|^2. \end{aligned} \tag{2.63}$$

Aus Gleichung (2.63) folgt, dass der *MSEP* als eine Funktion der Projektionsmatrizen G und H aufgefasst werden kann. Somit können mit verschiedenen G und H– unter Berücksichtigung der Nebenbedingung – verschiedene *MSEP* erreicht werden.

Damit folgt als Kriterium für eine vorhersageoptimale Klassifikation:

Definition 14 *Bei der Vorhersageoptimalen-Projektion wird das Kriterium*

$$J^{VOP}(G,H) = E_{Y|X}E_{Y_0|X_0}\|(Y_0H_{X,Y} - X_0(G_{X,Y}G'_{X,Y}X'(YH_{X,Y})))\|^2) \quad (2.64)$$

zweier Projektionsmatrizen $G \in \mathbb{R}^{d\times r}$ und $H \in \mathbb{R}^{K\times r}$ unter den Nebenbedingungen $(XG)'(XG) = I$ bzw. $(YH)'(YH) = I$ minimiert.

Definition 15 *Sei*

$$\tilde{M} = G(XG)'YH\left(C^2(1-C^2)\right)^{-\frac{1}{2}}, \quad (2.65)$$

wobei G und H nach (2.64) und C die Diagonalmatrix mit den c_i aus (2.41) der Diagonalen ist. Weiter sei

$$\eta_i = \mu_i\tilde{M}. \quad (2.66)$$

Dann erfolgt die Klassifikation von Beobachtungen x, unter Berücksichtigung des VOP Kriteriums mit

$$\mathbf{a}_k^{VOP} = (x\tilde{M} - \eta_k)^2 \quad (2.67)$$

$$\wedge\mathbf{a}^{VOP} = \arg\min_{k\in\{1,\ldots,K\}} \mathbf{a}_k^{VOP}. \quad (2.68)$$

Insgesamt wurde also eine Klassifikationsregel hergeleitet, deren Berechnung keine Verteilungsannahmen beinhaltet. Trotzdem lassen sich Optimalitätseigenschaften zeigen:

Satz 8 *Unter den Annahmen (A1) und (A2) sowie*

- *die Daten sind innerhalb der Klassen multivariat normalverteilt mit Erwartungswert μ_k und Kovarianzmatrix Σ_k,*

ist die Klassifikation nach (2.67) optimal.

Beweis: Aufgrund von (2.57) und (2.59) ist $\tilde{M}$ nach (2.65) äquivalent zu einer Projektion nach Fisher (Definition 8) und damit ist (2.67) identisch zu (2.18). Dann folgt die Behauptung mittels Korollar 4. □

Auch für die Klassifikation nach Projektion über das Vorhersage-Optimale-Projektion-Kriterium gibt es eine Aussage über die maximal notwendige Dimension vergleichbar zu Satz 4.

Lemma 4 *Sei $p \leq d$ der Rank von X. Weiterhin sei $r = \min(p, K-1)$.*

1. *Falls $K - 1 \geq p$ gilt, dann gibt es kein $\tilde{r} > r$ mit $(X\tilde{G})'(X\tilde{G}) = I_{\tilde{r}}$ mit $\tilde{G} \in I\!R^{d\times\tilde{r}}$.*

2. *Falls $K-1 < p$ gilt, so gibt es für jedes $\tilde{G} \in I\!R^{d\times\tilde{r}}$ mit $\tilde{r} > r$ und $(X\tilde{G})'(X\tilde{G}) = I_{\tilde{r}}$ ein $G \in I\!R^{d\times r}$ mit $(XG)'(XG) = I_r$ und*

$$\tilde{M} = \tilde{G}(X\tilde{G})'Y = GB = M$$

Beweis:

1. $r = \min(p, K-1) = p$. Angenommen $\tilde{r} > r$. Aber aus $(X\tilde{G})'(X\tilde{G}) = I_{\tilde{r}}$ folgt, dass $\tilde{r} \leq p = r$ und somit folgt ein Wiederspruch.

2. Aus $(X\tilde{G})'(X\tilde{G}) = I_{\tilde{r}}$ folgt, dass $\tilde{r} \leq p$. Aus $\tilde{B} = (X\tilde{G})'Y$ und $Y = X\tilde{G}\tilde{B}$ folgt, dass $\text{rang}(\tilde{B}) = r$, da

   - $\text{rang}(\tilde{B}) \leq \min(\tilde{r}, K-1) \leq \min(p, K-1) = r$
   - $\text{rang}(\tilde{B}) \geq \text{rang}(Y) = K - 1 \geq \min(p, K-1) = r$

gilt. Mit Hilfe einer Singulärwertzerlegung lässt sich $\tilde{B}$ schreiben als $\tilde{B} = UV', U \in I\!R^{\tilde{r}\times r}$ und $V \in I\!R^{r\times K}$ mit $U'U = I_r$ (Harville, 1997, S. 550). Mit $G = \tilde{G}U$ folgt, dass

$$(XG)'(XG) = U'(X\tilde{G})'(X\tilde{G})U = I_r$$

und

$$B = G'X'Y = U'\tilde{G}'X'Y = U'\tilde{B} = U'UV' = V'$$

und damit

$$M = GB = \tilde{G}UV' = \tilde{G}\tilde{B} = \tilde{M}. \quad \square$$

Damit ist auch bei einer Klassifikation über die vorhersageorientierte Projektion gezeigt, dass der Raum, in dem die Klassifikation stattfindet, höchstens die Dimension $K - 1$ besitzt.

2.6 Weitere Verfahren

In diesem Abschnitt werden weitere Verfahren zur linearen Dimensionsreduktion bei einem Klassifikationsproblem vorgestellt. Diese Verfahren wurden in der Literatur vorgeschlagen, werden aber in dieser Arbeit nicht näher betrachtet und nicht im zu entwickelnden adaptiven Verfahren verwendet. Dieser Abschnitt dient der Vollständigkeit und ermöglicht einen Einblick in die Ideen, die in den letzten zwanzig Jahren aufgekommen sind. Da die Anzahl der vorgeschlagenen Verfahren sehr groß ist, kann hier nur eine Auswahl besprochen werden. Dabei ist zu beachten, dass die Verfahren teilweise aus der Informatik, aus der Statistik und aus den Anwendungen stammen.

Ein Teil der entwickelten Verfahren bezieht sich direkt auf das Fisher-Kriterium (siehe (2.12)). Die meisten Vorschläge kommen dabei aus dem Feld der Informatik. So schlägt Okada & Tomita (1985) orthonormale Diskriminanz Vektoren (ODV) vor. Beim ODV Verfahren wird die Nebenbedingung ($G'\Sigma_W G = I$) ersetzt durch die

Nebenbedingung $G'G = I_r$. Damit wird die Berechnung der optimalen Projektionsmatrix schwieriger: Aus den Eigenvektoren von $\Sigma_W^{-1}\Sigma_B$ werden iterativ die neuen Diskriminanzvektoren gebildet, die dann mit Hilfe des Gram-Schmidt-Verfahrens post-orthonormalisiert werden. Hamamoto *et al.* (1993) zeigt, dass sich mit den orthonormalen Diskriminanz Vektoren (ODV) bessere Ergebnisse in den einzelnen Komponenten

$$J^F(g_{\cdot,i}) = \frac{g'_{\cdot,i}\Sigma_B g_{\cdot,i}}{g'_{\cdot,i}\Sigma_W g_{\cdot,i}}$$

des Fisher-Kriteriums (2.12) erzielen lassen. Insbesondere können so auch mehr als $K-1$ Projektionsvektoren gefunden werden. Allerdings ist das Verfahren nicht besser als das klassische, wenn die gesamte Projektionsmatrix $G = (g_{\cdot,1} \cdots g_{\cdot,r})$ betrachtet wird (siehe dazu S. 53, Satz 10). Da für die Klassifikation nach der Projektion die gesamte Projektionsmatrix benötigt wird, bringt dieses Verfahren keinen neuen Aspekt bei der Allokation.

Loog *et al.* (2001) nehmen auch das Fisher-Kriterium als Ausgangspunkt. Allerdings entwickeln sie unter der paarweisen Zerlegung von Σ_B (siehe Loog (1999))

$$\Sigma_B = \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \pi_i \pi_j (\mu_i - \mu_j)'(\mu_i - \mu_j) \tag{2.69}$$

ein Kriterium, welches den paarweisen Abstand von Klassenmitten als Gewicht mit in das Kriterium einbezieht. Dieser paarweise Ansatz kann mit dem der Optimalen-Separations-Projektion (Abschnitt 2.4) verglichen werden. Allerdings werden die Gewichte aufgrund der Abstände im d-dimensionalen Ursprungsraum bestimmt, so dass dieses Verfahren dem OSP Verfahren unterlegen ist, wo die Abstände im projizierten Raum gewichtet werden. Der Ansatz von Loog *et al.* (2001) löst somit nur das Problem, dass das Fisher-Kriterium ungeeignet ist, wenn eine Klasse weit von den anderen entfernt ist und damit die Eigenwerte von $\Sigma_W^{-1}\Sigma_B$ dominiert (Loog, 1999).

Yan *et al.* (2004) entwickeln mit Hilfe der paarweisen Zerlegung (2.69) ein Kriterium, das auf die Optimierung von

$$\text{spur}\,(G'(\Sigma_B - \Sigma_W)G)$$

hinausläuft. Wenn aber die Nebenbedingung $G'\Sigma_W G = I$ verwendet wird, ist dieses Kriterium äquivalent zu dem von Fisher (2.12). Auch andere Ansätze aus den Bereich der Informatik führen häufig wieder zurück auf das Fisher-Kriterium (Loog *et al.*, 2004).

Ein vielversprechender Ansatz zur Dimensionsreduktion aus der Statistik basiert darauf, eine Teststatistik auf Gleichheit der Wahrscheinlichkeitsverteilungen im projizierten Raum zu maximieren (Zhu & Hastie, 2003):

$$LR(g) = \frac{\max_{f_k} \prod_{i=1}^{K} \prod_{x_j:\{\mathcal{K}_j=i\}} f_k^{(g)}(g'x_j)}{\max_{f_k=f} \prod_{i=1}^{K} \prod_{x_j:\{\mathcal{K}_j=i\}} f_k^{(g)}(g'x_j)}, \tag{2.70}$$

wobei $f_k^{(g)}(\cdot)$ die marginale Dichte von Klasse k bei einer Projektion mit g ist. Unter Berücksichtigung von Orthogonalitätsbedingungen lassen sich mit Hilfe von (2.70) weitere Projektionsrichtungen finden. Abgesehen von numerischen Schwierigkeiten hat dieser Ansatz den Vorteil, generell anwendbar zu sein, sofern ein sinnvoller Schätzer für die Dichte $f_k^{(g)}(\cdot)$ bekannt ist. Daher kann dieser Ansatz mit den Ansatz des Minimalen-Fehler-Klassifikators (Abschnitt 2.3) verglichen werden. Aber Zhu & Hastie (2003) zeigen auch, dass dieser Ansatz unter den Annahmen von Satz 2 zu einer Projektion nach Fisher führt. Da es Ziel dieser Arbeit ist, die Probleme einer Dimensionsreduktion auch unter diesen Annahmen zu lösen, wird dieser auch nichtparametrisch mögliche Ansatz nicht weiter verfolgt.
Ein anderer Ansatz zur Dimensionsreduktion führt über zentrale oder effektive Unterräume (siehe als Übersichtsartikel Cook & Yin (2001)). Unter der Annahme von linearen Klassengrenzen führt dieser Ansatz zur Sliced-Inverse-Regression (SIR, (Li, 1991)). Allerdings führt auch SIR unter den Annahmen von Satz 2 zu einer Projektion nach Fisher (siehe Kent (1991) und Cook & Yin (2001)).

Aus der Statistik kommen Vorschläge, die versuchen, das Problem kollinearer Daten zu lösen. Kollinearität der Daten ist einer der Gründe, eine Dimensionsreduktion zu verwenden. Dabei kann die Lösung des Kollinearitätenproblems durch eine vorherige Dimensionsreduktion erfolgen (siehe Næs & Mevik (2001)). Zu den bekanntesten Verfahren zählt eine Projektion auf die Hauptkomponenten (Hauptkomponentenanalyse, siehe zum Beispiel Mardia *et al.* (1979, Seite 213ff.)) oder auf

die Partial-Least-Squares (PLS) Komponenten (siehe zum PLS Verfahren zum Beispiel Garthwaite (1994)). Da die Hauptkomponenten nur aufgrund der erklärenden Variablen $\mathcal{X}$ bestimmt werden und keine Verbindung zur Zielvariable der Klassen $\mathcal{K}$ haben, ist der Nutzen bei der Klassifikation häufig nicht sehr groß (McLachlan, 1992, Seite 197). Hauptkomponenten sind nur dann sinnvoll zur Erklärung von abhängigen Variablen, wenn die größte Variabilität in $\mathcal{X}$ auch die größten Unterschiede in der Zielvariable erklärt. Trotzdem gibt es Situationen, in denen eine Projektion auf die Hauptkomponenten vor einer Klassifikation gerechtfertigt ist (Chang, 1983). Beim Partial-Least-Squares-Verfahren werden die PLS Komponenten iterativ unter Berücksichtigung der Zielvariable berechnet (für den Algorithmus siehe Weihs & Jessenberger (1999, Seite 170f.)). Barker & Rayens (2003) zeigen, dass im Rahmen eines Klassifikationsproblems der Lösungsansatz von PLS darauf hinausläuft die Eigenvektoren der Matrix

$$\Sigma_{B^*} = \sum_{i=1}^{K} \pi_i^2 (\mu_i - \mu)'(\mu_i - \mu) \tag{2.71}$$

zu berechnen, wobei die Eigenvektoren auf die Länge 1 normiert werden. Man kann diese Lösung mit der Lösung nach Fisher (siehe Abschnitt 2.2) vergleichen, wo die Projektion mit Hilfe der (normierten) Eigenvektoren der Matrix $\Sigma_W^{-1}\Sigma_B$ erfolgt. Man beachte, dass das Σ_{B^*} aus (2.71) nur andere Gewichte (π_i^2 anstelle von π_i in Σ_B) benutzt. Daher unterscheiden sich die Lösungen von Fisher und PLS durch die unterschiedlichen Nebenbedingungen ($G'\Sigma_W G = I$ bzw. $G'G = I$), sowie das Gewicht, das den klassenweisen Abweichungen vom allgemeinen Erwartungswert ($(\mu_i-\mu)'(\mu_i-\mu)$) zugewiesen wird. Allerdings ist anders als bei den in Abschnitt 2.5 vorgestellten Verfahren nicht klar, warum dieser Ansatz besser sein sollte. So behaupten Barker & Rayens (2003) und Nocairi *et al.* (2005) auch nur eine Überlegenheit von Partial-Least-Squares-Komponenten gegenüber den Hauptkomponenten in Bezug auf die Klassifikationsgüte.

Ein weiterer Ansatz, das Kollinearitätenproblem zu lösen ist, die Schätzer innerhalb der Klassifikation zu stabilisieren. Auch wenn das nicht Thema der vorliegenden Arbeit ist, sei auf diese Ansätze kurz eingegangen. Innerhalb der linearen Diskrimi-

nanzanalyse (siehe Definition 7) beziehen sich diese Lösungsvorschläge meistens auf die Stabilisierung der Inversen von Σ_W^{-1}. Einer der bekanntesten Ansätze ist der der regularisierten Diskriminanzanalyse (RDA) nach Friedman (1989). Dabei wird unter anderem wie bei einer Ridge-Regression (siehe z.B. Golub *et al.* (1979)) nicht Σ_W invertiert, sondern $\Sigma_W + \rho I$, womit die aufgrund der Multikollinearität auftretenden numerischen Probleme bei geeigneten ρ umgangen werden können. Ähnlich wie in Abschnitt 2.5 ist es auch hier möglich, die Wahl des Ridge-Parameters ρ vorhersageorientiert zu bestimmen (Luebke *et al.*, 2004). In einem integrierten Ansatz von Næs & Indahl (1998) werden RDA, PLS, Hauptkomponenten und andere Verfahren zusammengefasst. Allerdings führt dieser Ansatz zu einer großen Anzahl freier Parameter, die nicht aufgrund von Daten geschätzt werden können. Außerdem kann mit diesem Ansatz auch keine Dimensionsreduktion erreicht werden.

Insgesamt kann man sagen, dass die meisten in der Literatur vorgeschlagenen Verfahren – insbesondere unter der Annahmen von Satz 2 – auf das klassische Kriterium nach Fisher (1936) zurückgeführt werden können und damit auch das in der Einleitung erwähnte Problem nicht lösen. Aus diesem Grund werden sie bei der Erstellung des adaptiven Verfahrens nicht berücksichtigt.

Kapitel 3

Schätzen bei Klassifikationsaufgaben

In Kapitel 2 wurden Verfahren zur linearen Dimensionsreduktion und Klassifikation vorgestellt. Dabei wurden die Verfahren und Kriterien zur Dimensionsreduktion bzw. Klassifikation mit Hilfe der zugrunde liegenden Populationsparameter entwickelt. Um die Verfahren anwenden zu können, muss geklärt werden, wie die jeweiligen Parameter geschätzt werden können (Abschnitt 3.1). Die Frage, wie die Fehlklassifikationswahrscheinlichkeit und damit die Güte der Verfahren geschätzt werden kann, wird im Abschnitt 3.2 behandelt. Abschließend wird in 3.3 gezeigt, wie das Optimum des jeweiligen Dimensionsreduktionsverfahrens gefunden werden kann.

3.1 Schätzen der Parameter der Verteilungen

Leider werden die Vorraussetzungen für die Anwendung von Klassifikationsverfahren in der Literatur häufig nicht sauber herausgearbeitet. Zum Beispiel werden oft anstelle der Populationsparameter die empirischen Werte eingesetzt. Dabei wird ignoriert, dass dies Schätzer sind, die je nach Vorraussetzungen und deren Erfüllung mehr oder weniger optimal sein können.

Es gibt viele Methoden, Schätzer zu finden. Die wahrscheinlich bekanntesten sind:

- Maximum-Likelihood,
- Momentenmethode,
- Bayes und Minimax

(vergleiche zum Beispiel Mood *et al.* (1974, S. 273ff.)). Während die Anwendung der Maximum-Likelihood und Momentenmethode keine Information außer der zugrundeliegenden Verteilung erfordert, benötigt die Anwendung der Bayes bzw. Minimax Methode Zusatzwissen: Die Bayes Methode benötigt mindestens eine Annahme über die Verteilung der Parameter (Posterior Ansatz) und eventuell sogar die Angabe einer Verlustfunktion (Risiko-Ansatz). Beim Minimax-Prinzip werden sowohl Annahmen über die Verteilung der Parameter als auch eine Verlustfunktion benötigt. Da in dieser Arbeit dieses situationsabhängige Vorwissen nicht bekannt ist, finden diese Schätzmethoden hier keine Anwendung.

Wichtige Gütekriterien für Schätzer sind (vergleiche Büning & Trenkler (1994, S. 29f.)):

- Erwartungstreue.
- Schwache und starke Konsistenz.
- Effizienz.

Im folgenden werden Schätzmethoden für die unbekannten Parameter eines Klassifikationsproblems vorgestellt.

A priori Wahrscheinlichkeiten

Geschätzt werden müssen im Rahmen einer Klassifikation die a priori Wahrscheinlichkeiten $\pi_i, \quad i = 1, \ldots K$ der Klassen. Wenn kein Vorwissen über die a priori

Wahrscheinlichkeit vorhanden ist, wird meistens der Schätzer

$$\hat{\pi}_i = \frac{n_i}{N} \tag{3.1}$$

verwendet. Dieser Schätzer ist unter der Annahme, dass die Beobachtungen $x_i, i = 1, \ldots N$ [1] unabhängig und mit der gleichen Wahrscheinlichkeit aus einer interessierenden Gesamtheit gezogen wurden, der Maximum-Likelihood Schätzer für π_i (McLachlan, 1992, Seite 31), wobei die zugrundeliegende Verteilung eine Multinomialverteilung ist. Die Multinomialverteilung ist ein Mitglied der Exponential-Familien (Shao, 1999, Seite 68). Aufgrund der Tatsache, dass $n_i, i = 1, \ldots, K$, eine minimal suffiziente Statistik ist, ist dieser Schätzer daher effizient (Mood *et al.*, 1974, Seite 326).
Sollte die Stichprobe nicht unabhänig mit der gleichen Wahrscheinlichkeit gezogen worden sein, müssen die Schätzer für die a priori Wahrscheinlichkeit angepasst werden. Lösungsansätze für dieses Problem finden sich zum Beispiel in McLachlan (1992, Seite 31ff.).

Populationsparameter

Zur Berechnung einer linearen Diskriminanzanalyse nach Definition 7 werden neben den geschätzten a priori Wahrscheinlichkeiten auch noch Schätzungen für die folgenden Populationsparameter benötigt:

- $\mu_k = E(\mathcal{X}|\mathcal{K} = k), \quad k = 1, \ldots K$ die Erwartungswertvektoren.
- $\Sigma_k = Cov(\mathcal{X}|\mathcal{K} = k), \quad k = 1, \ldots K$ die Kovarianzmatrix innerhalb der Klassen.

Satz 9 *Die Schätzer*

- $\hat{\mu}_k := \bar{x}_k = \frac{1}{n_k} \sum_{j=1}^{N} I_{\{k_j = k\}} x_j$

[1] i ist hier der Laufindex über die Beobachtungen

- $\hat{\Sigma}_k := \frac{1}{n_k} \sum_{j=1}^{N} I_{\{k_j=k\}}(x_j - \bar{x}_k)'(x_j - \bar{x}_k)$

sind

a) unter der Annahme unabhängig identisch verteilter Beobachtungen innerhalb einer Klasse die Momentenschätzer für μ_k und Σ_k.

b) unter der Annahme, dass die Beobachtungen innerhalb einer Klasse $x_i, i = 1, ..., n_k, k_i = k$, unabhängig identisch normalverteilt mit Erwartungswertvektor μ_k und Kovarianzmatrix Σ_k sind, die Maximum-Likelihood Schätzer der Populationsparameter.

Beweis:
a) siehe z.B. Shao (1999, S. 173).
b) siehe z.B. Giri (1996, S. 96). □

Dabei ist $\hat{\mu}_k$ erwartungstreu, $\hat{\Sigma}_k$ aber nur asymptotisch erwartungstreu (Giri, 1996, S. 99). Daher wird häufig antelle von $\hat{\Sigma}_k$ der unverzerrte Schätzer

$$S_k := \frac{n_k}{n_k - 1} \hat{\Sigma}_k \tag{3.2}$$

verwendet.

Korollar 7 *Die Schätzer $\bar{x}_k, S_k$ sind*

a) unter der Annahme unabhängig identisch verteilter Beobachtungen innerhalb einer Klasse stark konsistent.

b) unter der Annahme, dass die Beobachtungen innerhalb einer Klasse $x_i, i = 1, ..., n_k, k_i = k$ unabhängig identisch normalverteilt mit Erwartungswertvektor μ_k und Kovarianzmatrix Σ_k sind, stark konsistent und effizient.

Beweis:
a) Nach Satz 9 und der Anwendung des starken Gesetzes der großen Zahlen für den Momentenschätzer (Shao, 1999, S. 173).
b) Nach Satz 9 und der Tatsache, dass (μ_k, S_k) eine minimale, vollständige und suffiziente Statistik ist (Giri, 1996, S. 107). □

Insgesamt sind die Populationsparameter unter der Annahme, dass die Beobachtungen klassenweise unabhängig identisch verteilt sind, zumindest konsistent und unverzerrt. Bei normalverteilten Daten sind sie sogar effizient – wenn auch im Falle $d \geq 3$ unzulässig. Eine Alternative ist dann der Stein-Schätzer, siehe zum Beispiel Giri (1996, S. 113ff.).

Allerdings bleibt die Frage, was passiert, wenn die Beobachtungen diesen Annahmen nicht genügen:

- Nicht identisch verteilt: Falls die Verteilung innerhalb einer Klasse selbst eine Mischverteilung ist, bleibt unklar, was die Schätzer schätzen. Mischverteilungen haben häufig mehrere Modalwerte, so dass weder Stichprobenmittelwert noch Stichprobenkovarianz die Mischverteilung charakterisieren.
  Ein wichtiger Spezialfall von nicht identisch verteilten Beobachtungen sind Ausreißer. Es ist bekannt, dass Ausreißer – insbesondere, wenn sie mehr in einer Richtung als in den anderen vorkommen– die Schätzung des Mittelwertes verzerren. Eine noch größere Verzerrung erzeugen Ausreißer aber bei der Varianz (siehe zum Beispiel Miller (1986, Seite 10)). Die Schätzung der Varianz (oder im multivariaten Fall der Kovarianzmatrix) enthält quadratische Terme, und damit wird der Effekt der Abweichung der Ausreißer noch verstärkt. Daher sollten in Fällen, bei denen Ausreißer vorliegen, robuste Schätzer für die Populationsparameter verwendet werden, siehe McLachlan (1992, Seite 161ff.).

- Nicht unabhängig verteilt: Miller (1986, Seite 32f.) unterscheidet im Wesentlichen zwei Gründe für Abhängigkeit innerhalb der Daten: Einerseits durch Blockeffekte, andererseits durch Reiheneffekte, die sich in zeitlicher oder räumlicher Korrelationen der Beobachtungen auswirken können. Während

Blockeffekte (zum Beispiel hervorgerufen durch Untergruppen innerhalb der Stichprobe) als weitere Einflussvariablen ins Modell mit aufgenommen werden können, ist dies bei Reiheneffekten nicht möglich. Miller (1986, Seite 34) zeigt im univariaten Normalverteilungsfall, dass der Mittelwert bei einer seriellen Korrelation immer noch erwartungstreu geschätzt wird, die Varianz aber entweder über- oder unterschätzt wird – abhängig davon, ob negative oder positve Korrelation vorliegt.

Eine Verletzung der Annahmen, dass die Beobachtungen unabhängig und identisch verteilt sind, kann also schwerwiegende Folgen haben. Insbesondere die Varianz und damit die Kovarianz wird falsch geschätzt. Dies kann insbesondere bei der linearen Diskriminanzanalyse (Definition 7) zu verfälschten Ergebnissen führen, da die Inverse des Schätzers für die Kovarianzmatrix $\hat{\Sigma}_W = \sum_{i=1}^{K} n_i S_i/(N-K)$ verwendet wird (Mardia *et al.*, 1979, Seite 309).

3.2 Schätzen der Fehlklassifikationswahrscheinlichkeit

In Abschnitt 2.1.1 wurde die Fehlklassifikationswahrscheinlichkeit einer Klassifikationsregel $\mathbf{a}$ eingeführt (2.2):

$$e(\mathbf{a}) = \sum_{k=1}^{K} P(\mathbf{a}(\mathcal{X}) \neq k | \mathcal{K} = k) P(\mathcal{K} = k).$$

Diese Fehlklassifikationswahrscheinlichkeit wird auch allgemeine Fehlklassifikationswahrscheinlichkeit genannt (McLachlan, 1992, S. 17). Im Fall, dass n unabhängig identisch verteilte Realisierungen

$$D^n = ((x_1, k_1), (x_2, k_2), \ldots, (x_n, k_n))'$$

der Zufallsvariablen $(\mathcal{X}, \mathcal{K})$ zum Lernen bzw. Schätzen der Klassifikationsregel vorliegen, wird häufig die sogenannte bedingte Fehlklassifikationswahrscheinlichkeit

$$\mathcal{E}^c(\mathbf{a_n}) = \sum_{k=1}^{K} P(\mathbf{a_n}(\mathcal{X}) \neq k | \mathcal{K} = k, D^n) P(\mathcal{K} = k) \tag{3.3}$$

betrachtet (McLachlan, 1992, S. 17), wobei $\mathbf{a_n}$ die Klassifikationsregel, die mit n Daten geschätzt wurde, bezeichnet:

$$\mathbf{a_n} : \{\mathbb{R}^d \times \{1, \ldots, K\}\}^n \rightarrow \{1, \ldots, K\}$$

In (3.3) wird auch deutlich, dass $\mathcal{E}^c(\mathbf{a})$ als messbare Funktion von Zufallsvariablen selbst wieder eine Zufallsvariable ist (Devroye *et al.*, 1996, S. 2). $\mathcal{E}^c(\mathbf{a_n})$ mittelt über die Verteilung $\mathcal{F}$ von $(\mathcal{X}, \mathcal{K})$, aber die Daten D^n sind fest.

Gegeben die Beobachtungen D^n muss also die Realisierung der bedingten Fehlerrate $e^c(\mathbf{a_n})$ von $\mathcal{E}^c(\mathbf{a_n})$ geschätzt werden. Zu diesem Thema gibt es viele Verfahren und Untersuchungen (McLachlan, 1992, S. 338f.). Drei häufig verwendete Schätzer sind die Offensichtliche Fehlerrate, die Training- und Testfehlerrate sowie die Kreuzvalidierte Fehlerrate. Diese werden im Folgenden vorgestellt. Weitere Schätzer für die Fehlerrate basieren beispielsweise auf Bootstrapping oder Jackknife Schätzer (McLachlan, 1992, S. 344ff.).

3.2.1 Offensichtliche Fehlerrate

Die offensichtliche Fehlerrate wird auch Wiedereinsetzungsfehlerrate genannt. Dabei werden die Daten D^n, die zum Schätzen der notwendigen Parameter für die Berechnung einer Klassifikationsregel $\mathbf{a_n}$ verwendet werden, auch zum Schätzen der bedingten Fehlklassifikationswahrscheinlichkeit eingesetzt:

$$\widehat{e_o^c}(\mathbf{a_n}) = \frac{1}{n} \sum_{i=1}^{n} I_{\{\mathbf{a_n}(x_i) \neq k_i\}}. \tag{3.4}$$

Allerdings ist dieser Schätzer für die bedingte Fehlklassifikationswahrscheinlichkeit zu optimistisch und damit verzerrt (McLachlan, 1992, S. 339ff.): Dieselben Daten D^n

werden sowohl zum Schätzen der Parameter als auch zur Fehlerschätzung verwendet. Man kann damit nicht überprüfen, ob sich die Klassifikationsregel zu sehr an die Daten anpasst: So ist die offensichtliche Fehlerrate bei der Klassifikation mit Hilfe der Nächste-Nachbar-Allokation immer 0 (Devroye, 1988).

3.2.2 Training- und Testfehlerrate

Bei dieser Methode wird zusätzlich noch ein Testdatensatz

$$T^m = ((x_{n+1}, k_{n+1}), (x_{n+2}, k_{n+2}), \ldots, (x_{n+m}, k_{n+m}))'$$

von unabhängig identisch nach $\mathcal{F}$ verteilten Beobachtungen, die auch unabhängig von D^n sind, verwendet. Dann ist der Schätzer der Training- und Testmethode definiert als:

$$\widehat{e^c_{tt^m}}(\mathbf{a_n}) = \frac{1}{m} \sum_{j=1}^{m} I_{\{\mathbf{a_n}(x_{n+j}) \neq k_{n+j}\}}. \tag{3.5}$$

Dieser Schätzer ist unverzerrt, es gilt also

$$E\left(\widehat{e^c_{tt^m}}(\mathbf{a_n}) | \mathcal{E}^c(\mathbf{a_n}) = e^c(\mathbf{a_n})\right) = e^c(\mathbf{a_n}).$$

Außerdem kann man zeigen (Devroye *et al.*, 1996, S. 123f.), dass

$$P(|\widehat{e^c_{tt^m}}(\mathbf{a_n}) - e^c(\mathbf{a_n})| > \epsilon | D^n) \leq 2e^{-2m\epsilon^2} \tag{3.6}$$

für alle $\epsilon > 0$ gilt, der Schätzer ist also konsistent.

3.2.3 Kreuzvalidierte Fehlerrate

Bei der Bestimmung der kreuzvalidierten Fehlerrate wird ein vorhandener Datensatz D^n wiederholt in Trainings- und Testdatensatz aufgeteilt. Häufig wird dabei jeweils eine Beobachtung aus dem vorhandenen Datensatz entfernt und als Testbeobachtung verwendet (1-fache Kreuzvalidierung):

$$D^{n,i} = ((x_1, k_1), \ldots, (x_{i-1}, k_{i-1}), (x_{i+1}, k_{i+1}), \ldots, (x_n, k_n))'$$

$$\widehat{e^c_{cv}}(\mathbf{a_n}) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\mathbf{a_{n-1}}(x_i, D^{n,i}) \neq k_i\}} \tag{3.7}$$

$\widehat{e^c_{cv}}(\mathbf{a_n})$ ist ein unverzerrter Schätzer für $e^c(\mathbf{a_{n-1}})$ und nicht für $e^c(\mathbf{a_n})$ (Devroye *et al.*, 1996, S. 407). Da $e^c(\mathbf{a_n})$ in vielen Fällen für $n \to \infty$ nach Wahrscheinlichkeit konvergiert, sind die Unterschiede zwischen $e^c(\mathbf{a_{n-1}})$ und $e^c(\mathbf{a_n})$ für große n gering. Bei der k-fachen Kreuzvalidierung wird der Datensatz D^n in k etwa gleich mächtige, disjunkte Teile eingeteilt. Jeder Teil dient dann einmal der Testdatensatz während die anderen $k-1$ zum Schätzen der Parameter verwendet werden.

Die k-fache Kreuzvalidierung ist insbesondere bei kleinen Datensätzen der Training- und Testmethode vorzuziehen, da durch die Aufteilung in Training- und Testdaten bei der Erstellung der Klassifikationsregel Informationen verloren gehen. Bei der Schätzung von $e^c(\mathbf{a_n})$ durch $\widehat{e^c_{tt^m}}$ wird ein Testdatensatz verwendet. Also wird die Information dieser m Beobachtungen nicht zum Schätzen der Parameter der Klassifikationsregel $\mathbf{a_n}$ verwendet, wodurch die Schätzung dieser Parameter eine größere Varianz bekommt. Die k-fache Kreuzvalidierung hat fast so gute Schätzeigenschaften wie die Training- und Testfehlerrate, ist aber sehr rechenintensiv. Außerdem sind Aussagen über die Fehlerabschätzung von $\widehat{e^c_{cv}}$ zu $e^c(\mathbf{a_n})$ schwieriger (Devroye *et al.*, 1996, S. 125).

3.3 Finden der Optimalen Projektionen

Die in Kapitel 2 vorgestellten Kriterien erfordern unterschiedliche Optimierungsmethoden: Während das Optimum des Fisher-Kriteriums (Defintion 8) aus den geschätzten Populationsparametern direkt berechnet werden kann, werden zur Schätzung der Optimalen Projektion nach dem Minimaler-Fehler-Klassifikator (Definition 10) und den anderen Kriterien stochastische Optimierungsverfahren und rechenintensive Resample-Verfahren eingesetzt.

3.3.1 Optimierung des Fischer-Kriteriums

Das Kriterium nach Fisher lautet (siehe Defintion 8)

$$J^F(G) = \text{spur}\left((G'\Sigma_B G)^{-1} G'\Sigma_W G)\right).$$

Satz 10 *Sei Σ_W regulär und sei $G^{Fisher} := (g_{\cdot,1} \cdots g_{\cdot,r})$ mit*

$$r \leq rang(\Sigma_B) \leq rang(\Sigma_W)$$

und

$$\Sigma_W^{-1}\Sigma_B g_{\cdot,i} = \lambda_i g_{\cdot,i}, \lambda_i > 0, i = 1, \ldots r,$$

wobei

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0.$$

Dann optimiert G^{Fisher} das Fisher Kriterium (2.12). Weiterhin können die $g_{\cdot,i}$ so skaliert werden, dass gilt

$$G^{Fisher\prime}\Sigma_W G^{Fisher} = I_r. \tag{3.8}$$

Da der Beweis bekannt ist, wird er hier nur skizziert (vergleiche Wilks (1963, Seite 580): Durch Ableiten von $J^F(G)$ nach G (siehe z.B. Fukunaga (1990, Seite 448f.)) kann das Optimum als die Lösung eines generalisierten Eigenwertproblems (siehe auch z.B. Satz A7.5 aus Seber (1984, Seite 526f.)) gefunden werden, so dass das optimale G aus den r Eigenvektoren der Matrix $\Sigma_W^{-1}\Sigma_B$ zu den r größten Eigenwerten bestimmt werden kann. Da das Problem invariant zur Skalierung der $g_{\cdot,i}$ ist (siehe Bemerkung zum alternativen Kriterium (2.13)), können diese Eigenvektoren so skaliert werden, dass die Nebenbedingung (3.8) gilt. Das generalisierten Eigenwertproblem erfordert ein reguläres Σ_W. Bei singulären Σ_W kann als Alternative zu Σ_W^{-1} die Moore-Penrose-Inverse verwendet werden (siehe zum Beispiel Raudys & Duin (1998)).

3.3.2 Optimierung des Minimaler-Fehler-Kriteriums

Nach dem Minimaler-Fehler-Kriterium

$$J^{\mathrm{MFK}}(G) = e(\mathbf{a}, G) = P(\mathbf{a}(\mathcal{X}G) \neq \mathcal{K}), \tag{3.9}$$

das Röhl *et al.* (2002) vorstellen (siehe Definition 10), wird die Güte einer Projektion direkt mit der Fehlklassifikationswahrscheinlichkeit bestimmt. Aber da die Bestimmung des optimalen G nach Definition 10, wenn eine der Annahmen von Lemma 2 nicht gegeben sind, nicht mehr direkt möglich sein muss, werden an mehreren Stellen bei der praktischen Anwendung dieser Methode Approximationen und Schätzungen verwendet. Die Autoren stellen zwei verschiedene Ansätze zur direkten Minimierung der Fehlerrate einer Projektion vor, die sich durch die Art der Schätzung der Fehlerrate unterscheiden.

- MEC 1: Schätzt die Fehlerrate des Klassifikators $\mathbf{a}$ im projizierten Raum mittels Bootstrap Methoden.
- MEC 2: Schätzt (parametrisch) $f_i, i = 1, \ldots, K$ im Originalraum und berechnet mittels numerischer Integration oder Monte-Carlo Methoden die Fehlklassifikationswahrscheinlichkeit im projizierten Raum.

Ein Problem von MEC 2 ist, dass die Verteilung der Daten im hochdimensionalen Originalraum geschätzt werden muss. Solche Annahmen sind bei höherer Dimension in der Regel schwer zu bestätigen und Parameterschätzungen werden ebenfalls unsicherer. Diese zunehmende Unsicherheit liegt am so-genannten Fluch der Dimension: Eine d-dimensionale Kovarianzmatrix hat $\frac{(d+1)d}{2}$ freie und damit zu schätzende Parameter. Daher wird in dieser Arbeit nur die erste Variante, MEC 1, behandelt. Die optimale Projektionsmatrix bezüglich (2.25) wird sukzessive mit Hilfe von Simulated Annealing (Bohachevsky *et al.* (1986), Salamon *et al.* (2002)) gefunden. Dabei wird zuerst die optimale Projektion auf eine Dimension gefunden. Anschließend wird die zweite Projektionsrichtung gegeben die erste Projektion optimiert. Dabei wird die im ersten Schritt gefundene erste Spalte der Projektionsmatrix G

festgehalten und über die zweite Spalte optimiert. Dieses Vorgehen wird bis zur r-ten Dimension fortgesetzt. Simulated Annealing ist ein stochastisches, numerisches Optimierungsverfahren. Die verwendete Implementierung wird im Anhang (siehe Seite 115) vorgestellt.

3.3.3 Optimierung des Optimalen-Separations-Kriteriums

Das Optimale-Separations-Kriterium

$$J^{\mathrm{OSP}}(G) = \frac{K-1}{K} \sum_{i=1}^{K} \Phi\left(-\frac{1}{2} \min_{j=1,\ldots,K, j\neq i} \delta_M(i,j|G)\right), \tag{3.10}$$

benötigt ebenfalls eine numerische Optimierung. Ein Problem bei der Bestimmung des Optimalen G^{OSP} ist, dass (2.35) nicht überall nach G abgeleitet werden kann, sowie dass es lokale Minima gibt:

Beispiel: (Schervish, 1984) Sei $\mu_1 = (-1,-1), \mu_2 = (1,3), \mu_3 = (3,1)$ und $\Sigma_W = I_2$. Dann lassen sich alle G mit $r = 1$, die der Nebenbedingung (3.8)genügen, berechnen als $G = (\sin(\alpha), \cos(\alpha))'$, also als Funktion einer Variable. Das resultierende $J^{\mathrm{OSP}}(\alpha)$ ist in Abbildung 3.1 dargestellt.

In Abbildung 3.1 sieht man deutlich die lokalen Minima von (2.35) sowie die Tatsache, dass an den Stellen, in denen die jeweils nächste Klasse wechselt, die Funktion keine Ableitung besitzt. Daher wird zur Bestimmung des optimalen G nach Definition 11 Simulated Annealing (siehe S. 115) verwendet.

3.3.4 Optimierung des Vorhersageoptimale-Projektions-Kriteriums

Da die vorhersageoptimale Projektion

$$J^{\mathrm{Vop}}(G,H) = E_{Y|X} E_{Y_0|X_0} \|(Y_0 \hat{H}_{X,Y} - X_0(\hat{G}_{X,Y} \hat{G}'_{X,Y} X'(Y \hat{H}_{X,Y})))\|^2) \tag{3.11}$$

auf dem *MSEP* (2.63) basiert, ist J^{Vop} als eine Funktion von bedingten Erwartungswerten definiert.Die zugrunde liegende Verteilung von $\mathcal{X}, \mathcal{Y}$ ist allerdings zumindest

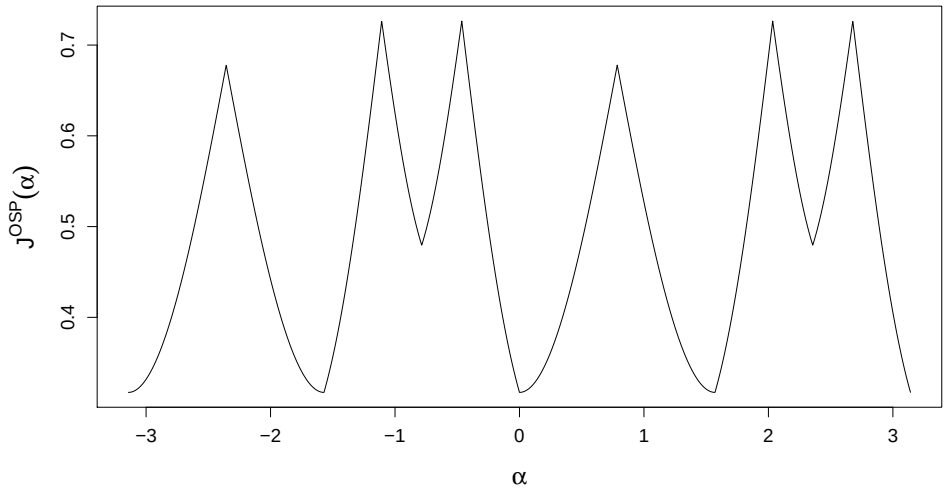

Abbildung 3.1: OSP Kriterium als Funktion einer Variable (Beispiel)

teilweise unbekannt, so dass der *MSEP* selbst geschätzt werden muss. Dies kann zum Beispiel über Bootstrap oder Kreuzvalidierungs Methoden erfolgen. Die Beurteilung einer Projektion über den *MSEP* mit G und H ist bei festem n_0 proportional zu dem Vorfaktor $\frac{1}{n_0}$, der somit nicht berücksichtigt werden muss. Da im Allgemeinen für beliebige passende Matrizen A, B gilt $\|AB\| \neq \|A\| \, \|B\|$ (Gegenbeispiel: $\|I_2 I_2\| = \|I_2\| = \sqrt{2} \neq 2 = \|I_2\| \, \|I_2\|$), muss die Optimierung von (2.64) über beide Matrizen gleichzeitig erfolgen. Da für die Optimalen Projektionen nach Definition 14 keine analytischen Lösung bekannt ist, wird wieder Simulated Annealing als numerische Optimierungsmethode verwendet. Dabei wird wie bei der Optimierung des Minimalen-Fehler-Kriteriums ein sukzessives Vorgehen verwendet: Erst wird die Projektion auf eine Dimension optimiert, dann die Projektion auf die zweite Dimension gegeben die erste. Wiederum wird dabei die erste Spalte von G und H festgehalten und die optimale zweite Spalte der Projektionsmatrizen gesucht. Dieses Vorgehen wird dann bis zur r-ten Dimension fortgesetzt.

Kapitel 4

Charakterisierung der Grundgesamtheit für eine Dimensionsreduktion

Bei der Konstruktion eines adaptiven Verfahrens zur linearen Dimensionsreduktion stellt sich zunächst die Frage, wie eine mögliche Ausgangslage aussehen kann. Insbesondere ist eine adäquate Beschreibung der vorliegenden Datensituation wichtig, da dies die Basis für die Auswahl des besten Verfahrens liefert.

Es gab mehrere europäische Forschungsprojekte mit dem Ziel, die Güte unterschiedlichster Klassifikationsverfahren auf verschiedenen Datensätzen zu evaluieren und daraus Regeln über ihre Performanz abzuleiten. Die Bekanntesten sind das StatLog (Michie *et al.*, 1994) und das METAL Projekt (siehe zum Beispiel Brazdil *et al.* (2003), Pfahringer *et al.* (2000)). Bei beiden Projekten wurde aufgrund von Kennzahlen des vorliegenden Datensatzes das beste Verfahren prognostiziert. Diese Empfehlungen (die beim METAL-Projekt in einem sogenannten Expertensystem münden) wurden auf Basis der Klassifikationsgüte der Verfahren auf realen Datensätzen ermittelt. In diesem Kapitel wird eine systematische Entwicklung von Kennzahlen eines Datensatzes für das Problem der linearen Dimensionsreduktion

im Rahmen einer linearen Klassifikation beschrieben. Diese Kennzahlen sollen dann anschließend für die Selektorstatistik des adaptiven Verfahrens verwendet werden können. Dies impliziert, dass sie aus den Daten bestimmbar sein müssen.

Hand (1997, S. 193) weist im Zusammenhang von Vergleichsstudien bei Klassifikationsaufgaben darauf hin, dass die Faktoren einer Simulationsstudie sorgfältig vor dem Hintergrund der zu untersuchenden Fragestellung ausgewählt werden müssen. Dies ist bisher leider nur selten der Fall (Hand, 1997, S. 193): die simulierten Daten werden häufig für das bevorzugte oder propagierte Verfahren erzeugt, und daher erfolgt kein fairer Vergleich (Hand, 1997, S. 191). Um eine solche Verzerrung zu vermeiden, wird in diesem Kapitel der Raum der möglichen Datensituationen beschrieben, innerhalb dessen zur Entwicklung des adaptiven Verfahrens die Kennzahlen eines Datensatzes variiert werden.

4.1 Statistischer Hintergrund von Verfahrensvergleichen

Die Grundgesamtheit Ω, die einem Klassifikationsproblem zugrunde liegt, besteht aus $K \geq 2$ Klassen $\Omega_1, \ldots, \Omega_K$, wobei gilt

$$\Omega_i \cap \Omega_j = \emptyset, \quad \forall i, j = 1, \ldots, K, i \neq j$$

und

$$\bigcup_{i=1}^{K} \Omega_i = \Omega$$

(siehe Abschnitt 2.1.1). Bei unterschiedlichen Klassifikationsaufgaben (z.B. Kreditwürdigkeit und Gesundheitsstatus) unterscheiden sich sowohl die zugrunde liegende Grundgesamtheit Ω als auch die Verteilung $\mathcal{F}$ der Daten D^n der aus der Grundgesamtheit gewonnenen Zufallsvariablen $(\mathcal{X}, \mathcal{K})$.

Da die Güte eines Klassifikators $\mathbf{a_n}$ gemessen mit Hilfe der bedingten Fehlklassifika-

tionswahrscheinlichkeit (siehe (3.3))

$$\mathcal{E}^c(\mathbf{a_n}|D^n) = \sum_{k=1}^{K} P(\mathbf{a_n}(\mathcal{X}) \neq k|\mathcal{K} = k, D^n)P(\mathcal{K} = k)$$

eine messbare Funktion der Zufallsvariablen $(\mathcal{X}, \mathcal{K})$ darstellt, ist die Güte selber eine Zufallsvariable. Die Verteilung dieser Zufallsvariable hängt neben dem Klassifikator $\mathbf{a_n}$ auch von den Daten D^n und damit von der gemeinsamen Verteilung $\mathcal{F}$ der Zufallsvariablen $(\mathcal{X}, \mathcal{K})$ des Klassifikationsproblems ab. Bei einem statistischen Vergleich von Klassifikations- bzw. Dimensionsreduktionsverfahren werden die Verteilungen der Zufallsvariablen $\mathcal{E}^c(\mathbf{a_n^j}|D^n), j = 1, \ldots, L$ von L Klassifikatoren verglichen. Wird ein adaptives Verfahren verwendet, wird aufgrund der Information, die aus D^n gewonnen werden kann, der Klassifikator $\mathbf{a_n^l}$ ausgewählt, bei dem der Erwartungswert des bedingten Fehlers,

$$E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n^l})) \tag{4.1}$$

minimal ist. Dafür werden Abbildungen von dem Raum aller möglichen Klassifikationsprobleme $\Omega^* = \bigcup_j \Omega^j$ mit einer Familie von Verteilungen $\mathbf{F}$ mit $\mathcal{F}^j \in \mathbf{F}$ nach $\mathbb{R}^c$ gesucht, die es erlauben, Aussagen über die Verteilung von $\mathcal{E}^c(\mathbf{a_n}|D^n)$ bzw. $E(\mathcal{E}^c(\mathbf{a_n}))$zu treffen. Eine Realisierung einer solchen Abbildung wird hier Datensituation genannt.

Definition 16 *Eine Datencharakteristik ist eine Eigenschaft der Stichprobe, die zum Lernen bzw. Schätzen eines Verfahrens verwendet wird. Eine Datencharakteristik ist relevant, wenn sie einen Einfluss auf die relative oder absolute Güte eines mit der Stichprobe geschätzten Verfahrens hat. Die Datensituation ist die Gesamtheit aller Datencharakteristika.*

Zum Begriff Datencharakteristik siehe auch Rendell & Cho (1990) und Aha (1992). Formal ist eine Datensituation bei einer Klassifikation das Ergebnis einer Abbildung $g(\cdot)$:

$$g : g(D^n) : \mathbb{R}^{n\times(d+1)} \rightarrow \mathbb{R}^c. \tag{4.2}$$

Datencharakteristika in diesem Sinne sind häufig Parameter bzw. deren Schätzer einer Verteilung.

Beispiel: Die Schätzer für die Parameter μ, σ sowie der Stichprobenumfang n sind Datencharakteristika einer Realisierung von n unabhängig identisch normalverteilten Daten:

$$u_i \sim_{u.i.v.} \mathbb{N}(\mu, \sigma).$$

Dann ist die Datensituation $g(u_1, u_2, \ldots, u_n)$ definiert als

$$g(u_1, u_2, \ldots, u_n) = (\hat{\mu}, \hat{\sigma}, n).$$

Der aus der vorliegenden Stichprobe konkret ermittelte Wert von $\hat{\mu}$ ist aber nicht relevant für die Güte der Maximum-Likelihood-Schätzmethode (siehe Abschnitt 3.1) für den Erwartungswert, wenn diese mit Hilfe des mittleren quadratischen Fehlers bestimmt wird, da gilt

$$E\left((\mu - \hat{\mu})^2\right) = Var(\hat{\mu}) = \frac{\sigma^2}{n}.$$

Aufgrund dieser Definition sind Datencharakteristika verwandt mit suffizienten Statistiken. Bei einer suffizienten Statistik ist die Verteilung der Daten gegeben die Statistik unabhängig von den unbekannten Parametern einer Verteilung. Da bei Klassifikationsaufgaben die Verteilung der bedingten Fehlklassifikationswahrscheinlichkeit meistens unbekannt ist (McLachlan, 1992, S. 338), wird hier anstelle von suffizienten Statistiken der Begriff der Datencharakteristik verwendet.

4.2 Datencharakteristika bei linearer Dimensionsreduktion in der Klassifikation

Um im vorliegenden Fall einer linearen Dimensionsreduktion zur Klassifikation Datencharakteristika zu bestimmen, sei zunächst der Fall betrachtet, dass die Daten

bedingt die Klassen multivariat normalverteilt sind. Zur Vereinfachung werden wieder die Annahmen von Seite 7 verwendet:

(A1) Die a priori Wahrscheinlichkeiten sind gleich:
$\pi_1 = \cdots = \pi_K = \frac{1}{K}$.

(A2) Die Kovarianzmatrizen innerhalb der Klassen sind gleich:
$\Sigma_1 = \ldots = \Sigma_K = \Sigma_W$.

(A3) Die gemeinsame Kovarianzmatrix Σ_W ist regulär, es existiert also Σ_W^{-1}.

Dann beschreiben die verschiedenen Möglichkeiten, in denen die Klassenmitten μ_i zueinander stehen können, als relevante Datencharakteristika das Klassifikationsproblem. Um die jeweilige Lage der Mittelwertsvektoren zu ermitteln, wird in Abschnitt 4.2.1 eine (invariante) Repräsentation der möglichen Konfigurationen hergeleitet.

In Kapitel 3 wurden Schätzverfahren und Schätzer für interessierende Parameter vorgestellt. Alle Schätzer sowie auch die Güte der Schätzungen hängen von der Anzahl der Beobachtungen n der Daten D^n ab. Die Anzahl der Beobachtungen ist damit eine weitere Datencharakteristik.

Die zugrundeliegende Verteilung innerhalb der Klassen hat natürlich auch einen Einfluss auf die relative Güte der Verfahren (siehe Michie *et al.* (1994) oder Brazdil *et al.* (2003)). Um diesen Effekt systematisch zu untersuchen, muss ein Simulationssystem verwendet werden, in dem die bedingten Verteilungen abweichend von der Normalverteilung variiert werden können. Eine solche Charakterisierung von Verteilungen wird im Abschnitt 4.2.2 vorgestellt.

Eine weitere Datencharakteristik ist die Kovarianzmatrix Σ_W innerhalb der Klassen. Eine für die Klassifikation relevante Charakteristik von Σ_W wird in Abschnitt 4.2.3 entwickelt.

4.2.1 Konfigurationen der Mittelwertsvektoren

Das Problem der Lage der Klassenmitten ist bisher sehr wenig untersucht worden. In Rendell & Cho (1990) wird der Einfluss der Lage von zwei Klassen auf die Güte von zwei Varianten von Entscheidungsbäumen ermittelt, wobei jede Klasse aus Mischverteilungen besteht. Wettschereck & Dietterich (1995) untersuchen die Performanz von nächsten-Nachbarn und nächsten-Hyperrechteck bei drei unterschiedlichen Konstellationen im $\mathbb{R}^2$. Friedman (1989) erzeugt zwei Konstellationen: Einerseits liegen die Unterschiede der Mittelwerte im Unterraum mit hoher Varianz, andererseits in einem Unterraum mit niedriger Kovarianz. Dabei konnte kein Unterschied bei der Klassifikation mit Hilfe der LDA (ohne Dimensionsreduktion) festgestellt werden (Friedman, 1989). Allerdings kann dieses Ergebnis mit Hilfe der vereinfachende lineare Abbildung nach McCulloch (1986) erklärt werden (siehe unten). Daher muss für die Charakteristik der Konfiguration der Mittelwertsvektoren erst die theoretische Grundlage entwickelt werden.

Aufgrund von Satz 4 kann ohne Beschränkung der Allgemeinheit davon ausgegangen werden, dass $d = K - 1$ ist. Häufig wird die Dimension d bei realen Daten größer als $K - 1$ sein. Im Beispiel der multivariaten Bestimmung von Konjunkturphasen (Kapitel 7) ist die Ausgangsdimension $d = 13$ und die Anzahl der Klassen $K = 4$. Bei regulären Σ_W und im Falle von $d > K - 1$ kann die sogenannte vereinfachende lineare Abbildung nach McCulloch (1986) angewendet werden:

1. $z = x\Sigma_W^{-\frac{1}{2}}$. Die Klasse i in z hat dann den Erwartungswert $\nu_i = \mu_i\Sigma_W^{-\frac{1}{2}}$, und die gemeinsame Kovarianz ist I_d.

2. $y = z - \bar{\nu}$ wobei $\bar{\nu}$ der Mittelwert der ν_i ist. Der Erwartungswert von Klasse i in y ist jetzt $\alpha_i = \nu_i - \bar{\nu}$. Die Kovarianz ist unverändert.

3. Sei $A = (\alpha_1' \cdots \alpha_K')'$. Mit Hilfe einer Singulärwertzerlegung ist $A = UDV'$ wobei U eine orthonormale $K \times d$ und V eine orthonormale $d \times d$ Matrix ist. D ist eine Diagonalmatrix mit den Einträgen in der Diagonalen $\|\lambda_1\| \geq \|\lambda_2\| \geq \cdots \geq \|\lambda_r\| \geq 0 = \cdots = \|\lambda_d\| = 0$.

4. $t = yV$. Dann ist der Erwartungswert von Klasse i in t die i-te Zeile von UD. Die Kovarianz innerhalb der Klassen ist immer noch I_d.

Falls die zugrunde liegenden Verteilung von $\mathcal{X}$ eine multivariate Normalverteilung ist, ist die Verteilung von $\mathcal{T}$ ebenfalls eine Normalverteilung, da sich die vereinfachende lineare Abbildung als Linearkombination darstellen lässt und damit die Normalverteilung erhalten bleibt (Mardia *et al.*, 1979, S. 62). Weiterhin sind die d Variablen von $\mathcal{T}$ unabhängig verteilt mit Varianz 1 und der Erwartungswert der Klassen in den Variablen $r+1, r+2, \ldots, d$ ist 0. Daher kann ohne Informationsverlust bei der Fehlerrate von $\mathcal{X}$ auf die ersten $r = K-1$ Variablen von $\mathcal{T}$ projiziert werden (McCulloch, 1986). Allerdings muss beachtet werden, dass dazu die Matrix $\Sigma_W^{-\frac{1}{2}}$ benötigt wird, was bei der konkreten Berechnung numerische Probleme aufwerfen kann (siehe Abschnitt 4.2.3). Im Folgenden wird davon ausgegangen, dass die vereinfachende lineare Abbildung durchgeführt wurde, dass also gilt $d = K-1$ und $\Sigma_W = I_{K-1}$. Dann beschränkt sich das Problem der Charakterisierung der möglichen Mittelwertkonstellationen darauf, die Lage von K Punkten (Mittelwerten) im $K-1$ dimensionalen Raum zu beschreiben.

Die verschiedenen Dimensionsreduktionsverfahren, die in Kapitel 2 vorgestellt wurden und zur Konstruktion eines adaptiven Verfahrens herangezogen werden, sind invariant bezüglich Rotation, Reflexion und Translation:

Lemma 5 *Das Fisher-Kriterium, das Minimale-Fehler-Kriterium, das Optimale-Separations-Projektionskriterium und das Kriterium bei einer vorhersageoptimalen Projektion sind invariant gegenüber Rotation, Reflexion und Translation der Daten, denn es gilt*

$$J(G)_{\mathcal{X}} = J(R'G)_{\mathcal{X}R+v}, \tag{4.3}$$

wobei $R^{-1} = R'$ und damit $R'R = RR' = I$ ist.

Beweis:
Sei $\mathcal{Y} = \mathcal{X}R + v$, wobei $\Sigma_W(\mathcal{X}) = I$ und $\pi_1 = \cdots = \pi_K$. Aus $R'R = I$ folgt, dass

$\Sigma_W(\mathcal{Y})$ ebenfalls die Einheitsmatrix I ist. Weiterhin gilt für die Kovarianzmatrix im Bildraum:

$$\begin{aligned} Cov(\mathcal{Y}(R'G)|\mathcal{K}=i) &= Cov((\mathcal{X}R+v)(R'G)|\mathcal{K}=i) \\ &= G'RCov(\mathcal{X}R+v|\mathcal{K}=i)R'G \\ &= G'RR'Cov(\mathcal{X}|\mathcal{K}=i)RR'G \\ &= G'Cov(\mathcal{X}|\mathcal{K}=i)G =: \Sigma_W^G. \end{aligned}$$

- Fisher-Kriterium: Hier muss gezeigt werden, dass $(R'G)'\Sigma_B(\mathcal{Y})(R'G) = G'\Sigma_B(\mathcal{X})G$ gilt (siehe (2.12)).

$$\begin{aligned} (R'G)'\Sigma_B(\mathcal{Y})(R'G) &= (R'G)'[\frac{1}{K}\sum_{i=1}^{K}(\mu_i R+v-(\mu R+v))' \\ &\quad (\mu_i R+v-(\mu R+v))](R'G) \\ &= (R'G)'\left(\frac{1}{K}\sum_{i=1}^{K}R'(\mu_i-\mu)'(\mu_i-\mu)R\right)(R'G) \\ &= (R'G)'R'\Sigma_B(\mathcal{X})R(R'G) \\ &= G'\Sigma_B(\mathcal{X})G. \end{aligned}$$

- Minimaler-Fehler-Kriterium: Die Allokation (und damit die mögliche Fehlerrate) ändert sich nicht, da bei Allokation mittels LDA in (2.25) gilt

$$\begin{aligned} f_k(y(R'G)) &= |2\frac{1}{K}\Sigma_W^G|^{-0.5} \\ &\quad \exp(-\frac{1}{2}(y-(\mu_k R+v))(R'G)(\Sigma_W^G)^{-1}(y-(\mu_k R+v))(R'G)') \\ &= |2\frac{1}{K}\Sigma_W^G|^{-0.5} \\ &\quad \exp(-\frac{1}{2}(x-\mu_k)R(R'G)(\Sigma_W^G)^{-1}(x-\mu_k)R(R'G)') \\ &= |2\frac{1}{K}\Sigma_W^G|^{-0.5} \\ &\quad \exp(-\frac{1}{2}(x-\mu_k)G(\Sigma_W^G)^{-1}G'(x-\mu_k)') \\ &= f_K(xG). \end{aligned}$$

- Optimale-Separations-Kriterium: Bei dem Optimale-Separations-Kriterium muss gezeigt werden, dass $\delta_M^2(i,j|R'G)^{\mathcal{Y}} = \delta_M^2(i,j|G)^{\mathcal{X}}$ gilt (siehe (2.35)).

$$\begin{aligned}\delta_M^2(i,j|R'G)^{\mathcal{Y}} &= ((\mu_i R + v - (\mu_j R + v))R'G)\,(\Sigma_W^G)^{-1} \\ &\quad ((\mu_i R + v - (\mu_j R + v))R'G)' \\ &= (\mu_i - \mu_j)RR'G(\Sigma_W^G)^{-1}((\mu_i - \mu_j)RR'G)' \\ &= \delta_M^2(i,j|G)^{\mathcal{X}}.\end{aligned}$$

- Vorhersageoptimale-Projektion-Kriterium: Aufgrund der Mittelwertbereinigung spielt die Verschiebung v keine Rolle. Der Rest folgt aus

$$(XR)(R'G)(R'G)'(XR)' = X(RR')GG'(RR')X' = XGG'X' \tag{4.4}$$

  in (2.64).

Daher sind die Werte der Kriterien invariant zu Rotation, Reflexion und Transformation. □

Dadurch, dass es für jede Projektionsmatrix G in $\mathcal{X}$ eine äquivalente Projektion $R'G$ in $\mathcal{Y}$ gibt, sind auch die Schätzer für die jeweils optimalen Projektionen äquivalent. Da alle Beobachtungen die gleiche bijektive Abbildung erfahren, sind die Allokationen und damit die Fehlerraten identisch.

Durch die Abbildung $\mathcal{Y} = \mathcal{X}R + v$ mit $R^{-1} = R'$ wird eine Äquivalenzrelation im $\mathbb{R}^{K\times(K-1)}$ induziert. Daher genügt es, Repräsentanten der Äquivalenzklassen zu untersuchen. Dryden & Mardia (1998, S. 57) nennen den Raum, der sich auf diese Äquivalenzklassen bezieht, den Reflexion-Größe-und-Form Raum. Um die Repräsentanten der Äquivalenzklassen erfassen zu können, werden Erkenntnisse der Geometrie und der statistischen Analyse von Formen herangezogen. Lele & Richtsmeier (2001) definieren die Form eines Objektes (hier K Punkte im $\mathbb{R}^{K-1}$) als die Charakteristik, die invariant unter Verschiebung, Rotation oder Reflexion des Objektes ist. Die Distanzmatrix Δ der $\frac{K(K-1)}{2}$ Abstände der Punkte voneinander (Lele & Richtsmeier, 2001, S. 74) ist eine Repräsentation der Punkte, die invariant innerhalb einer solchen Äquivalenzklasse ist. Umgekehrt kann aus einer Distanzmatrix

Δ – sofern sie die Eigenschaften einer Distanzmatrix ($\Delta_{ii} = 0, \Delta_{ij} = \Delta_{ji} \geq 0$) erfüllt – unter gewissen Bedingungen wieder ein Repräsentant der Äquivalenzklasse gewonnen werden (siehe z.B. (Mardia *et al.*, 1979, S. 397f.)).

Außerdem ist das Problem der Beschreibung der Konfiguration invariant gegenüber der Bezeichnung der Klassen. Das heißt, dass die Klassifikationsverfahren invariant zu Permutationen der Zeilen der Matrix der Mittelwertsvektoren Ξ sind und damit gegenüber den entsprechenden Permutationen innerhalb der Distanzmatrix Δ. Innerhalb der statistischen Analyse von Formen bezieht man sich auf nicht-markierte Landmarken. Eine Permutations-invariante Beschreibung der Distanzmatrix kann durch die normierten zentralen (empirischen) Momente η_j erfolgen. Die normierten zentralen empirischen Momente der Distanzmatrix sind also die gesuchte Beschreibung der Mittelwertsvektoren, da erstens die Distanzmatrix invariant gegenüber Verschiebung, Rotation oder Reflexion ist, und zweitens die empirischen Momente invariant gegenüber Permutationen sind. Um die Invarianz gegenüber Permutationen zu zeigen seien zur Vereinfachung der Notation die Abstände mit $x_\cdot$ anstelle von $\Delta_{\cdot\cdot}$ bezeichnet. Dann sind die normierten zentralen (empirischen) Momente η_j definiert als:

$$\eta_1(n) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4.5}$$

$$\eta_2(n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \eta_1)^2 \tag{4.6}$$

$$\eta_j(n) = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \eta_1)^j}{\sqrt{\eta_2}^{\,j}}, \quad j = 3, 4, \ldots \tag{4.7}$$

Lemma 6 *Die Folge $\eta_j(n), j = 1, 2, \ldots$ bestimmt bis auf Permutationen die disjunkte Menge $\{x_i : x_i \in I\!R, i = 1, \ldots, n\}$.*

Beweis:

Die diskrete Dichtefunktion $f(x)$ sei definiert als

$$f(x) = \begin{cases} \frac{1}{n} & \text{wenn} \quad x \in \{x_i : x_i \in I\!R, i = 1, \ldots, n\} \\ 0 & \text{sonst.} \end{cases}$$

Somit bestimmt die Menge $\{x_i : x_i \in I\!R, i = 1, \ldots, n\}$ die Funktion $f(x)$ eineindeutig. Für die so konstruierte Dichtefunktion existieren alle Momente und die momenterzeugende Funktion. Wenn aber die momenterzeugende Funktion existiert, so bestimmen die Momente eindeutig die Dichtefunktion (Mood *et al.*, 1974, Seite 81) und damit die Menge $\{x_i\}$ als die Punkte, an denen die Dichtefunktion größer als Null ist. Diese Argumentation ist unabhängig davon, ob die x_i Realisierungen einer Zufallsvariablen sind oder nicht (Mood *et al.*, 1974, Seite 78). □

Die ersten beiden normierten zentralen Momente sind der Mittelwert (η_1) und die Varianz (η_2) einer Stichprobe $x_1, \ldots, x_n$. η_3 ist die Schiefe der Stichprobe und η_4 die Wölbung (Johnson & Lowe, 1979). Es ist klar, dass die zentralen Momente $\eta_3, \eta_4, \ldots$ nach Konstruktion invariant zur Lage und Skalierung der x_i sind.

Dabei sind folgende Ungleichungen für Schiefe und Wölbung bekannt:

$$0 \leq \eta_4 - \eta_3^2 - 1 \tag{4.8}$$

$$\eta_4 \leq n \tag{4.9}$$

(siehe Johnson & Lowe (1979)).

Getrennt für jede Achse werden Momente häufig in der Bilderkennung und Bildbeschreibung eingesetzt (Prokop & Reeves, 1992). Dabei werden zur Charakterisierung des Bildes meistens nur die ersten drei bis vier Momente benötigt (siehe auch die Orginalarbeit zum Einsatz von Momenten in der Bilderkennung Hu (1962)). In dieser Arbeit werden die normierten zentralen Momente zur Beschreibung des Abstandes zwischen den Klassenmitten verwendet. Insbesondere die Momente η_2, η_3, η_4 eignen sich zur Beschreibung der meisten interessanten Konstellationen im hier untersuchten Fall von vier Klassen. Dazu seien ein paar Beispielkonstellationen betrachtet:

1. Alle Mittelwertsvektoren haben (in etwa) den gleichen Abstand: η_2 minimal, η_3, η_4 beliebig.

2. Drei Mittelwertsvektoren sind nah, einer von den anderen entfernt (Ausreißerklasse): Dann gibt es drei kleine und drei große Abstände, und damit ist

$\eta_3 \approx 0$ und η_4 minimal (nahe bei 1).

3. Je zwei Klassen sind nah beieinander: Dann gibt es zwei kleine und vier große Abstände, und damit ist $\eta_3 < 0$ und η_4 klein.

4. Nur zwei Klassenmitten sind (relativ) nah, die anderen Klassen haben einen in etwa gleich großen Abstand: In dieser Situation ist η_3 sehr klein und η_4 eventuell groß.

5. zwei Klassenmitten sind nah, die beiden anderen Klassen liegen einander über die Achse der beiden nahen Klassenmitten gegenüber: Es gibt einen kleinen, vier mittlere und einen großen Abstand, also ist η_4 groß (relativ zu η_3).

Um die Zusammenhänge zwischen Konstellation und den normalisierten zentralen Momenten der Abstände zu verdeutlichen, wurde von jeder Konstellation eine Basis von Abständen gewählt. Innerhalb der Basis beträgt der kleine Abstand 2, ein großer 6 (bzw. 10 in Konstellation 5). Zu jedem Abstand innerhalb der Basis wurde unabhängig eine standardnormalverteilte Zufallszahl addiert, so dass recht große Abweichungen von der Basis möglich sind. Abbildung 4.1 zeigt die Verteilungen der dritten und vierten normalisierten zentralen Momente der Konstellationen 2,3,4,5. Sie verdeutlichen, dass man mit η_3, η_4 diese Konstellationen von vier Klassenmitten im $\mathbb{R}^3$ unterscheiden kann.

4.2.2 Charakterisierung von Verteilungen

Viele statistische Verfahren sind optimal unter der Annahme einer Normalverteilung, so zum Beispiel die lineare Diskriminanzanalyse (siehe Korollar 1). Die häufige Verwendung der Annahme einer Normalverteilung kann in vielen Situationen mit Hilfe des zentralen Grenzwertsatzes (Shao, 1999, S. 47) motiviert werden. Da im Rahmen einer linearen Dimensionsreduktion außerdem gilt, dass $\mathcal{Z}$ in $\mathcal{Z} = \mathcal{X}G$ eine Linearkombination der Ausgangsvariablen $\mathcal{X}$ ist, kann unter schwachen Voraussetzungen gezeigt werden, dass $\mathcal{Z}$ bei N und d gegen unendlich approximativ

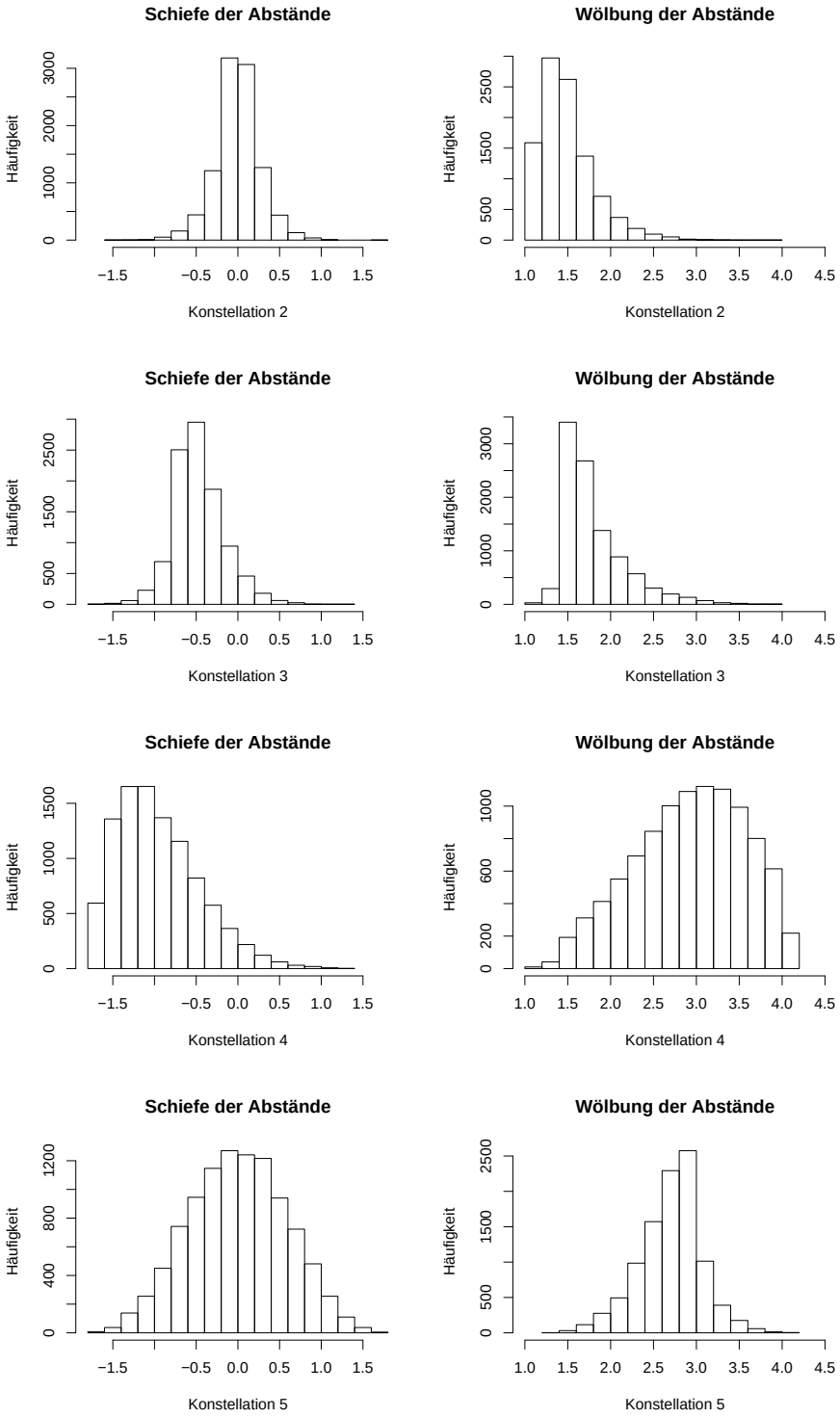

Abbildung 4.1: Normalisierte zentrale Momente der Abstände von vier Punkten im $\mathbb{R}^3$ in verschiedenen Situationen

normalverteilt ist (Diaconis & Freedman, 1984). Im endlichen Fall stellt sich aber die Frage, wie empfindlich die Verfahren auf die Verletzung der Annahme einer Normalverteilung reagieren. Daher ist es notwendig, für den Lernprozess wann welches Verfahren das Beste ist, Verteilungen zu generieren, die in einer bekannten Weise von der Normalverteilung abweichen. Es gibt verschiedene Systeme nicht normalverteilte Verteilungen zu erzeugen (siehe z.B. Tadikamalla (1980)) wie das Johnson- (Johnson, 1949) oder das Fleishman-System (Fleishman, 1978). Bei diesen Systemen werden (univariat) normalverteilte Zufallsvariablen so transformiert, dass die neue Zufallsvariable bestimmte, normalisierte dritte und vierte zentrale Momente besitzt. Konkret werden also bestimmte Werte für die Schiefe und Wölbung angestrebt, die mehr oder weniger von denen der Normalverteilung (Schiefe 0 und Wölbung 3) abweichen. Gleichzeitig geben Schiefe und Wölbung auch Hinweise auf die Art der Abweichung von der Normalverteilung (Miller, 1986, S. 14f.).

Aufgrund der kompakteren Darstellbarkeit und Programmierbarkeit wird das Fleishman-System (Fleishman, 1978) verwendet, um unterschiedliche Werte der Verteilung einer Zufallsvariablen auf der Ebene der Schiefe und Wölbung zu erhalten.

Definition 17 *Sei $\mathcal{U}$ eine standardnormalverteilte Zufallsvariable, so führt die Potenztransformation*

$$\mathcal{V} = a + b\,\mathcal{U} + c\,\mathcal{U}^2 + d\,\mathcal{U}^3 \tag{4.10}$$

zu einer Zufallsvariable mit einer Verteilung im System von Fleishman (1978).

Der Vorteil des Systems von Fleishman ist, dass man Gleichungssysteme aufstellen kann, die die Parameter a, b, c, d ins Verhältnis zu den Erwartungswerten von Mittelwert, Varianz, Schiefe und Wölbung setzen (Fleishman, 1978). Da man o.B.d.A. den Mittelwert auf 0 und die Varianz auf 1 setzen kann, ist es möglich, für die meisten möglichen Werte von Schiefe und Wölbung numerisch die passenden Parameter a, b, c, d zu bestimmen und dann anschließend Mittelwert und Varianz anzupassen. Sehr niedrige Werte für die Wölbung können nicht realisiert werden. Dies ist ein Nachteil des Systems von Fleishman (Fleishman, 1978). Abbildung 4.2 zeigt die

geschätzten Dichten für Verteilungen nach den Fleishman-System, wobei sowohl die Standardnormalverteilung (Schiefe 0, Wölbung 3) als auch Zufallszahlen mit einer geringeren Wölbung (Schiefe 0 und Wölbung 2) und schiefe Daten (Schiefe 1, Wölbung 3.5) mit Hilfe des Fleishman-System generiert wurden.

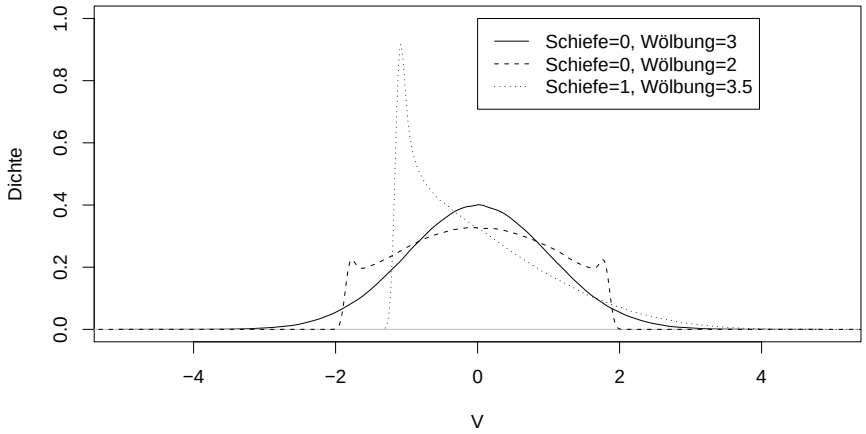

Abbildung 4.2: Verschiedene Dichten erzeugt mit dem Fleishman-System

Um die Konvergenzgeschwindigkeit der empirischen Momente nach dem Fleishman-System zu überprüfen, wurde der mittlere quadratische Fehler

$$MQF(i,n) = \frac{1}{N}\sum_{j=1}^{N}(E(\eta_i) - \eta_i^j(n))^2 \tag{4.11}$$

mit $\eta_i^j(n)$ nach (4.7) als das empirische, normierte, zentrale Moment von n Beobachtungen in der j-ten Stichprobe betrachtet. In Abbildung 4.3 sind die mittleren quadratischen Fehler für die obigen Beispiele aufgrund von 10000 zufälligen Stichproben aufgetragen. Man sieht, dass man schon bei relativ kleinem n den mittleren quadratischen Fehler vernachlässigen kann.

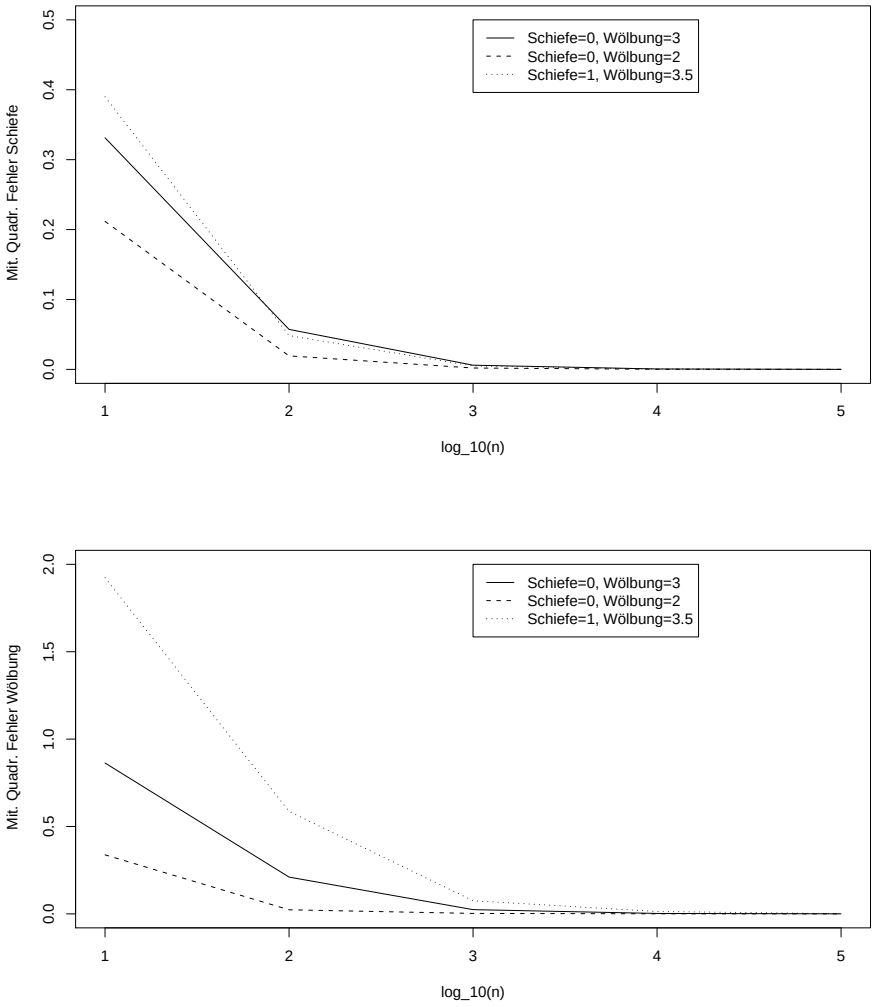

Abbildung 4.3: Mittlere quadratische Fehler bei Zufallszahlen, die mit den Fleishman-System erzeugt werden

4.2.3 Kollinearität der Variablen

Innerhalb der vereinfachenden linearen Abbildung nach McCulloch (1986) (siehe Seite 62) werden die Daten mittels $\Sigma_W^{-\frac{1}{2}}$ so skaliert, dass die transformierten Variablen linear unabhängig mit Kovarianzmatrix I verteilt sind. Da bei vielen Verfahren wie zum Beispiel der linearen Diskriminanzanalyse Σ_W^{-1} berechnet werden muss, ist es für das adaptive Verfahren wichtig zu wissen, wie sich die verschiedenen Verfahren bei unterschiedlichen Konditionen von Σ_W verhalten.

Es ist bekannt, dass der numerische Fehler, der bei der Inversion auftreten kann, von der Konditionszahl κ der Matrix Σ_W abhängt (Belsley *et al.*, 1980, Seite 173ff.). Dabei gilt

$$\kappa(\Sigma_W) := \frac{\max\{\lambda_i\}}{\min\{\lambda_i\}}, \tag{4.12}$$

wobei $\{\lambda_i\}$ die Menge der Eigenwerte von Σ_W bezeichnet. Allgemein ist die Kondition einer Matrix eine Maßzahl für die Kollinearität der Daten. Kollinearität entsteht dadurch, dass die einzelnen Variablen in $\mathcal{X}$ korreliert sind (Belsley *et al.*, 1980, S. 86). Um verschiedene Kovarianzmatrizen mit verschiedenen Konditonszahlen zu erzeugen, wird daher ein Zusammenhang zwischen der Konditionszahl κ und der durchschnittlichen Korrelation ρ zwischen den Variablen betrachtet.

Dabei wird im Folgenden angenommen, dass die Korrelation zwischen allen Variablen gleich ist. Im vorliegenden Fall der Untersuchung von vier Klassen in drei Variablen hat Σ_W dann folgende Gestalt:

$$\Sigma_W = \begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}. \tag{4.13}$$

Dann lassen sich die Eigenwerte von Σ_W berechnen als

$$\begin{aligned}
\|\Sigma_W - \lambda I\| &= \begin{vmatrix} 1-\lambda & \rho & \rho \\ \rho & 1-\lambda & \rho \\ \rho & \rho & 1-\lambda \end{vmatrix} \\
&= \begin{vmatrix} 1-\lambda & \rho & \rho \\ -1+\rho+\lambda & 1-\rho-\lambda & 0 \\ \rho & \rho & 1-\lambda \end{vmatrix} \\
&= \begin{vmatrix} 1+\rho-\lambda & \rho & \rho \\ 0 & 1-\rho-\lambda & 0 \\ 2\rho & \rho & 1-\lambda \end{vmatrix} \\
&= (1-\rho-\lambda) \begin{vmatrix} 1+\rho+\lambda & \rho \\ 2\rho & 1-\lambda \end{vmatrix} \\
&= (1-\rho-\lambda)\left[(1+\rho-\lambda)(1-\lambda) - 2\rho^2\right] \\
&= (1-\rho-\lambda)^2(1+2\rho-\lambda).
\end{aligned}$$

Damit ist $\lambda_1 = 1-\rho$ ein Eigenwert mit der Vielfachheit 2 und $\lambda_2 = 1+2\rho$ ein Eigenwert mit der Vielfachheit 1. Daraus folgt für $\rho \geq 0$, dass

$$\kappa = \frac{1+2\rho}{1-\rho}. \tag{4.14}$$

(Die Konditionszahl ist für $\rho = 1$ und $\rho = -0.5$ nicht definiert. In diesem Fall ist Σ_W singulär.) Damit kann ρ in Abhängigkeit von der gewünschten Kondition bestimmt werden:

$$\rho = \frac{\kappa - 1}{\kappa + 2}.$$

Kapitel 5

Selektorstatistik in der linearen Dimensionsreduktion

Nachdem in Kapitel 4 mögliche Einflussgrößen für die Güte eines Verfahrens zur linearen Dimensionsreduktion im Rahmen einer Klassifikation hergeleitet wurden, wird in diesem Kapitel die Entwicklung einer passenden Selektorstatistik beschrieben. Mit dieser wird aufgrund einer vorhandenen Datensituation im Klassifikationsproblem entschieden, welches der Verfahren

- Dimensionsreduktion mit dem Fisher-Kriterium
- Dimensionsreduktion mit dem Minimaler-Fehler-Kriterium
- Dimensionsreduktion mit dem Optimale-Separations-Projektion-Kriterium
- Dimensionsreduktion mit dem Vorhersageoptimale-Projektion-Kriterium

zur Anwendung kommen soll um eine möglichst geringe Fehlklassifikationsrate zu erreichen. Die Selektorstatistik wird mit Hilfe einer Simulationsstudie entwickelt. Im Abschnitt 5.2 wird der Aufbau der Simulationsstudie beschrieben. Anschließend wird in den Abschnitten 5.3 und 5.4 eine Auswertungsstrategie hergeleitet und vorgestellt.

5.1 Adaptive Verfahren

Grundidee eines adaptiven Verfahrens ist es, die wichtige und verfügbare Information aus dem vorliegenden Datensatz zu extrahieren, um dann mit Hilfe dieser Information ein geeignetes Verfahren auszuwählen [S. 2](Büning, 1991). Der hier verwendete Begriff der Datensituation (siehe Definition 16) entspricht dabei der wichtigen und verfügbaren Information im Sinne von Büning (1991). Das adaptive Vorgehen ist damit zweistufig:

1. Analyse der Daten D^n und aufgrund der dadurch ermittelten Datencharakteristik $g(D^n)$ Entscheidung für einen Klassifikator $\mathbf{a_n}^l$.
2. Anwendung (Schätzen) des Klassifikators $\mathbf{a_n}^l$ auf D^n.

Adaptive Vorgehen sind nicht neu in der statistischen Praxis. So wird häufig überprüft, ob die Daten den Modellannahmen entsprechen (Miller, 1986). Ist dies nicht der Fall, wird entweder die Auswertungsmethode verändert, oder die Daten werden geeignet transformiert. Diese Arbeitsschritte können im adaptiven Verfahren automatisiert und zusammen gefasst werden.

5.2 Methodik der Simulationsstudie

Allgemein können mit Hilfe einer Simulationsstudie auch in sehr komplexen Problemen Erkenntnisse über interessierende Zusammenhänge gewonnen werden, die mittels herkömmlicher analytischer Methoden nur sehr schwer oder gar nicht aufgedeckt werden können. Bei der Entwicklung eines adaptiven Verfahrens muss der Zusammenhang zwischen den Einflussvariablen, das heißt hier der Datensituation, mit der Güte der Verfahren festgestellt werden. Das Teilgebiet der statistischen Versuchsplanung (siehe z.B. Hinkelmann & Kempthorne (1994)) beschäftigt sich mit der Frage, wie mit minimalen Aufwand (Anzahl von Experimenten) möglichst viel Information über die Beziehung zwischen den Einflussvariablen und der Zielgröße

gewonnen werden kann. Es ist auch möglich, im Rahmen der statistischen Versuchsplanung mehrere Zielgrößen zu beachten, aber das wird hier nicht benötigt. Anders als in der „klassischen“ Versuchsplanung, in der physisch Experimente durchgeführt werden, müssen bei Experimenten, die ausschließlich mit einem Computer berechnet oder simuliert werden, Prinzipien wie Blockbildung oder Randomisation häufig nicht beachtet werden (Santner *et al.*, 2003, Seite 123).

Da im Vorhinein unbekannt ist, in welchem Unterraum der Datensituation welches Verfahren das Beste ist, sollten die möglichen Datensituationen gleichmäßig überprüft werden. Die Versuchspunkte sollten den Raum der möglichen Datensituationen daher gut repräsentieren. Dazu können raumfüllende Versuchspläne (Santner *et al.*, 2003, Seite 125) verwendet werden. Ein Beispiel für einen solchen Versuchsplan ist das Latin-Hypercube-Design.

Definition 18 *Sei jeder Einflussfaktor in m Zellen (Intervalle oder Stufen) eingeteilt. Ein Versuchsplan ist ein Latin-Hypercube-Design (LHD), wenn bei m Versuchspunkten jede Zelle jedes Faktors einmal Teil eines Versuchspunktes ist.*

Damit ist im LHD jede Randverteilung eines Faktor eine diskrete Gleichverteilung. Latin-Hypercube-Versuchspläne sind eine Erweiterung von lateinischen Quadraten (Hinkelmann & Kempthorne, 1994, S. 315ff.), bei denen in jeder Spalte und jeder Zeile alle Stufen auftreten, siehe auch Abbildung 5.1. LHDs werden häufig im Rahmen von Computerexperimenten verwendet (siehe z.B Welch *et al.* (1992)), allerdings füllt nicht jeder Latin-Hypercube-Versuchsplan den Raum gleichmäßig aus, siehe Beispiel 2 in Abbildung 5.1. Daher werden häufig mehrere LHDs generiert und aufgrund eines Kriteriums wird dann eines davon ausgewählt. In dieser Arbeit wird derjenige Versuchsplan verwendet, der den größten minimalen Abstand hat. Ein solcher Versuchsplan wird auch „maximin-Abstand-Versuchsplan“ (Santner *et al.*, 2003, Seite 138) genannt.

Definition 19 *Ein Versuchsplan* $\mathcal{D} \in \mathbf{D} \subset I\!R^{o \times d}$ *mit Versuchspunkten* $\{f_1, \ldots, f_o\}$,

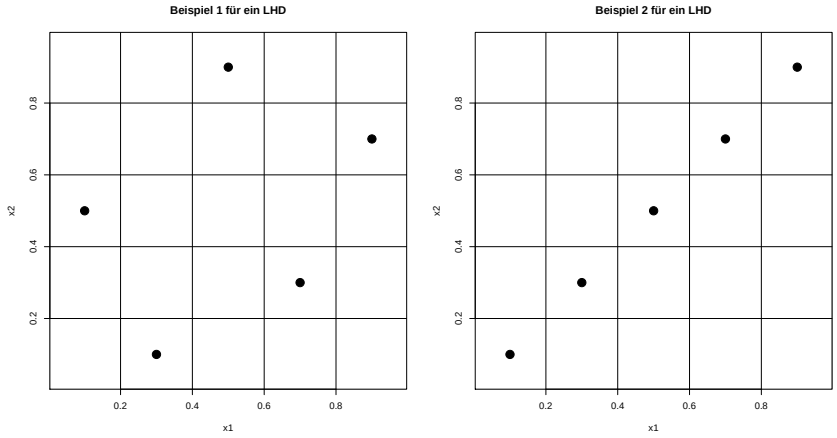

Abbildung 5.1: Beispiele für Latin Hypercube Designs bei 5 Versuchen und 2 Faktoren

$f_i \in \mathbb{R}^d$ ist ein maximin-Abstand-Versuchsplan ($\mathcal{D}_{Mm}$), wenn gilt

$$\min_{f_i, f_j \in \mathcal{D}_{Mm}} \delta_E(f_i, f_j) = \max_{\mathcal{D} \in \mathbf{D}} \min_{f_i, f_j \in \mathcal{D}} \delta_E(f_i, f_j). \tag{5.1}$$

Dabei ist $\mathbf{D}$ *die Menge aller zulässigen Versuchspläne.*

Im vorliegenden Fall ist $\mathbf{D}$ eine Menge, deren Elemente Latin-Hypercube-Versuchspläne sind, wobei gilt $\mathbf{D} \subset [0,1] \times [0,1] \times \cdots \times [0,1] \subset \mathbb{R}^d$.

Die interessierenden Einflussfaktoren sind (vergleiche Kapitel 4):

- Lage der klassenspezifischen Mittelwertsvektoren zueinander, die durch die ersten vier normierten zentralen (empirischen) Momente (4.7) beschrieben wird:
  - Mittlerer Abstand der Mittelwerte: f_1
  - Varianz der Abstände der Mittelwerte: f_2
  - Schiefe der Abstände der Mittelwerte: f_3
  - Wölbung der Abstände der Mittelwerte: f_4
- Kondition der (gemeinsamen) Kovarianzmatrix: f_5

- Anzahl der Beobachtungen: f_6
- Verteilung der Daten, variiert durch:
  - Schiefe der Verteilung innerhalb der Klassen: f_7
  - Wölbung der Verteilung innerhalb der Klassen: f_8

  Dabei ist die Verteilung innerhalb der Klassen für alle Klassen gleich. Daher unterscheiden sich die Klassen nur durch den Lokationsparameter.

Diese Einflussfaktoren sind Datencharakteristika von D^n mit der zugrundeliegenden Verteilung $\mathcal{F}$. Insbesondere haben diese Faktoren den Vorteil, dass sie auch bei vorliegenden Daten gemessen werden können und nicht nur im Rahmen einer Simulationsstudie variiert werden können.

Allerdings ist der Wertebereich, in dem Versuche durchgeführt werden können, wegen der inhärenten Eigenschaften von Abständen (keine negativen Abstände, Dreiecksungleichung) sowie an Schiefe und Wölbung nicht rechteckig. Daher müssen die Werte in $\mathcal{D}$ transformiert werden. Sei dazu $\mathcal{D}$ hier ein Versuchsplan mit Werten in $[0,1]^8$, dann werden die einzelnen Werte im Versuchsplan wie folgt transformiert um sinnvolle Werte bei den Einflussfaktoren zu erzielen:

- $f_1 \to 1 + f_1 \cdot (-4\Phi^{-1}(0.1) - 1)$, wobei Φ die Verteilungsfunktion der Standardnormalverteilung ist. Mit dieser Transformation wird erreicht, dass der minimale mittlere Abstand der Klassenmitten mindestens 1 ist, während im maximalen Fall der mittlere Klassenabstand 5.13 ist. Damit ist die mittlere Fehlklassifikationswahrscheinlichkeit in diesem Fall zwischen 2 Klassen ca. 0.005 (vergleiche (2.32)).
- $f_2 \to f_2 \cdot \frac{2}{3} f_1^2$. Damit entspricht das Maximum des Wertebereiches von f_2 der Varianz die vorliegt, wenn alle Klassenmitten gleichmäßig verteilt auf einer Geraden liegen.
- $f_3 \to f_3 \cdot 2\sqrt{3 + \frac{1}{5}} - \sqrt{3 + \frac{1}{5}}$. Damit wird der Wertebereich von f_3 auf die maximal mögliche Schiefe bei 4 Klassenmitten bzw. 6 Abständen eingeschränkt.

- $f_4 \rightarrow f_4 \cdot (3 + 0.4 \cdot f_3^2 - (f_3^2 + 1)) + (f_3^2 + 1)$. Mit dieser Transformation von f_4 wird der in Abhängigkeit von f_3 mögliche Wertebereich der Wölbung von 6 Abständen approximiert.

- $f_5 \rightarrow [\sqrt{f_5} \cdot (10^5 - 1) + 1]$. f_5 ist nach der Transformation dreiecksverteilt mit Werten zwischen 1 und 10^5, wobei hohe Konditonszahlen von Σ_W häufiger vorkommen, da vermutlich die Verfahren in diesen Bereich größere Schwierigkeiten haben werden.

- $f_6 \rightarrow [500 - \sqrt{f_6} \cdot (500 - 10)]$. Mit dieser Transformation wird eine Dreieicksverteilung von f_6 zwischen 10 und 500 erzielt, so dass im vermutlich schwierigeren Bereich mit geringen Fallzahlen mehr Versuchspunkte liegen.

- $f_7 \rightarrow f_7 \cdot 3$. Die Schiefe der Daten variiert damit zwischen 0 und 3.

- $f_8 \rightarrow f_8 \cdot (20 - (f_7^2 + 1)) + (f_7^2 + 1)$. Der Wertebereich der Wölbung der Daten hängt nach (4.8) von der Schiefe der Daten ab. Die maximale Wölbung wurde auf 20 festgesetzt.
  Wenn man den möglichen Wertebereich für die Schiefe und Wölbung der Daten mit Abbildung 4.2 vergleicht wird deutlich, wie groß die möglichen Abweichungen von der Normalverteilung im vorliegenden LHD sein können.

Ein realisierter Versuchsplan nach diesen Transformationen ist weiter hinten in Abbildung 6.1 zu finden.

5.3 Vergleich von Klassifikationsverfahren

Erste Untersuchungen zum Vergleich verschiedener Klassifikationsverfahren finden sich in Dietterich (1998), wobei dort mehrere gebräuchliche Testverfahren empirisch evaluiert werden. In Hothorn *et al.* (2005) wird ein theoretischer Rahmen für solche Tests vorgestellt, der hier für die vorliegende Situation angepasst und präzisiert wird.

Ausgangspunkt ist die bedingte Fehlerrate (3.3)

$$\mathcal{E}^c(\mathbf{a_n}, D^n) = \sum_{k=1}^{K} P(\mathbf{a_n}(\mathcal{X}) \neq k | \mathcal{K} = k, D^n) P(\mathcal{K} = k),$$

eine Zufallsvariable, deren Verteilung von den Klassifikationsverfahren $\mathbf{a_n}$ und den Daten D^n, die unabhängig identisch verteilt aus der Verteilung $\mathcal{F}$ gezogen werden, abhängt.

Aussagen über die Verteilung der Fehlerraten sind dann mit Hilfe von mehreren Stichproben D^n möglich (Hothorn *et al.*, 2005), die alle aus derselben Verteilung $\mathcal{F}$ von $(\mathcal{X}, \mathcal{K})$ gewonnen werden. Damit können, bedingt $\mathcal{F}$ (bei Hothorn *et al.* (2005) datengenerierender Prozess genannt) klassische statistische Methoden für die Verteilung $\mathcal{E}^c(\mathbf{a_n})$ des Klassifikators $\mathbf{a_n}$ angewendet werden. Insbesondere können die Verteilungen $\mathcal{E}^c(\mathbf{a_n^i}), i = 1, \ldots, L$, von L Klassifikatoren bedingt die Verteilung $\mathcal{F}$ verglichen werden (Hothorn *et al.*, 2005).

Um ein optimales adaptives Verfahren zu entwickeln, muss für vorliegende Verteilungen $\mathcal{F}$ getestet werden, ob ein, und wenn ja welcher der zur Verfügung stehenden Klassifikatoren $\mathbf{a_n}^i, i = 1, \ldots, L$ der signifikant beste ist. Da die Realisierungen $e^c(\mathbf{a_n^i})$der Zufallsvariablen $\mathcal{E}^c(\mathbf{a_n^i}, D^n)$ nicht beobachtet werden können, sondern diese selber geschätzt werden müssen (siehe Abschnitt 3.2), ist ein direkter Vergleich von $\mathcal{E}^c(\mathbf{a_n^i}, D^n)$ nicht möglich. Daher wird der erwartete Fehler

$$E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^i)), \quad i = 1, \ldots, L, \tag{5.2}$$

gegeben Klassifikator $\mathbf{a_n}^i$ über die Verteilung $\mathcal{F}$ untersucht. Der Test für ein bestes Verfahren wird über die Hypothesen

$$H_0 \quad : E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^1)) = E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^2)) = \cdots = E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^L)) \tag{5.3}$$

vs.

$$H_1 \quad : \exists l \in \{1, ..., L\} : \quad E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^l)) < E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^j)) \quad \forall j \neq l \tag{5.4}$$

entwickelt. Dies ist ein multipler Test (siehe z.B. Hochberg & Tamhane (1987)), der

aus $L-1$ einzelnen Test besteht:

$$\forall \quad j \neq l, \quad l,j \in \{1,...,L\} : H_0^j : E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^l)) \geq E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^j)) \tag{5.5}$$

vs.

$$\forall \quad j \neq l, \quad l,j \in \{1,...,L\} : H_1^j : E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^l)) < E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^j)). \tag{5.6}$$

Um ein globales Testniveau mit einer Irrtumswahrscheinlichkeit α_G einzuhalten, muss beachtet werden, dass

a) Potentiell jedes der L Verfahren das beste sein kann,

b) Das beste Verfahren mit $L-1$ anderen Verfahren verglichen wird. Die Hypothese H_0^j in (5.5) daher $L-1$ mal getestet wird.

Damit das globale Fehlerniveau eingehalten werden kann, wird das Niveau der einzelnen Hypothesen (5.5) in b) mit der Methode nach Bonferroni-Holm (Holm, 1979) angepasst: Die einzelnen Hypothesen werden jeweils zum Niveau

$$\frac{\alpha}{L-1}, \frac{\alpha}{L-2}, \cdots, \frac{\alpha}{1}$$

getestet, wobei die Hypothesen H_0^j bei vorliegenden Daten entsprechend ihrer p-Werte aufsteigend angeordnet werden (Holm, 1979).
Um a) zu entsprechen, wird das Niveau für die Hypothesen durch

$$\alpha = \frac{\alpha_G}{L} \tag{5.7}$$

angepasst.

Als Statistik zum Testen der Hypothese

$$H_0^j : E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^l)) \geq E_{\mathcal{F}}(\mathcal{E}^c(\mathbf{a_n}^j))$$

kann der Mittelwert von t Schätzern, die mit Hilfe der Trainings- und Testmethode (siehe (3.5), Seite 51) $\widehat{e^c_{tt^m}}(\mathbf{a_n}^j)$,

$$\overline{\widehat{e^c_{tt^m}}(\mathbf{a_n}^j)} = \frac{1}{t}\sum_{i=1}^{t} \widehat{e^c_{tt_i^m}}(\mathbf{a_n}^j, D_i^n) \tag{5.8}$$

herangezogen werden, da unter H_0^j gilt

$$\begin{aligned}
E(\overline{\widehat{e^c_{tt^m}}(\mathbf{a_n}^j)}) &= \frac{1}{t}\sum_{i=1}^{t} E(\widehat{e^c_{tt_i^m}}(\mathbf{a_n}^j, D_i^n)) \\
&= \frac{1}{t}\sum_{i=1}^{t} E(E(\widehat{e^c_{tt_i^m}}(\mathbf{a_n}^j, D_i^n)|\mathcal{E}^c(\mathbf{a_n}^j, D_i^n) = e^c(\mathbf{a_n}^j, D_i^n))) \\
&= \frac{1}{t}\sum_{i=1}^{t} E(e^c(\mathbf{a_n}^j, D_i^n)) \\
&= \frac{1}{t}\sum_{i=1}^{t} E(e^c(\mathbf{a_n}^l, D_i^n)) \\
&= E(\overline{\widehat{e^c_{tt^m}}(\mathbf{a_n}^l)}).
\end{aligned}$$

Da keinerlei Information über die Verteilung von $\mathcal{E}^c(\mathbf{a_n}^j, D_i^n)$ vorliegt, kann kein parametrischer Test verwendet werden. Außerdem wird vermutet, dass die geschätzten Fehlerraten $\widehat{e^c_{tt_i^m}}(\mathbf{a_n}^j, D_i^n)$ schief verteilt sind (Hothorn *et al.*, 2005). Um den Einfluss der Varianz in den Trainingsdaten D_i^n bzw. in den Testdaten T_i^m beim Vergleich zu minimieren, werden alle Verfahren j bei der i-ten Stichprobe von $\widehat{e^c_{tt_i^m}}(\mathbf{a_n}^j, D_i^n)$ auf den gleichen Trainings- und Testdaten verglichen. Damit sind die einzelnen $\widehat{e^c_{tt_i^m}}(\mathbf{a_n}^j, D_i^n)$ für die Verfahren j nicht mehr unabhängig. Trotzdem kann die Hypothese mit Hilfe eines Permutationstests für verbundene Stichproben überprüft werden (Edgington, 1995). Bei dem hier verwendeten Permutationstest wird jeweils die Zuordnung der Ergebnisse zu den Verfahren bei den t Realisierungen von $\widehat{e^c_{tt_i^m}}(\mathbf{a_n}^j, D_i^n)$ je Datensatz D_i^n, T_i^m permutiert. Unter der Nullhypothese

$$E(\mathcal{E}^c(\mathbf{a_n}^1)) = E(\mathcal{E}^c(\mathbf{a_n}^2)) = \cdots = E(\mathcal{E}^c(\mathbf{a_n}^L))$$

ändert sich der Erwartungswert der Teststatistik für H_0^j

$$\overline{\widehat{e^c_{tt^m}}(\mathbf{a_n}^l)} - \overline{\widehat{e^c_{tt^m}}(\mathbf{a_n}^j)}, \quad \forall j \neq l, j \in \{1, ..., L\} \tag{5.9}$$

bei Permutation der Verfahren nicht (siehe oben). Der Anteil an Permutationen der Verfahren, die einen gleichen oder größeren Wert als die im Experiment errechnete Teststatistik ergeben, ist dann der p-Wert für die Nullhypothese. Da es bei dem

Vergleich zweier Verfahren und t Beobachtungen 2^t Permutationen gibt, kann es nötig sein, nur eine zufällige Auswahl der Permutationen zu berechnen. Ist der p-Wert kleiner als das vorgegebene Signifikanzniveau, kann die Alternative,

$$E(\mathcal{E}^c(\mathbf{a_n}^l)) < E(\mathcal{E}^c(\mathbf{a_n}^j)), \quad \forall j \neq l, j \in \{1, ..., L\}$$

zum jeweiligen Niveau angenommen werden. Beachte, dass die paarweisen Hypothesen

$$H_0^j : E(\mathcal{E}^c(\mathbf{a_n}^l)) \geq E(\mathcal{E}^c(\mathbf{a_n}^j))$$

mächtiger als die Hypothesen

$$E(\mathcal{E}^c(\mathbf{a_n}^l)) = E(\mathcal{E}^c(\mathbf{a_n}^j))$$

sind. So ist nicht nur eine Permutationen der Verfahren unter der Nullhypothese möglich, sondern unter der Nullhypothese sind auch kleinere berechnete Werte für $\overline{\widehat{e^c_{ttm}}(\mathbf{a_n}^j)}$ bzw. größere berechnete Werte für $\overline{\widehat{e^c_{ttm}}(\mathbf{a_n}^l)}$ erlaubt. Wenn nur die Verfahren permutiert und nicht zusätzlich die berechneten Werte verändert werden, wird der Test konservativ (Edgington, 1995, S. 322), so dass das Niveau eingehalten wird.

Insgesamt ist es mit diesem theoretischen Rahmen möglich, statistisch signifikante Aussagen über die Güte von Verfahren bei gegebener Datensituation zu treffen. Das Vorgehen zum Verfahrensvergleich bei einer gegebenen Datensituation kann wie folgt zusammengefasst werden:

1. Generiere t Realisierungen $D_i^n, T_i^m, i = 1, \ldots, t$, der Verteilung $\mathcal{F}$ nach der gewünschten Datensituation.
2. Schätze die L Verfahren auf den Daten $(\mathbf{a_n}^j, D_i^n), i = 1, \ldots, t, j = 1, \ldots, L$.
3. Schätze die bedingte Fehlerrate der Verfahren $\widehat{e^c_{tt_i^m}}(\mathbf{a_n}^j, D_i^n), i = 1, \ldots, t, j = 1, \ldots, L$.
4. Bestimme das beste Verfahren l als $l = \arg\min_j \overline{\widehat{e^c_{ttm}}(\mathbf{a_n}^j)}$.

5. Überprüfe die Hypothese $H_0^j : E(\mathcal{E})^c(\mathbf{a_n}^l) \geq E(\mathcal{E})e^c(\mathbf{a_n}^j)$ gegen die Alternative $H_1^j : E(\mathcal{E})^c(\mathbf{a_n}^l) < E(\mathcal{E})^c(\mathbf{a_n}^j)$ für alle $j \neq l, j \in \{1, ..., L\}$, mit Hilfe eines Permutationstestes.
   Notiere die p-Werte

$$P_j = \hat{P}\left(\widehat{e^c_{ttm}(\mathbf{a_n}^l)} - \widehat{e^c_{ttm}(\mathbf{a_n}^j)} \leq 0 | E(\mathcal{E}^c(\mathbf{a_n}^l)) \geq E(\mathcal{E}^c(\mathbf{a_n}^j))\right).$$

6. Ordne die p-Werte an:

$$P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(L-1)},$$

   und vergleiche diese nach der Methode nach Bonferroni-Holm mit dem jeweiligen Signifikanzniveau. Wenn

$$P_{(1)} < \frac{\alpha}{L-1} \wedge P_{(2)} < \frac{\alpha}{L-2} \wedge \cdots \wedge P_{(L-1)} < \frac{\alpha}{1}$$

   gilt, dann ist Verfahren l zum Niveau αL das beste Verfahren bei der vorliegenden Datensituation.

5.4 Bestimmung der Selektorstatistik

Nachdem in Abschnitt 5.2 der Aufbau der Simulationsstudie zur Erzeugung von verschiedenen Datensituationen, welche durch die Versuchspunkte $f \in \mathcal{D}$ beschrieben werden, vorgestellt wurde, und in Abschnitt 5.3 eine Auswertungsstrategie je Datensituation f entwickelt wurde kann nun ein Verfahren zur Entwicklung der Selektorstatistik vorgeschlagen werden:

Die möglichen Datensituationen von D^n mit Verteilung $\mathcal{F}$ werden in dem Versuchsplan $\mathcal{D}$ variiert. Für jeden Versuchspunkt $f \in \mathcal{D}$ wird überprüft, ob es ein zum Niveau α signifikantes bestes Verfahren $l \in \{1, \ldots, L\}$ gibt. Eine Selektorstatistik $\mathbf{S}$ ist eine Funktion von D^n, deren Wertemenge eine Einteilung der zu verwendenden Verfahren ermöglicht (Büning, 1991). In der vorliegenden Arbeit werden die Dimensionsreduktionskriterien bezüglich ihrer Fehlerrate bei gegebener Datensituation

eingeteilt. Das gesamte Vorgehen (Bestimmung der Selektorstatistik und Einteilung der Wertemenge) kann als Klassifikation im Sinne von Abschnitt 2.1.1 verstanden werden:

$$\mathbf{s} : \mathbf{s}(g(D^n)) : I\!R^{n\times(d+1)} \rightarrow \{1, \ldots, L\}, \tag{5.10}$$

wobei $g(D^n)$ die Datencharakteristika von D^n ermittelt, also hier

$$g : g(D^n) : I\!R^{n\times(d+1)} \rightarrow I\!R^8 \tag{5.11}$$

mit

$$g_1(D^n) = \eta_1(\bar{x}) = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}\hat{\delta}_E(i,j), \tag{5.12}$$

$$g_2(D^n) = \eta_2(\bar{x}) = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}(\hat{\delta}_E(i,j) - \eta_1(\bar{x}))^2, \tag{5.13}$$

$$g_3(D^n) = \eta_3(\bar{x}) = \frac{\frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}(\hat{\delta}_E(i,j) - \eta_1(\bar{x}))^3}{\sqrt{\eta_2(\bar{x})}^3}, \tag{5.14}$$

$$g_4(D^n) = \eta_4(\bar{x}) = \frac{\frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1,j\neq i}^{K}(\hat{\delta}_E(i,j) - \eta_1(\bar{x}))^4}{\sqrt{\eta_2(\bar{x})}^4}, \tag{5.15}$$

$$g_5(D^n) = \kappa(\hat{\Sigma}_W), \tag{5.16}$$

$$g_6(D^n) = \frac{n}{K}, \tag{5.17}$$

$$g_7(D^n) = \sum_{i=1}^{K}\hat{\pi}_i\sum_{j=1}^{K-1}|\frac{1}{n_i}\sum_{m=1}^{n_i}\frac{(x_m^j - \bar{x}_i^j)^3}{\sqrt{\frac{1}{n_i}\sum_{m=1}^{n_i}(x_m^j - \bar{x}_i^j)^2}^3}|, \tag{5.18}$$

$$g_8(D^n) = \sum_{i=1}^{K}\hat{\pi}_i\sum_{j=1}^{K-1}|\frac{1}{n_i}\sum_{m=1}^{n_i}\frac{(x_m^j - \bar{x}_i^j)^4}{\sqrt{\frac{1}{n_i}\sum_{m=1}^{n_i}(x_m^j - \bar{x}_i^j)^2}^4}|. \tag{5.19}$$

Bis auf $g_5(D^n)$ werden alle Datencharakteristika berechnet, nachdem die vereinfachende lineare Abbildung nach McCulloch (1986) angewendet wurde (siehe auch Seite 62).

Als Selektor (bzw. Klassifikator) $\mathbf{s}$ aufgrund der Statistik $\mathbf{S} = g(D^n)$ wird im Rahmen des StatLog Projektes ein Entscheidungsbaum (Breiman *et al.*, 1984) verwendet, siehe Michie *et al.* (1994, S. 201f.). Entscheidungsbäume haben den Vorteil, dass

sie eine Interpretation der Ergebnisse erleichtern (Hastie *et al.*, 2001, S. 267). Das Ergebnis eines Entscheidungsbaumes kann dann in Regeln umgeschrieben werden (Michie *et al.*, 1994, S. 202). Ein weiterer Vorteil ist, dass Entscheidungsbäume auf keiner Annahme über die Verteilung der Daten beruhen, insbesondere werden keine unimodalen Daten vorausgesetzt. Aus diesen Gründen wird die Selektorstatistik in dieser Arbeit automatisch mit Hilfe eines Entscheidungsbaumes gewonnen. Dieser Entscheidungsbaum wird mit den unterschiedlichen Datensituationen des Latin-Hypercube-Design gelernt und kann dann zur Bestimmung des optimalen Verfahrens auf den jeweils vorliegenden Daten angewendet werden.

Kapitel 6

Adaptive lineare Dimensionsreduktion

In den vorhergehenden Kapiteln wurden Verfahren zur Klassifikation und Dimensionsreduktion vorgestellt und diskutiert. Damit aus diesem Verfahrenspool ohne aufwändige Versuche das passende Verfahren für einen gegebenen Datensatz ausgewählt werden kann, wurde die Methodik zur Entwicklung eines adaptiven Verfahrens beschrieben (Kapitel 5). Im vorliegenden Kapitel wird ein adaptives Verfahren für die Klassifikation im 4-Klassen-Fall entwickelt und evaluiert. Dabei werden die Annahmen (A1), (A2) und (A3) (siehe S. 7 und 61) sowie

(A4) Die Verteilung der Daten innerhalb der Klassen unterscheidet sich durch den Lokationsparameter

verwendet. Dabei ist (A4) eine stärkere Annahme als die Annahme identischer Kovarianzmatrizen (A3).

6.1 Design der Simulationsstudie

Zur Entwicklung des adaptiven Verfahrens werden drei unabhängige Latin-Hypercube Versuchspläne verwendet: Die Ergebnisse des ersten Versuchsplanes werden zum Schätzen (Lernen) der Selektorstatistik verwendet, wohingegen die Ergebnisse des zweiten zum Validieren herangezogen werden. Mit Hilfe der Ergebnisse des dritten Versuchsplanes wird das adaptive Verfahren unabhängig von den Daten, die zur Entwicklung verwendet wurden, getestet.

Trotz der Transformationen innerhalb eines Versuchsplanes (siehe Seite 79) ist nur ca. jeder vierte Versuchspunkt realisierbar. Häufigste Ursache hierfür ist, dass die Varianz bei den Abständen der Klassenmitten im Verhältnis zum mittleren Klassenabstand zu groß ist, so wird entweder die Dreiecksungleichung nicht eingehalten, oder ein negativer Abstand müsste verwendet werden. Die zweidimensionalen Randverteilungen der realisierten 271 Versuchspunkte eines Versuchsplanes mit 1000 geplanten Versuchspunkten zeigt Abbildung 6.1.

Man kann in Abbildung 6.1 sehr gut die Abhängigkeit der prinzipiell möglichen Realisierungen von Schiefe (g_3) und Wölbung (g_4) bei 6 Abständen erkennen. Außerdem sieht man, dass mehr Versuche bei höheren Konditionszahlen (g_5) sowie geringeren Anzahl an Beobachtungen zum Schätzen (g_6) der Verfahren durchgeführt wurden (vergleiche Seite 79). Der realisierte Versuchsplan zum Schätzen der Selektorstatistik besteht aus 1066 Versuchspunkten für unterschiedliche Datensituationen. Der ursprüngliche Versuchsplan enthielt 4000 Versuchspunkte. Bei den Versuchsplänen zum Validieren und Testen der Selektorstatistik und des adaptiven Verfahrens wurden 285 bzw. 271 Versuchspunkte von 1000 ursprünglichen Versuchspunkten eines maximin-Abstand Latin-Hypercube-Versuchsplans realisiert. Gemäß der Datensituation, die ein Versuchspunkt beschreibt, wurden die Daten erzeugt und die Fehlerrate der Verfahren auf 4000 unabhängigen Testbeobachtungen geschätzt (vergleiche Abschnitt 3.2). Die Anzahl der Trainingsbeobachtungen variierte je nach Datensituation. Nach Gleichung (3.6) ist damit die Wahrscheinlichkeit, dass die wah-

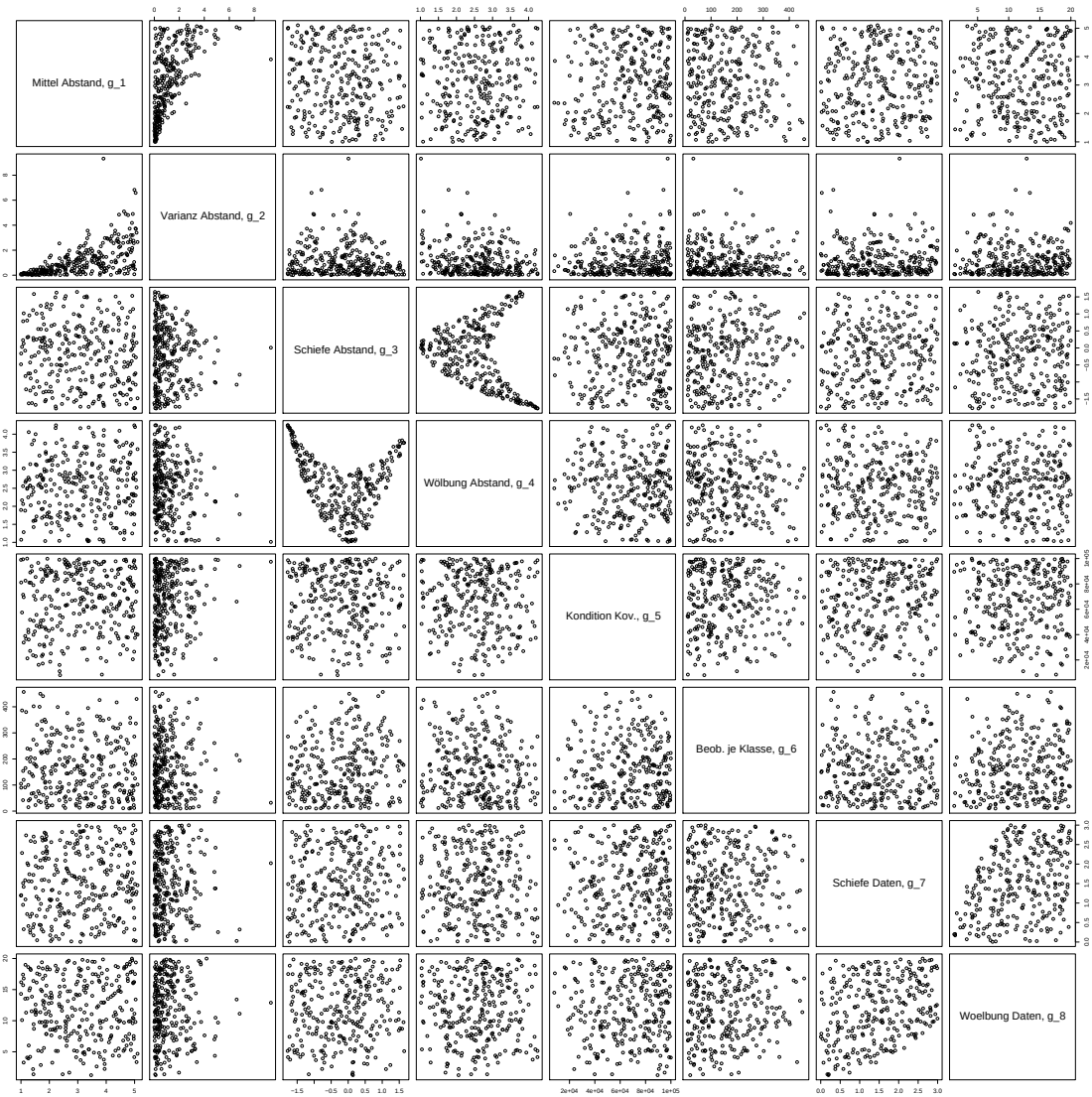

Abbildung 6.1: Realisierte Versuchspunkte des Validierungs-Versuchsplanes

re, bedingte Fehlerrate um mehr als 0.025 von der geschätzten Fehlerrate abweicht, kleiner als 0.015. Damit sind die geschätzten Fehlerraten hinreichend genau für die Entwicklung der Selektorstatistik. Außerdem wurde jeder Versuchspunkt 30 mal wiederholt, so dass insgesamt $30 \cdot (1066 + 285) = 40530$ Datensätze zur Entwicklung des adaptiven Verfahrens verwendet wurden. Die Versuchspunkte decken ein breites Spektrum der möglichen Datensituationen ab.

6.2 Ergebnisse für die Selektorstatistik

Aufgrund der vorgeschalteten Anwendung der vereinfachenden lineare Abbildung nach McCulloch (1986) (siehe Seite 62) und Satz 4 müssen im hier untersuchten Fall von vier Klassen nur die Selektorstatistiken bei einer Dimensionsreduktion auf $r = 1$ bzw. $r = 2$ Dimensionen betrachtet werden. In den meisten Anwendungen wird eine Projektion auf $r = 2$ Dimensionen verwendet werden. Trotzdem kann es in bestimmten Fällen (z.B. extreme Multikollinearität, Dichteschätzungen im projezierten Raum) sinnvoll sein, auch eine Projektion auf nur eine Dimension zu untersuchen. Daher wird die Selektorstatistik auch für diesen Fall entwickelt, so dass der Anwender bei Bedarf auch hier eine adaptive lineare Dimensionsreduktion durchführen kann.

6.2.1 Selektorstatistik für die Projektion auf eine Dimension

In 469 von 1066 Datensituationen, gibt es bei der Projektion auf eine Dimension ein zum Niveau 0.1 signifikant bestes Verfahren. Tabelle 6.1 zeigt, dass zur Projektion auf eine Dimension in den meisten Datensituationen das Minimale-Fehler-Kriterium zum besten Ergebnis führt. Am zweit häufigsten ist das Optimale-Separations-Kriterium das beste Kriterium, während die anderen beiden Kriterien seltener das signifikant beste Ergebnis erreichen. Allerdings sollte dieses Ergebnis

Methode	Häufigkeit
Fisher-Kriterium	41
MFK	258
OSP	139
VOP	31

Tabelle 6.1: Häufigkeiten als bestes Verfahren bei Projektion auf eine Dimension

nicht überbewertet werden, da das beste Kriterium von der jeweils vorliegenden Datensituation abhängt, und der verwendete Versuchsplan – obwohl er ein breites Spektrum von Datensituationen abdeckt – nicht repräsentativ für real vorkommende Datensituationen sein muss. Daher ist die Selektorstatistik, die die Verfahren adaptiv auswählt, von größerem Interesse. Der Entscheidungsbaum, der der Selektorstatistik zugrunde liegt, ist in Abbildung 6.2 angegeben. Man erkennt, dass sowohl Datencharakteristika der Verteilung (z.B. Wölbung der Daten, g_8) als auch Charakteristika, die die Abstände der Klassenmitten beschreiben (g_1 – g_4), bei der Zuordnung der besten Methode zu der gegebenen Datensituation wichtig sind, da sie in der Selektorstatistik in Abbildung 6.2 verwendet werden.

Aufgrund der Häufigkeiten für das beste Verfahren (vergleiche Tabelle 6.1), ist es nicht verwunderlich, dass das Minimale-Fehler-Kriterium zur Dimensionsreduktion am häufigsten in den Blättern des Entscheidungsbaumes (Abbildung 6.2) das beste Kriterium darstellt. Andere Kriterien führen vor allem in Datensituationen mit einer geringen Anzahl von Trainingsbeobachtungen (g_6) oder bei der Kombination einer hohen Wölbung (g_8) und geringen Schiefe (g_7) in den Verteilungen der Daten zum besten Ergebnis. Insgesamt überwiegt bei einer Wölbung nach Transformation von über 12.42 leicht das OSP-Kriterium als bestes Kriterium, während bei einer geringeren Wölbung das MFK-Kriterium vorzuziehen ist.

Die Fehlklassifikationsrate der adaptiven Selektorstatistik auf den Versuchsplänen zur Validierung beträgt 0.24. Die Aussagekraft dieser Fehlerrate ist aber nur begrenzt, da sich das adaptive Verfahren unter Umständen nur für ein leicht schlechte-

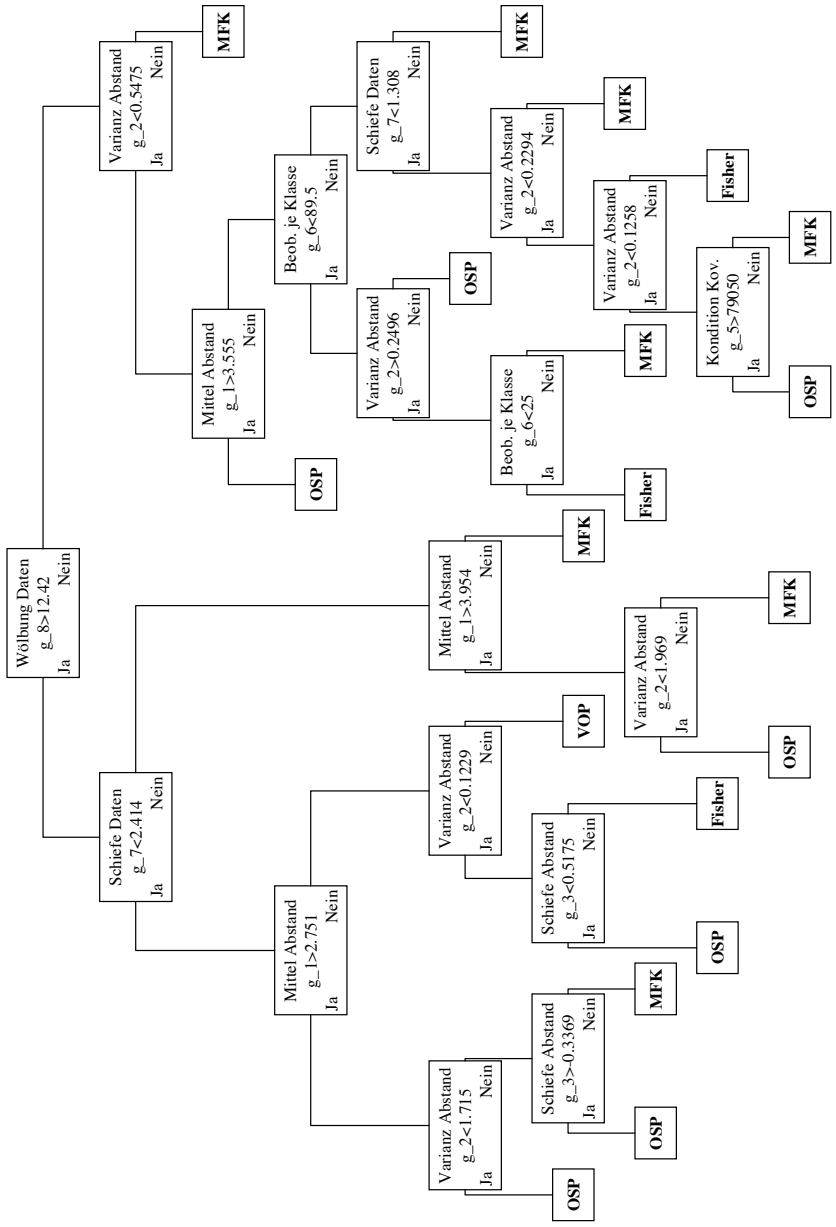

Abbildung 6.2: Selektorstatistik bei Projektion auf eine Dimension

res Dimensionsreduktionsverfahren entscheidet und damit die Unterschiede bei der Anwendung nur minimal sind. Daher wird auf dem Testdatensatz überprüft, inwieweit das adaptive Verfahren den anderen Verfahren insgesamt überlegen ist. Konstruktionsbedingt wird das adaptive Verfahren in keiner Datensituation das alleinig beste sein, da immer eines der bekannten Verfahren verwendet wird. Die Hypothesen zum Testen einer globalen Überlegenheit des adaptiven Vorgehen lauten:

$$\forall \quad j \in \{\text{Fisher}, \text{MFK}, \text{OSP}, \text{VOP}\} : H_0^j : E_{(\cdot)}(\mathcal{E}^c(\mathbf{a_n}^{\text{adapt}})) \geq E_{(\cdot)}(\mathcal{E}^c(\mathbf{a_n}^j))$$

vs.

$$\forall \quad j \in \{\text{Fisher}, \text{MFK}, \text{OSP}, \text{VOP}\} : H_1^j : E_{(\cdot)}(\mathcal{E}^c(\mathbf{a_n}^{\text{adapt}}) < E_{(\cdot)}(\mathcal{E}^c(\mathbf{a_n}^j)).$$

Diese Hypothesen werden mit Hilfe eines Permutationstests (siehe Seite 80) überprüft. Beachte, dass der Test nicht mehr bedingt die Verteilung $\mathcal{F}$ der Datensituation D^n durchgeführt wird, sondern über alle im Versuchplan auftretenden Datensituationen. Dadurch wird die globale Güte der adaptiven Dimensionsreduktion getestet. Die zusammengesetzte Hypothese H_0^j wird zum Niveau 0.005 abgelehnt. Damit ist gezeigt, dass das adaptive Verfahren über den vom Versuchsplan aufgespannten Raum der Datensituationen das signifikant beste Verfahren ist.

6.2.2 Selektorstatistik für die Projektion auf zwei Dimensionen

Bei der Projektion auf zwei Dimensionen gibt es in 417 von 1066 Datensituation ein zum Niveau 0.1 signifikant bestes Verfahren (Tabelle 6.2).

Der Entscheidungsbaum, der der Selektorstatistik für eine Projektion auf zwei Dimensionen zugrunde liegt, ist in Abbildung 6.3 gezeigt. Auch bei der Projektion auf zwei Dimensionen spielen sowohl Datencharakteristika der Verteilung als auch die Lage der Klassenmitten zueinander eine Rolle. Insgesamt sind aber die Datencharakteristika der Verteilung der Abstände der Klassenmitten zueinander (g_1 – g_4) näher an der Wurzel des Baumes zu finden als im Fall der Projektion auf eine

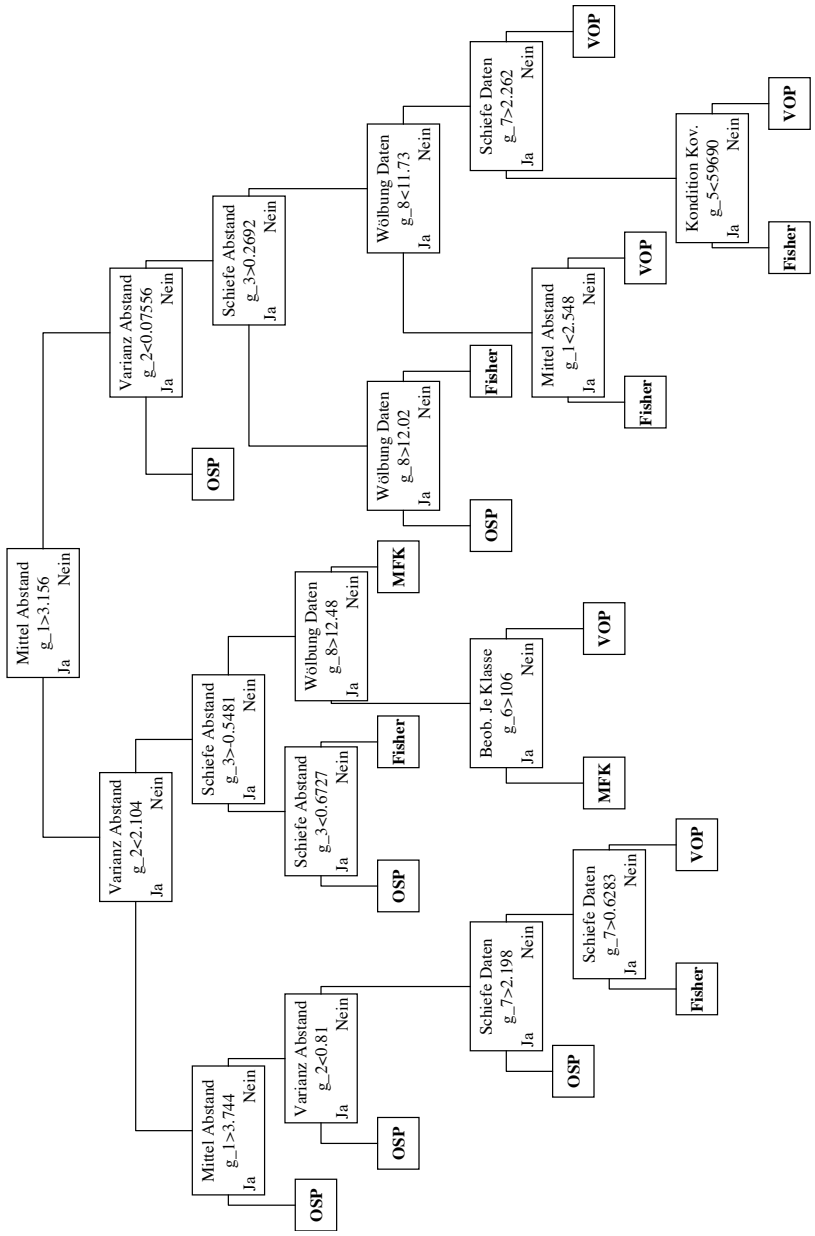

Abbildung 6.3: Selektorstatistik bei Projektion auf zwei Dimensionen

Methode	Häufigkeit
Fisher Kriterium	70
MFK	26
OSP	257
VOP	64

Tabelle 6.2: Häufigkeiten als bestes Verfahren bei Projektion auf zwei Dimensionen

Dimension (vergleiche Abbildung 6.3 mit Abbildung 6.2). Bei einem mittleren Abstand der Klassenmitten (g_1) nach Transformation von über 3.156 überwiegt das OSP Kriterium, während es bei einem geringeren Abstand der Klassenmitten nur bei sehr geringer Varianz der Abstände (g_2), also einem annähernd gleichseitigen Simplex, den anderen Kriterien überlegen ist.

Die Fehlerrate auf den Datensituationen zur Validierung der Selektorstatistik beträgt 0.28, ist also leicht schlechter als bei der Selektorstatistik für die Projektion auf eine Dimension. Darüber hinaus wird auch bei der Projektion auf zwei Dimensionen die globale Güte des adaptiven Verfahrens mit Hilfe eines Permutationstests überprüft. Um die zusammengesetzte Hypothese zu testen werden dabei wiederum die Ergebnisse der einzelnen Datensituationen verwendet.

$$\forall \quad j \in \{\text{Fisher}, \text{MFK}, \text{OSP}, \text{VOP}\} : H_0^j : E_{(\cdot)}(\mathcal{E}^c(\mathbf{a_n}^{\text{adapt}})) \geq E_{(\cdot)}(\mathcal{E}^c(\mathbf{a_n}^j))$$

$vs.$

$$\forall \quad j \in \{\text{Fisher}, \text{MFK}, \text{OSP}, \text{VOP}\} : H_1^j : E_{(\cdot)}(\mathcal{E}^c(\mathbf{a_n}^{\text{adapt}}) < E_{(\cdot)}(\mathcal{E}^c(\mathbf{a_n}^j)).$$

Auch bei der Projektion auf zwei Dimensionen kann die Hypothese der nicht-Überlegenheit des adaptiven Verfahrens zum Niveau 0.005 abgelehnt werden. Es zeigt sich also auch in diesen Fall eine signifikante globale Überlegenheit des adaptiven Vorgehens.

6.3 Zusammenfassung Selektorstatistik und adaptives Verfahren

Die Regeln, die einer Selektorstatistik im Fall von vier Klassen für die Dimensionsreduktion zugrunde liegen, sind komplex (siehe die Entscheidungsbäume in Abbildung 6.2 und 6.3). Von den in Kapitel 4 hergeleiteten Datencharakteristika können mit Hilfe der Methoden aus Kapitel 5 bis auf die Wölbung der Abstände der Klassenmitten (g_4) alle als wichtig für die Wahl des besten Verfahrens identifiziert werden. Es stellt sich heraus, dass sowohl Datencharakteristika, welche die Lage der Klassen zueinander beschreiben, als auch solche, die die Verteilung der Daten innerhalb der Klassen beschreiben, einen Einfluss auf die relative Güte der Verfahren haben. Dabei unterscheiden sich die Selektorstatistiken bei der Wahl des Verfahrens bei Projektion auf eine Dimension bzw. auf zwei Dimensionen.

Bei Projektion auf eine Dimension ist in den meisten Fällen das Minimale-Fehler-Kriterium am Besten. Bei der Projektion auf zwei Dimensionen wird dagegen das Optimale-Separations-Kriterium am häufigsten als das beste Kriterium für eine lineare Dimensionsreduktion zur Klassifikation identifiziert. Dieser Unterschied kann durch die Umsetzung der Verfahren erklärt werden. Die Optimierung der Projektion nach dem Minimale-Fehler-Kriterium wird aufgrund der hohen Rechenanforderungen schrittweise durchgeführt (siehe Seite 54). Anders als zum Beispiel beim Fisher Kriterium muss die gemeinsame Projektionsmatrix aber nicht mehr optimal sein. Der Rechenaufwand bei diesem Kriterium ist sehr hoch und dehalb ist das schrittweise Vorgehen bei der Optimierung dieses Kriteriums notwendig. Die beiden anderen Kriterien, das Kriterium nach Fisher und das vorhersageoptimale Projektionskriterium, sind nur dann in Datensituationen optimal, wenn in diesen die Varianz der Abstände der Klassenmitten gering aber nicht minimal ist. Geometrisch bedeutet dies, dass die Klassenmitten weder einen gleichseitigen Simplex bilden, noch dass es ein oder mehrere Klassen gibt, die im Verhältnis weit von den anderen entfernt liegen. Für das bekannte Fisher-Kriterium ist dieses Ergebnis schon in Loog (1999) bzw. Röhl *et al.* (2002) zu finden. Die bei der Entwicklung des VOP Kriteriums er-

hoffte erhöhte Stabilität gegenüber Abweichungen von der Normalverteilung konnte leider nicht erreicht werden.

Das mit Hilfe der Selektorstatistiken aus Abbildung 6.2 bzw. 6.3 entwickelte adaptive Verfahren zur Dimensionsreduktion in der Klassifikation ist sowohl bei der Projektion auf eine Dimension als auch bei der Projektion auf zwei Dimensionen über alle getesteten Datensituationen das signifikant global beste Verfahren. Da das adaptive Verfahren immer ein bekanntes Verfahren auswählt ist es in keiner Datensituation das alleinig beste. Konstruktionsbedingt existiert immer ein gleich gutes Verfahren, nämlich eben jenes Verfahren welches das adaptive Vorgehen in der Datensituation verwendet. Über die unterschiedlichen Datensituationen hinweg ist es aber insbesondere besser als die jeweils am häufigsten optimalen Kriterien MFK und OSP, so dass unter den Annahmen (A1) bis (A4) die Verwendung des adaptiven Vorgehens empfohlen werden kann, ohne das vorher die einzelnen Kriterien aufwändig evaluiert werden müssen.

Kapitel 7

Klassifikation von Konjunkturphasen

In diesen Kapitel werden die verschiedenen gebräuchlichen Klassifikationsverfahren, die im Kapitel 2 vorgestellt wurden, sowie das neue adaptive Vorgehen bei der Bestimmung von Konjunkturphasen eingesetzt. Die betrachteten Daten wurden dabei vom Rheinisch Westfälischen Institut für Wirtschaftsforschung (RWI) aus Essen ermittelt. Ziel ist es, mit Hilfe mehrerer makroökonomischer Variablen den Status der Konjunktur, die Konjunkturphase, zu bestimmen. Dabei ist die Konjunktur in vier Phasen eingeteilt:

1. Aufschwung
2. Oberer Wendepunkt
3. Abschwung
4. Unterer Wendepunkt

Die Phasen werden stets nacheinander in obiger Reihenfolge durchlaufen, wobei nach dem unteren Wendepunkt wieder ein Aufschwung folgt, so dass eine zyklische

Variable	Bezeichnung
Bruttosozialprodukt (real)	BSP91JW
Privater Verbrauch (real)	CP91JW
Anteil Staatsdefizit am Bruttosozialprodukt (%)	DEFRATE
Abhängig Erwerbstätige	EWAJW
Anteil Außenbeitrag am Bruttosozialprodukt (%)	EXIMRATE
Geldmenge M1	GM1JW
Investitionen in Ausrüstungsgüter (real)	IAU91JW
Investitionen in Bauten (real)	IB91JW
Lohnstückkosten	LSTKJW
Preisindex des Bruttosozialprodukts	PBSPJW
Preisindex des privaten Verbrauchs	PCPJW
Kurzfristiger Zinsatz (nominal)	ZINSK
Langfristiger Zinssatz (real)	ZINSLR

Tabelle 7.1: Ausgewählte Variablen zur Konjunkturphasenbestimmung

Bewegung entsteht (Heilemann & Münch, 1996). Die Dauer der einzelnen Konjunkturphasen variiert, so dass auf einen langen Aufschwung ein kurzer oberer Wendepunkt folgen kann. Dies führt dazu, dass die Häufigkeiten, in denen die einzelnen Phasen (Klassen) auftreten, nicht gleich sind. In der vorliegenden Untersuchung wurden 13 von den Ökonomen des RWI aus inhaltlichen Gründen ausgewählte, makroökonomische Variablen verwendet, um die jeweilige Konjunturphase der Bundesrepublik Deutschland zu bestimmen. Die verwendeten Variablen sind in Tabelle 7.1 angegeben. Die Daten wurden quartalsweise erhoben und erstrecken sich vom zweiten Quartal 1961 bis einschließlich des vierten Quartals 2000. In Tabelle 7.1 beziehen sich Variablen, deren Bezeichnung mit „JW “ endet, auf die Wachstumsrate der jeweiligen Variablen gegenüber dem Vorjahresquartal.

7.1 Bestimmung der Fehlerrate bei Konjunkturdaten

Bei der Bestimmung von Konjunkturphasen ist man interessiert an einer möglichst genauen Bestimmung der unmittelbar folgenden konjunkturellen Phasen, beispielsweise des nächsten Jahres. Daher ist für Ökonomen an dieser Stelle die so genannte Ex-Post-Ante-Fehlerrate interessant, um die Performanz verschiedener Klassifikationsverfahren zu vergleichen. Die Ex-Post-Ante-Fehlerrate misst retrospektiv die Vorhersagegüte der Verfahren:

$$epa(t; pre) = \frac{\sum_{i=t}^{\min(t+pre,T)} I_{\{k_i \neq \hat{k}_i^t\}}}{\min(pre, T-t)}, \tag{7.1}$$

wobei

- $k_i, \hat{k}_i$ die wahre bzw. die geschätzte Klasse (Konjunkturphase) der Beobachtung i ist,
- pre die Anzahl der Folgebeobachtungen, für die die Klasse bestimmt werden soll,
- t die letzte Beobachtung, mit der das Klassifikationsmodell geschätzt (gelernt) wird,
- T die letzte Beobachtung im Datensatz

ist. Mit dieser retrospektiven Vorhersagegüte zum Zeitpunkt t kann dann ein Schätzer für die Fehlklassifikationsrate für die nächsten pre Beobachtungen entwickelt werden

$$\widehat{err}_{t_0} = \sum_{i=0}^{i=\lfloor \frac{T-t_0}{pre} \rfloor} w(i \cdot pre + t_0) epa(i \cdot pre + t_0; pre). \tag{7.2}$$

Dabei ist t_0 die erste Beobachtung, ab der die Klassifikationsregel berechnet wird. Mit Hilfe des Terms $i \cdot pre + t_0$ in (7.2) wird erreicht, dass jede Beobachtung nur

einmal zum Testen der Regel verwendet wird. Es werden zum Zeitpunkt t immer die nächsten *pre* Beobachtungen auf Grund aller vorhergehenden Beobachtungen klassifiziert. Anschließend werden die Testbeobachtungen zum Trainingsdatensatz hinzugefügt und *pre* neue Beobachtungen werden mit Hilfe des vergrößerten Testdatensatz klassifiziert. In der vorliegenden Arbeit bedeuten (7.1) und (7.2), dass beispielsweise alle Daten bis einschließlich 1979 zum Lernen und Schätzen des Klassifikators verwendet werden und dieser auf den Daten von 1980 getestet wird. Anschließend werden die Daten bis einschließlich 1980 zum Lernen und Schätzen sowie die von 1981 zum Testen verwendet.

Es sind verschiedene Gewichte $w(t)$ in (7.2) möglich

1. Konstant über die Zeit: $w(t)_K = \frac{pre}{T-t_0}$,
2. Ansteigendes Gewicht im Laufe der Zeit mit höherer Anzahl an Trainingsbeobachtungen: $w(t)_T = \frac{t}{\sum_{t=t_0}^{T-1} t}$.

Da $w(\cdot)_T$ die Fehlerraten in jüngster Zeit höher gewichtet, entspricht es mehr dem Anforderungscharakter in der hier betrachteten Konjunkturphasenbestimmung.

7.2 Ergebnisse bei der Bestimmung der Konjunkturphase

Als problematisch bei der Anwendung der betrachteten Verfahren erwies sich die ungleiche Verteilung der Häufigkeiten der einzelnen Konjunkturphasen. Da bei der Optimierung des Minimaler-Fehler-Kriteriums Stichproben gezogen werden und in jeder dieser Stichproben alle Phasen (Klassen) zur Bestimmung der innerhalb der Optimierung notwendigen Matrizen auftreten müssen, erfordert die MFK Methode hier eine relativ große Stichprobe. Ein Schätzen der optimalen Projektion war hier erst möglich, nachdem fast drei Konjunkturzyklen durchlaufen waren (Ende 1973).

Daher konnte die Ex-Post-Ante-Fehlerrate (7.2) erstmalig mit allen Verfahren für die vier Quartale des Jahres 1974 bestimmt werden.

Der Selektorstatistik, die der Verfahrensauswahl bei den Konjunkturdaten zugrunde liegt, ist in Abbildung 7.1 angegeben. Bei der Projektion auf eine Dimension wählt das adaptive Verfahren aufgrund der Datensituation bis einschließlich $t = 1993$ das MFK Kriterium zur Vorhersage der Konjunkturphasen aus. Da sich die Datensituation mit steigender Anzahl an Trainingsbeobachtungen ändert, wird ab 1994 das OSP Kriterium zur Vorhersage verwendet. Die entscheidende Änderung zur Vorhersage ab 1994 ist die Verringerung der Varianz der Klassenabstände (g_2).

Die Selektorstatistik bei der Projektion auf zwei Dimensionen ist in Abbildung 7.2 dargestellt. Bei der Projektion auf zwei Dimensionen wird bis 1992 das Vorhersageoptimale Projektionskriterium verwendet. Ab 1993 wird das OSP Kriterium aufgrund der Adaption zur Konjunkturphasenbestimmung herangezogen. Für den Verfahrenswechsel ist hier entscheidend, dass sich der mittlere Abstand der Klassenmitten (g_1) erhöht, so dass ab der Wurzel des Baumes (siehe Abbildung 7.2) ein anderer Weg eingeschlagen wird. Die Daten für die Selektorstatistik sind in Tabelle B.1 im Anhang angegeben. Die Fehlerraten (7.2) bei der Projektion auf eine Dimension sind in Tabelle 7.2 zu finden. Das Ergebnis für das adaptive Verfah-

Gewicht	adapt1	fisher	mfk	osp	vop
w_K	0.306	0.278	0.306	0.287	0.296
w_T	0.285	0.243	0.281	0.261	0.249

Tabelle 7.2: Fehlklassifikationsraten bei Projektion auf eine Dimension

ren bei Projektion auf eine Dimension zur Bestimmung von Konjunkturphasen ist unbefriedigend. Ursache hierfür ist die schlechte Performanz des MFK Kriteriums, welches bis 1993 für das adaptive Vorgehen verwendet wird. Wie oben erwähnt, hat das MFK-Verfahren Probleme bei ungleichen Klassengrößen, allerdings wurde bei der Entwicklung der Selektorstatistik (siehe Kapitel 6) explizit die Annahme (A1) von gleichen a priori Wahrscheinlichkeiten verwendet (siehe Seite 61 und Seite 88).

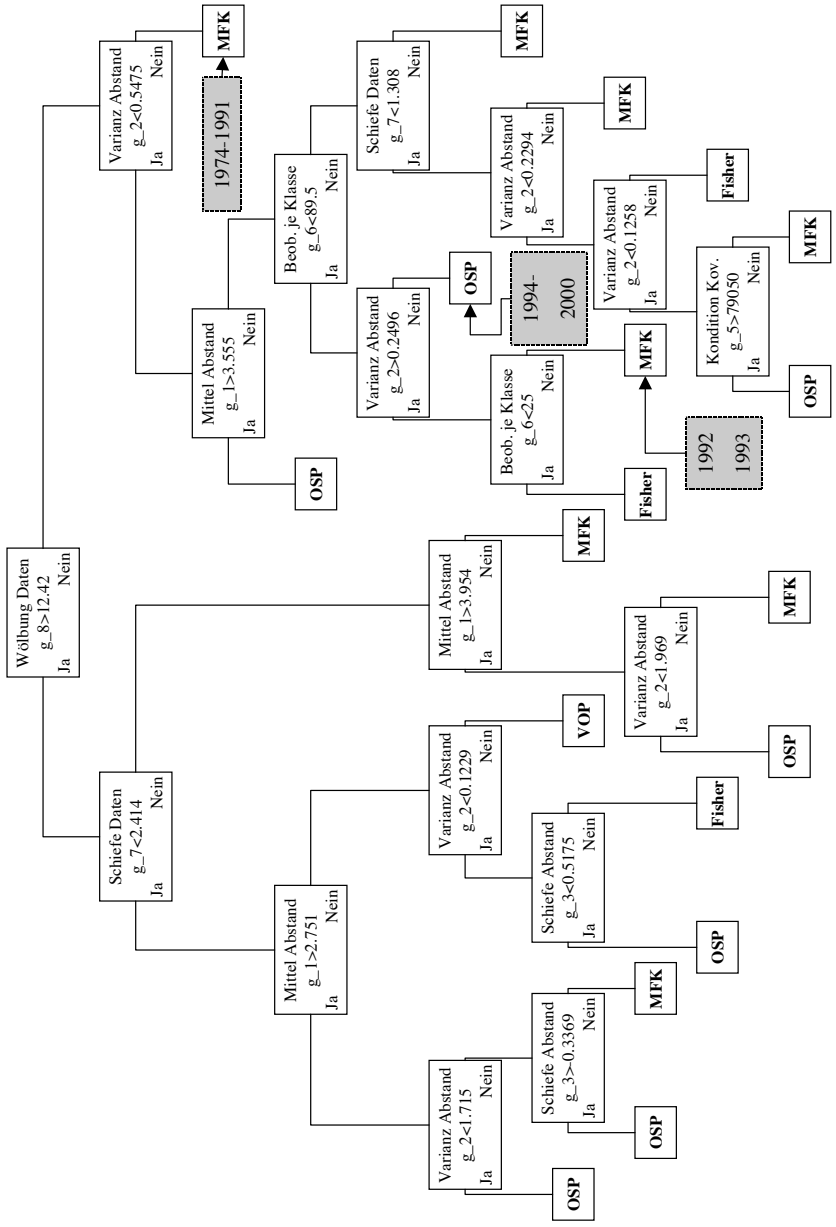

Abbildung 7.1: Ergebnisse der Selektorstatistik auf die Konjunkturdaten bei Projektion auf eine Dimension

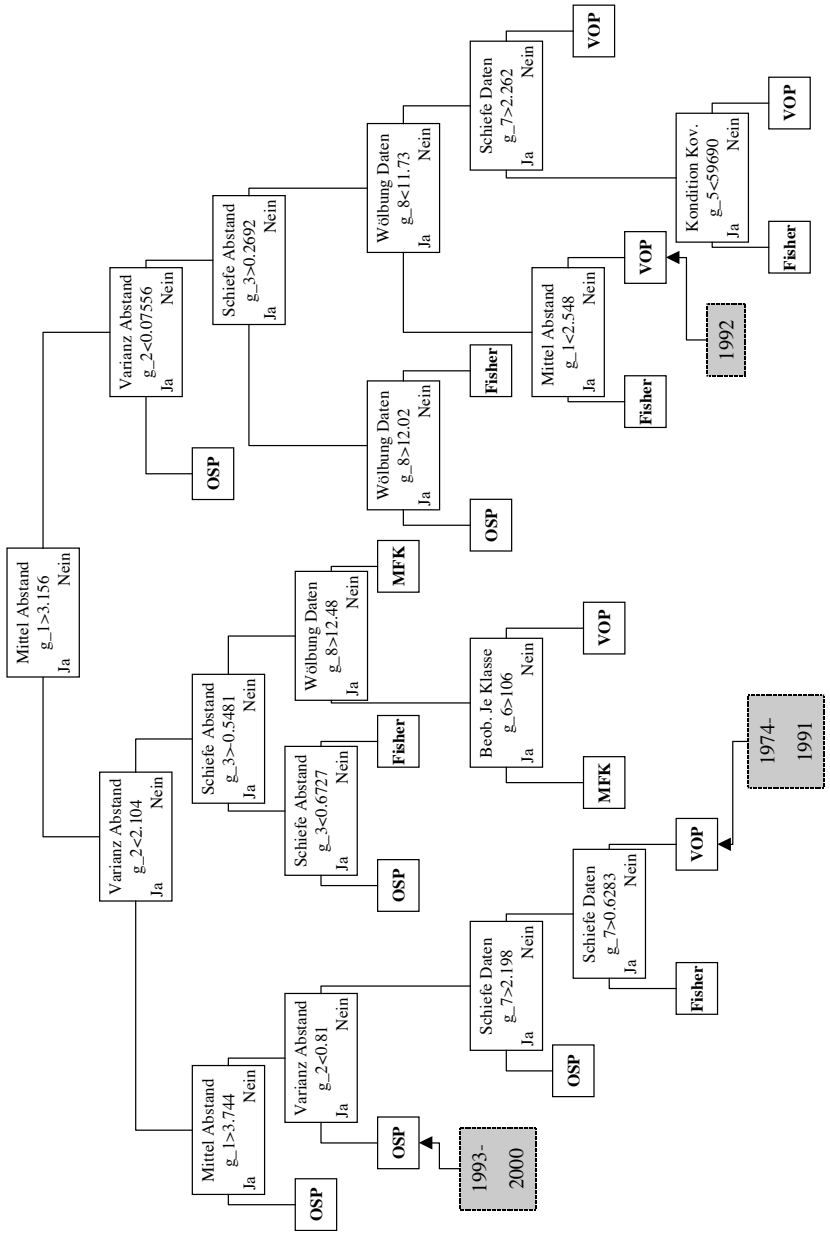

Abbildung 7.2: Ergebnisse der Selektorstatistik auf die Konjunkturdaten bei Projektion auf zwei Dimensionen

Diese Annahme ist im vorliegenden Datensatz nicht erfüllt, da die Klassen unterschiedlich stark besetzt sind. So gibt es 80 Quartale mit einem Aufschwung, aber nur 18 mit der Phase des oberen Wendepunktes. Des weiteren gibt es 38 Quartale des Abschwungs und 23 Quartale mit einen unteren Wendepunkt. Problematisch ist weiterhin die Berechnung der Klassengröße (g_6) als mittlere Klassengröße. Dies führt dazu, dass für die Jahre 1992 und 1993 das MFK Kriterium verwendet wird ($g_6 \geq 25$, vergleiche Abbildung 7.1), obwohl die beiden Wendepunktphasen weniger als 25 Beobachtungen enthalten. Würde das adaptive Verfahren anstelle des MFK das alternative Fisher-Kriterium verwenden, würde sich das Ergebnis ändern. Dann würde das adaptive Vorgehen insgesamt die gleiche Fehlerrate erreichen wie das Fisher-Kriterium (siehe Tabelle B.2 im Anhang), das global beste Verfahren bei der Projektion auf eine Dimension bei den vorliegenden Daten. Dabei würde das adaptive Vorgehen nur für zwei Jahre das Fisher-Kriterium verwenden. Daher ist die schlechte Performanz des adaptiven Vorgehens auf die spezifischen Eigenschaften in diesen zwei Jahren zurückzuführen.

Die Ergebnisse für die Fehlerraten (7.2) bei Projektion auf zwei Dimensionen sind in Tabelle 7.3 angegeben. In diesem Fall ist das adaptive Vorgehen das Beste der getes-

Gewicht	adapt2	fisher	mfk	osp	vop
w_K	0.111	0.185	0.269	0.120	0.130
w_T	0.088	0.182	0.227	0.093	0.115

Tabelle 7.3: Fehlklassifikationsraten bei Projektion auf zwei Dimensionen

teten Verfahren. Auch hier zeigt sich, dass das MFK Kriterium Probleme mit dem vorliegenden Datensatz hat. Im Vergleich zu Tabelle 7.2 verbessern sich die Fehlerraten deutlich. Einzig die des MFK Kriterium weisen nur eine leichte Verbesserung auf. Das adaptive Vorgehen ist hier nahe am Optimum: Würde man retrospektiv jeweils das Verfahren auswählen, welches die geringste Fehlklassifikation ermöglicht, so würde nur eine Beobachtung weniger fehlklassifiziert werden.

Ein Vorteil der Ex-Post-Ante-Fehlerrate (7.1) ist die Darstellung der Fehlerraten

in ihrer zeitlichen Struktur womit zeitliche Phänomene bei der Klassifikation beobachten werden können. Die Zeitreihe der Ex-Post-Ante-Fehlerrate bei der Konjunktuphasenklassifikation mit Hilfe des adaptiven Vorgehens zeigt Abbildung 7.3. Die Ergebnisse der Ex-Post-Ante Fehlerrate für alle Verfahren sind in den Tabellen B.2 und B.3 im Anhang zu finden. Auf Abbildung 7.3 ist gut zu erkennen, dass das

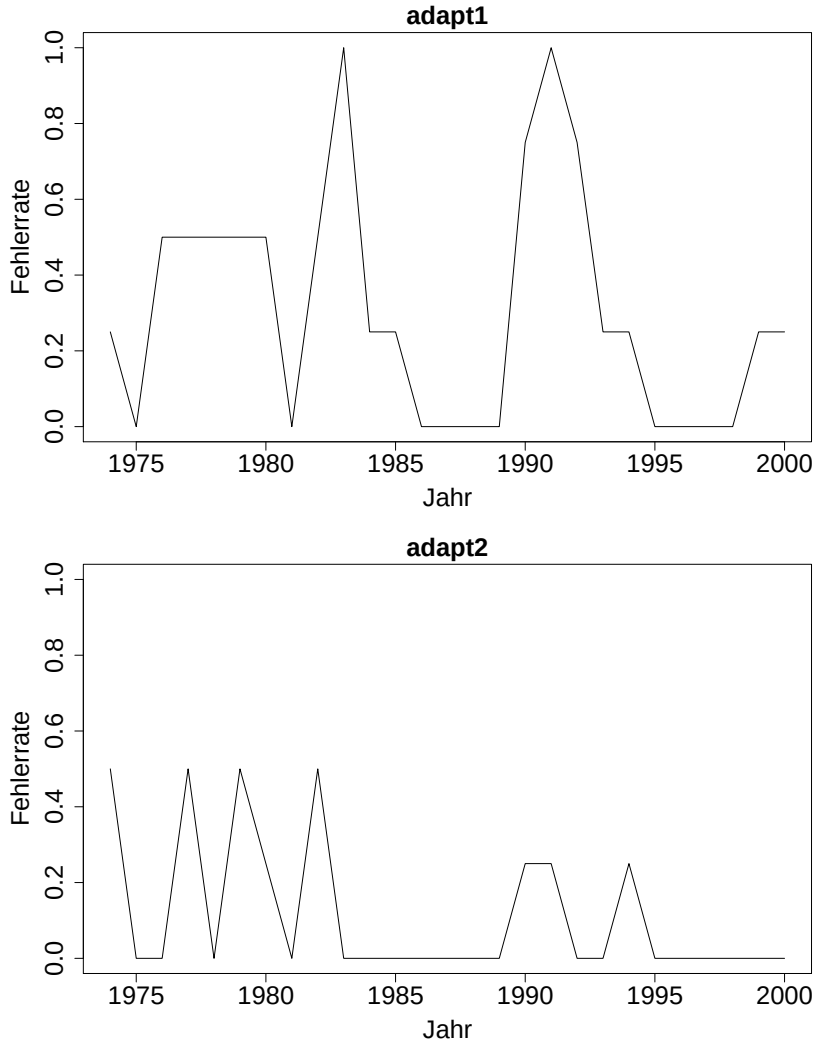

Abbildung 7.3: Zeitreihe der Ex-Post-Ante Fehlerrate (7.2)

adaptive Vorgehen – wie auch die einzelnen Verfahren – sehr große Schwierigkeiten mit der korrekten Klassifikation zum Zeitpunkt der Wiedervereinigung Deutschlands (1990) und der hohen Änderungsraten im Jahr nach der Wiedervereinigung

1991 haben. Da die meisten verwendeten Variablen sich auf Änderungen gegenüber dem Vorjahr beziehen (siehe Tabelle 7.1), sind viele Werte in den vier Quartalen von 1991 ein Vielfaches der sonstigen Änderungsraten. Beispielsweise beträgt die Änderungsrate der abhängig Erwerbstätigen (EWAJW) im zweiten Quartal 1991 fast 38%, während sie im selben Quartal von 1990 3% betrug. Weiterhin gibt es Probleme bei der Klassifikation, die auf die Nachwirkung der ersten (1973/74) sowie der zweiten (1979/1980) Ölkrise zurückzuführen sind. Diese Ereignisse führten sicherlich zu einer Änderung innerhalb der Konjunktur der Bundesrepublik. Statische statistische Klassifikationsverfahren wie die hier verwendeten haben in Fällen, in denen sich die zugrundeliegende Verteilung $\mathcal{F}$ ändert, Schwierigkeiten. Eine Abhilfe könnten hier die neu entwickelten lokalen Modelle liefern, die sich auch auf die lineare Diskriminanzanalyse anwenden lassen (Czogiel *et al.*, 2006).

Ein Vorteil der Dimensionsreduktion auf zwei Dimensionen ist die dann mögliche Visualisierung der Daten. In Abbildung 7.4 ist das Ergebnis der Projektion nach dem für den ganzen Datensatz besten Vorgehen (d.h. mit Hilfe des Optimale Separations Kriteriums) mitsamt der Partition des Bildraumes in die einzelnen Konjunkturphasen angegeben. Man kann gut erkennen, dass die Wendepunktphasen oberer Wendepunkt (2) sowie unterer Wendepunkt (4) keine gemeinsame Klassengrenze haben. Dies ist nachzuvollziehen, da sie sowohl zeitlich und auch im Konjunkturzyklus am weitesten voneinander entfernt liegen. Weiterhin ist auch in dieser Darstellung zu erkennen, dass die Aufschwungsphase (1) am stärksten besetzt ist. Im Originalraum liegen die Mittelwertsvektoren des oberen Wendepunktes (2) und die des Abschwungs (3) sowohl nach der euklidischen als auch nach der Metrik nach Mahalanobis am nächsten beieinander. Daher sind diese Phasen am schwersten zu separieren. Bei der Projektion nach dem OSP Kriterium ist die Fehlklassifikationswahrscheinlichkeit zwischen dem oberen Wendepunkt und dem Abschwung am geringsten von den Verfahren: 0.11 gegenüber beispielsweise 0.17 bei einer Projektion nach dem Fisher-Kriterium.

Mit Hilfe des adaptiven Verfahrens konnte die insgesamt geringste Fehlerrate bei der Klassifikation von Konjunkturphasen innerhalb der vorgestellten Verfahren erreicht

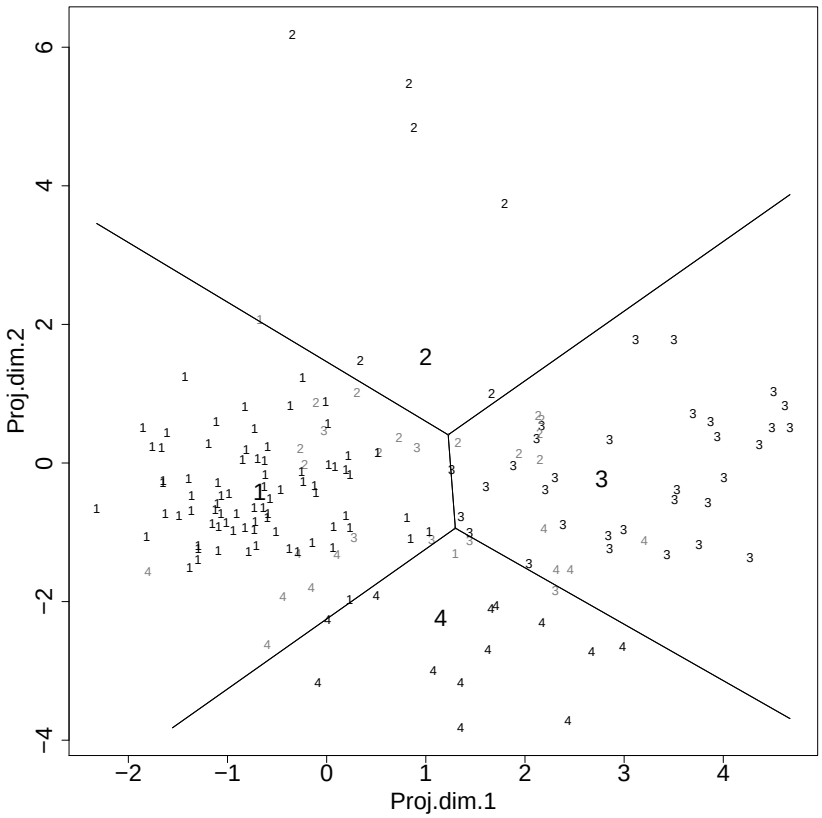

Abbildung 7.4: Partitionsplot der Konjunktudaten nach Projektion mit Hilfe des OSP Kriteriums

werden. Insbesondere ist das adaptive Vorgehen im Fall der Projektion auf zwei Dimensionen besser als die ausschließliche Verwendung eines der zur Auswahl verwendeten Verfahren. Bei der Projektion auf eine Dimension erweist sich die Beschreibung der Stichprobengröße mit nur einer Kennzahl als problematisch, da das MFK Kriterium bei sehr gering besetzten Klassen systematisch kein oder ein nicht optimales Ergebnis produziert. Die dem adaptiven Vorgehen zugrundeliegende Selektorstatistik deutet darauf hin, dass sich neben der Stichprobengröße weitere Datencharakteristika im Laufe der Zeit ändern (siehe Abbildungen 7.1 und 7.2). Der Verlauf der Ex-Post-Ante-Fehlerrate liefert weitere Hinweise auf eine Strukturveränderung der einzelnen Konjunkturphasen, die sich mit ökonomischen Besonderheiten erklären lassen. Neben der geringen Fehlerrate und der damit erhöhten Sicherheit bei der

Anwendung der Klassifikation auf neue Beobachtungen mit noch unbekannter Konjunkturphase, konnten mit Hilfe des adaptiven Vorgehens die schon vorhandenen Beobachtungen der Konjunktur in den Abbildungen besser verdeutlicht werden.

Kapitel 8

Zusammenfassung und Ausblick

Die Möglichkeit eines Vergleich von Dimensionsreduktionsverfahren in der Klassifikation auf Basis einer Selektorstatistik wurde vorgestellt. Darauf aufbauend konnte das jeweils beste Verfahren gewählt werden. Dabei wurde die Selektorstatistik mit Hilfe eines Versuchsplanes über den möglichen Datenraum, der Basis eines Klassifikationsproblems ist, gewonnen. Aufgrund der entwickelten Beschreibung der Datensituationen konnte erstmals – unter bestimmten restriktiven Annahmen – die Performanz der Verfahren systematisch über den möglichen Datenraum evaluiert werden.

Dazu wurden verschiedene Kriterien zur Dimensionsreduktion und daraus hervorgehende Klassifikationsverfahren dargestellt und entwickelt. Im Anschluss wurden die Optimierung über diese Kriterien und die dafür notwendigen Parameterschätzungen besprochen. In der Literatur wird nicht immer zwischen den theoretischen, auf die zugrundeliegende Verteilungen zurückgehenden Untersuchungen, und den Fall einer gegebenen Stichprobe unterschieden. Allerdings werden in der Literatur häufig keine systematischen Untersuchungen über die Performanz der Verfahren durchgeführt. Statt dessen werden die Verfahren anhand von einiger Beispiele (simuliert oder real) getestet, ohne dass über das verwendete Beispiel hinaus Aussagen getroffen werden können. Deshalb wurde in dieser Arbeit der Begriff der Datensituation

verwendet. Die Struktur der vorliegenden Daten soll damit systematisch und vergleichbar erfasst werden, damit weitergehende Analysen der Verfahren ermöglicht werden können. Um zu ermitteln, welches der vorhandenen Verfahren und Kriterien bei einer vorliegenden Datensituation das beste ist, muss der Raum der möglichen Datensituationen erfasst und beschrieben werden. Auf diesem Gebiet liegen bisher nur sehr wenige Arbeiten vor. Insbesondere die Beschreibung der Lage der Klassen zueinander mittels vergleichbarer Kennzahlen stellte eine große Hürde bei der Erfassung der Datensituation dar, da diverse Invarianzen des Problems berücksichtigt werden müssen. Nachdem aber eine Reihe von Datencharakteristika, die die Datensituation abbilden, entwickelt wurden, konnte der Raum der Datensituationen erfasst werden und es konnten unterschiedliche Datensituationen innerhalb dieses Raumes mit Hilfe von statistischer Versuchsplanung generiert werden. Durch das Herleiten eines Wahrscheinlichkeitsmodells für die Güte der Klassifikation nach einer Dimensionsreduktion wurde erreicht, dass statistisch fundierte Aussagen über die relative Güte der einzelnen Verfahren getroffen werden können. Zusammen mit der systematischen Generierung von Daten gemäß unterschiedlicher Datensituationen konnte ein adaptives Vorgehen mit Hilfe einer Klassifikation des jeweils signifikant besten Verfahrens aufgrund der Datensituation entwickelt werden. Somit kann ohne aufwändiges Ausprobieren anhand einer vorgeschalteten Bestimmung der Datensituation mit Hilfe der Selektorstatistik das beste Verfahren bestimmt und angewandt werden. Dabei erwiesen sich fast alle definierten Datencharakteristika als relevant für die Auswahl des besten Verfahrens.

Die Ergebnisse in der Anwendung eines solchen adaptiven Vorgehens zeigen im Fall der Klassifikation von vier Klassen eine signifikante Überlegenheit dieses adaptiven Vorgehens über alle getesteten Datensituationen hinweg. Konstruktionsbedingt ist das adaptive „Verfahren“ nie das alleinig beste, so dass es in jeder Datensituation ein mindestens gleich gutes Verfahren gibt. Im Allgemeinen ist dieses Vorgehen aber angebracht, da gezeigt werden konnte, dass die Fehlklassifikationsrate bei ausschließlich adaptivem Vorgehen signifikant geringer ist, als wenn ausschließlich eines der anderen Verfahren verwendet wird.

Die Anwendung des adaptiven Vorgehens auf die quartalsweise Zuordnung von vier Konjunkturphasen aufgrund von makroökonomischen Merkmalen der Bundesrepublik Deutschland führt zu zweierlei Erkenntnissen: Einerseits kann auch bei realen Daten das beste Ergebnis mit Hilfe der adaptiven Auswahl des Verfahrens erzielt werden, andererseits fehlen Datencharakteristika, die unterschiedliche Häufigkeiten der Klassen beschreiben und bei der Kriterienselektion berücksichtigen. Bei der Entwicklung des adaptiven Vorgehens wurde die Annahme von gleichen a priori Wahrscheinlichkeiten verwendet. Die Verletzung dieser Annahme führte bei Auswahl des Verfahrens aufgrund der Selektorstatistik bei der Projektion auf eine Dimension zu suboptimalen Ergebnissen. Speziell das Minimale-Fehler-Kriterium (MFK) erwies sich als anfällig bei einzelnen gering besetzten Klassen. Im Falle der Projektion auf zwei Dimensionen wurde das anfällige MFK Verfahren nicht selektiert. Somit konnte hier das insgesamt beste Ergebnis bezogen auf die Ex-Post-Ante-Fehlerrate erreicht werden. Hier werden von den vier Verfahren zur Auswahl nur die im Zuge dieser Arbeit neu entwickelten Verfahren von der Selektorstatistik verwendet: Das Optimale-Separations-Kriterium (OSP) sowie die vorhersageoptimale Projektion. Das OSP Kriterium ist dem MFK Kriterium in Bezug auf die notwendige Rechenzeit weit überlegen und wird auch ab dem Jahr 1994 zur Klassifikation von Konjunkturphasen bei der Projektion auf nur eine Dimension vom adaptiven Verfahren ausgewählt. Das VOP Kriterium konnte die in dieses Verfahren gesetzten Hoffnungen nur zum Teil erfüllen, allerdings ist es sehr flexibel und ermöglicht die Verwendung weiterer Aspekte innerhalb einer Klassifikation. So kann beispielsweise explizit die zeitliche Struktur von Daten berücksichtigt werden oder eine Variablenselektion im Ausgangsraum unterstützt werden (Luebke & Weihs, 2005b).
Die Betrachtung der Zeitreihe der neu entwickelten Ex-Post-Ante-Fehlerrate für zeitliche Daten zeigte die Auswirkung von Ereignissen wie der Wiedervereinigung auf das Ergebnis einer statistischen Klassifikation. Auch die Visualisierung der Projektion auf zwei Dimensionen konnte zum besseren Verstehen der Daten beitragen, indem es zeigt, dass in der Projektion die Wendepunktphasen weiter voneinander entfernt sind.

Die Ergebnisse der adaptiven linearen Dimensionsreduktion auf den Konjunkturdaten Deutschlands führen zu den ersten Weiterentwicklungen des adaptiven Vorgehens, nämlich der Berücksichtigung der einzelnen Klassenhäufigkeiten in der Datensituation. Weiterhin wird in der definierten Datensituation nicht betrachtet, ob innerhalb der Klassen unterschiedliche Kovarianzen auftreten, da in dieser Arbeit nur lineare Klassifikatoren verwendet wurden. In diesen Zusammenhang könnte auch das OSP Kriterium für die quadratische Diskriminanzanalyse erweitert werden. Ein weiterer Aspekt zukünftiger Arbeit kann die direkte adaptive Auswahl unterschiedlicher Klassifikationsverfahren (ohne Dimensionsreduktion) wie beispielsweise Entscheidungsbäume, Stützvektormethode oder Neuronale-Netze sein. Diese Klassifikationsverfahren sind in sich sehr unterschiedlich aufgebaut, so dass hier interessante Ergebnisse erwartet werden können. Dazu müssen dann natürlich zum Teil andere Datencharakteristika herangezogen werden. Für solche Analysen kann aber der in der vorliegenden Arbeit entwickelte theoretische Bezugsrahmen sowie das prinzipielle methodische Vorgehen herangezogen werden.

Anhang A

Simulated Annealing

Für die Kriterien

- Minimaler-Fehler
- Optimale-Separations-Projektion
- Vorhersageoptimale-Projektion

sind die optimalen Projektionsvektoren bzw. Matrizen unbekannt und es existieren keine analytischen Lösungen. In solchen Fällen wird probiert die optimale Lösung mittels Suchalgorithmen zu approximieren. Bei einem solchen Vorgehen muss besonders auf lokale Optima geachtet werden, da viele numerischen Suchalgorithmen nicht robust gegen lokale Optima sind. Stochastitsche Suchalgorithmen wie Genetische-Algorithmen oder Simulated-Annealing können je nach Konfiguration, lokale Optima überspringen und haben deshalb eine erhöhte Chance das globale Optimum zu finden. In dieser Arbeit wird Simulated Annealing (vergleiche Salamon *et al.* (2002)) in einer Version von Press *et al.* (1992) verwendet, die auf dem Nelder-Mead-Suchverfahren aufbaut.
Die jeweiligen Zielfunktionen $J(\cdot)$ der Kriterien sollen minimiert werden: geschätzte Fehlerrate (MFK), geschätzte obere Fehlerschranke (OSP) bzw. mittlerer quadratischer Vorhersagefehler (VOP). Daher werden die Optimierungsmethoden für den

Fall der Funktionsminimierung beschrieben. Sollte die Funktion über Matrizen optimiert werden, so müssen diese vektorisiert werden.

A.1 Nelder-Mead

Das Verfahren von Nelder-Mead (Nelder & Mead, 1965) ist eine Erweiterung des Simplex-Verfahrens und damit ein Verfahren zur Optimierung einer Zielfunktion, das ohne die Ableitung der zu optimierenden Funktion auskommt.
Das Verfahren von Nelder-Mead optimiert den Funktionswert einer Funktion von k Variablen iterativ mittels eines $k+1$ dimensionalen, flexiblen Simplex. Jeder Eckpunkt des Simplex besteht zur Iteration t aus einem k dimensionalen Vektor $w_t^j, j = 1, \cdots, k+1$. Sei $w_t^{(min)}$ der Punkt des Simplex mit dem kleinsten Funktionswert und $w_t^{(max)}$ der Punkt mit dem größten Funktionswert. Der Schwerpunkt des Simplex ohne den Punkt $w_t^{(max)}$ ist dann $w_t^s = \frac{1}{k}(\sum_{j=1}^{k+1} w_t^j - w_t^{(max)})$.
Zu jeder Iteration t finden dann folgende Schritte statt (vergleiche auch Press *et al.* (1992), Seite 408ff.):

1. *Reflexion*
   $w_t^{(max)}$ wird am Zentroid w_t^s gespiegelt. $w_t^{(test)} = 2w_t^{(s)} - w_t^{(max)}$.

2. *Expansion*
   Wenn der neue Punkt $w_t^{(test)}$ besser als der bisher beste Punkt $w_t^{(min)}$ ist, dann gehe weiter in Richtung von $w_t^{(test)}$. Dann ist $w_t^{(test2)} = 2w_t^{(test)} - w_t^{(s)}$. Falls $J(w_t^{(test2)}) < J(w_t^{(min)})$ ist, dann wird $w_t^{(max)}$ durch $w_t^{(test2)}$ ersetzt, andernfalls durch $w_t^{(test)}$. Weiter geht es mit Schritt 1 und $t = t + 1$.

3. *Kontraktion*
   Wenn $w_t^{(test)}$ besser als der schlechteste Punkt, aber schlechter als die anderen k Punkte ist, soll der Vektor $(w_t^{(max)} - w_t^{(min)})$ schrumpfen. Der Punkt $w_t^{(max)}$ wird durch den Punkt $w_t^{(neu)} = \frac{1}{2}w_t^s + \frac{1}{2}w_t^{(max)}$ ersetzt. Es geht mit Schritt 1 und $t = t + 1$ weiter.

4. *Reduktion*
   Sollte $w_t^{(test)}$ schlechter als alle bisherigen Punkte im Simplex sein, werden alle Punkte des Simplex in Richtung des besten Punktes verschoben. $w_t^j = w_t^{(min)} + \frac{1}{2}(w_t^j - w_t^{(min)}), j = 1, \cdots, k+1$. Es wird zu Schritt 1 mit $t = t + 1$ gegangen.

Die Konvergenz des Verfahrens kann zum Beispiel dadurch überprüft werden, indem festgestellt wird, wie weit die Eckpunkte des Simplex noch vom Schwerpunkt entfernt sind. Die Schritte Spiegeln, Expansion, Kontraktion und Reduktion beschreiben den Übergang von einem Simplex zum Nächsten, geben also eine Übergangsvorschrift an.

A.2 Simulated Annealing

Beim Simulated-Annealing handelt es sich um ein rechenintensives Optimierungsverfahren, dass sich an der Thermodynamik orientiert (Bohachevsky *et al.*, 1986). Dabei wird ausgenutzt, dass sich bei hoher Temperatur Moleküle mehr bewegen können als bei niedriger Temperatur. Wenn die Temperatur langsam sinkt und sich die Moleküle weniger bewegen können, wird ein optimales Gleichgewicht erreicht.
Die hier implementierte Version basiert auf der Version von Press *et al.* (1992) (siehe dort Seite 451ff.), die eine Erweiterung des Suchverfahrens von Nelder-Mead darstellt.
Der Algorithmus sieht wie folgt aus:

1. Erzeuge einen Startsimplex.
2. Passe die Vektoren bzw. Matrizen der Simplexpunkte gegebenenfalls einer Nebenbedingung an.
3. Berechne für jeden Punkt im Simplex den Wert des interessierenden Kriteriums J.

4. Addiere zu den Funktionswerten an den aktuellen Simplexpunkten eine Zufallszahl:

$$J_{temp}(w) = J(w) + t|log(u)|.$$

Dabei ist u eine auf (0,1) gleichverteilte Zufallsvariable.

5. Erzeuge einen neuen Punkt (w_{neu}) mittels der Nelder-Mead Übergangsfunktion (siehe oben). Zur Bestimmung des besten, zweitschlechtesten und schlechtesten Punktes im Simplex verwende die Funktionswerte von $J_{temp}(\cdot)$.

6. Passe den neuen Vektor bzw. die neue Matrix w_{neu} gegebenenfalls einer Nebenbedingung an.

7. Akzeptiere den neuen Punkt nach Nelder-Mead, wobei der Funktionswert des neuen Punktes

$$J_{temp}(w_{neu}) = J(w_{neu}) - t|log(u)|$$

ist.

8. Wiederhole die Schritte 2-4 l Mal. Verringere dann die Temperatur von t auf $c \cdot t$, wobei $0 \leq c < 1$ ist.

9. Wiederhole die Schritte 2-5 m Mal.

Der entscheidende Unterschied zum Verfahren von Nelder-Mead liegt darin, dass – mit einer bestimmten Wahrscheinlichkeit – neue Punkte auch dann akzeptiert werden, wenn sie schlechter als die bisherigen Punkte des Simplex sind. Ein besserer Punkt wird immer akzeptiert. Dieses kann geschehen, weil die Funktionswerte der alten Punkte im Simplex in Schritt 4 erhöht werden und der Funktionswert des neuen Punktes in Schritt 7 verringert wird. Damit geht es in diesem Verfahren nicht nur „bergab“ sondern mit einer bestimmbaren Wahrscheinlichkeit auch „bergauf“. Daher ist die Möglichkeit gegeben, sich aus einem lokalen Minima zu befreien. Allerdings verringert sich im Laufe der Zeit die Wahrscheinlichkeit, dass ein schlechterer Punkt akzeptiert wird, so dass das Verfahren nach einer gewissen Zeit gegen das nächste Minimum konvergiert.
Das Ergebnis des Simulated Annealing hängt von verschiedenen Faktoren ab.

- dem Startsimplex
- der Starttemperatur t_0
- dem Kühlparameter c
- den Iterationen l und m
- dem Zufall (U)

Als Ausgangsvektor des Startsimplexes wird in dieser Arbeit die optimale Projektion nach Fisher verwendet, da das Verfahren nach Fisher der bisherige Standard ist, und somit einen guten Ausgangspunkt liefern sollte. Die Starttemperatur wird in der Regel hoch gewählt, damit das Verfahren zu Beginn noch viel Freiraum hat. Hier wurde die maximal mögliche Fehlerrate (1) oder Abweichung ($K-1$) gewählt. Für den Kühlparameter und die Anzahl der Iterationen gibt es keine heuristischen oder theoretischen Lösungen. Aufgrund allgemein guter Erfahrungen wurde $c = 0.8$, $l = 100$ und $t = 50$ gewählt. Bei diesen Parameterkonstellationen konnte sehr häufig beobachtet werden, dass der Algorithmus konvergierte.

Anhang B

Ergebnisse Konjunkturphasenbestimmung

In diesen Anhang sind die Ergebnisse für die Bestimmung der Konjunkturphasen angegeben (siehe Kapitel 7). Tabelle B.1 gibt die Ergebnisse der Selektorstatistik aufgrund der das optimale Verfahren ausgewählt wird für das jeweilige Vorhersagejahr an. Die Tabellen B.2 und B.3 geben die Ex-Post-Ante-Fehlerraten (siehe Gleichung (7.1)) bei einer Vorhersage von $pre = 4$ Quartalen an.

Vorhersagejahr	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8
1974	3.35	1.12	−0.13	1.61	2002.91	12.50	0.52	2.79
1975	3.27	1.11	−0.01	1.65	1887.38	13.50	0.56	2.68
1976	3.29	1.17	−0.07	1.56	1654.04	14.50	0.42	2.63
1977	3.29	1.05	−0.16	1.54	1477.48	15.50	0.44	2.61
1978	3.18	1.06	−0.18	1.45	1295.49	16.50	0.32	2.36
1979	3.24	1.04	−0.24	1.47	1249.35	17.50	0.40	2.57
1980	3.25	1.19	−0.23	1.38	1219.93	18.50	0.47	2.57
1981	3.34	1.12	−0.23	1.39	1154.40	19.50	0.53	2.75
1982	3.41	1.20	−0.29	1.44	1264.06	20.50	0.46	2.69
1983	3.30	1.03	−0.22	1.53	1377.45	21.50	0.50	2.78
1984	3.28	0.95	−0.29	1.53	1378.92	22.50	0.50	2.82
1985	3.25	0.84	−0.33	1.58	1205.43	23.50	0.50	2.88
1986	3.33	0.86	−0.31	1.55	1233.88	24.50	0.51	2.92
1987	3.39	0.85	−0.28	1.54	1149.53	25.50	0.47	3.00
1988	3.44	0.88	−0.25	1.52	1120.08	26.50	0.38	3.01
1989	3.48	0.89	−0.25	1.51	1120.66	27.50	0.37	2.96
1990	3.53	0.92	−0.26	1.50	1064.27	28.50	0.36	3.01
1991	3.54	0.93	−0.19	1.55	967.32	29.50	0.36	3.06
1992	3.14	0.26	−0.14	1.73	1455.52	30.50	0.49	2.90
1993	3.22	0.25	−0.09	1.66	1478.34	31.50	0.51	2.93
1994	3.26	0.23	−0.10	1.60	1425.60	32.50	0.58	2.92
1995	3.26	0.22	−0.04	1.52	1359.26	33.50	0.55	2.92
1996	3.25	0.21	0.06	1.46	1314.63	34.50	0.46	2.91
1997	3.29	0.21	0.09	1.53	1316.44	35.50	0.46	2.92
1998	3.33	0.21	0.08	1.57	1242.16	36.50	0.47	2.92
1999	3.34	0.19	0.08	1.61	1113.23	37.50	0.46	2.90
2000	3.33	0.17	0.17	1.87	967.81	38.50	0.47	2.85

Tabelle B.1: Ergebnisse der Selektorstatistik

Vorhersagejahr	adapt1	fisher	mfk	osp	vop
1974	0.25	0.25	0.25	0.25	0.75
1975	0.00	0.00	0.00	0.00	0.00
1976	0.50	0.50	0.50	0.50	0.50
1977	0.50	0.50	0.50	0.50	0.50
1978	0.50	0.75	0.50	0.00	0.00
1979	0.50	0.50	0.50	0.50	0.50
1980	0.50	1.00	0.50	0.75	1.00
1981	0.00	0.00	0.00	0.50	0.75
1982	0.50	0.25	0.50	0.25	0.50
1983	1.00	0.75	1.00	0.75	0.75
1984	0.25	0.75	0.25	0.75	0.75
1985	0.25	0.00	0.25	0.75	0.00
1986	0.00	0.00	0.00	0.00	0.00
1987	0.00	0.00	0.00	0.00	0.00
1988	0.00	0.00	0.00	0.00	0.00
1989	0.00	0.00	0.00	0.00	0.00
1990	0.75	0.25	0.75	0.25	0.25
1991	1.00	1.00	1.00	0.00	1.00
1992	0.75	0.25	0.75	0.25	0.25
1993	0.25	0.00	0.25	1.00	0.00
1994	0.25	0.25	0.75	0.25	0.25
1995	0.00	0.00	0.00	0.00	0.00
1996	0.00	0.00	0.00	0.00	0.00
1997	0.00	0.00	0.00	0.00	0.00
1998	0.00	0.00	0.00	0.00	0.00
1999	0.25	0.25	0.00	0.25	0.00
2000	0.25	0.25	0.00	0.25	0.25

Tabelle B.2: Ex-Post-Ante Fehlerraten bei Projektion auf eine Dimension

	adapt2	fisher	mfk	osp	vop
1974	0.50	0.50	0.25	0.50	0.50
1975	0.00	0.00	0.00	0.00	0.00
1976	0.00	0.00	0.75	0.25	0.00
1977	0.50	0.50	0.75	0.50	0.50
1978	0.00	0.00	0.00	0.00	0.00
1979	0.50	0.50	0.50	0.50	0.50
1980	0.25	0.50	0.50	0.25	0.25
1981	0.00	0.00	0.50	0.00	0.00
1982	0.50	0.50	0.25	0.50	0.50
1983	0.00	0.00	1.00	0.00	0.00
1984	0.00	0.00	0.50	0.00	0.00
1985	0.00	0.00	0.50	0.00	0.00
1986	0.00	0.00	0.00	0.00	0.00
1987	0.00	0.00	0.00	0.00	0.00
1988	0.00	0.00	0.00	0.00	0.00
1989	0.00	0.00	0.00	0.00	0.00
1990	0.25	0.25	0.25	0.25	0.25
1991	0.25	1.00	0.50	0.25	0.25
1992	0.00	0.00	0.00	0.00	0.00
1993	0.00	0.00	0.75	0.00	0.00
1994	0.25	0.25	0.25	0.25	0.25
1995	0.00	0.00	0.00	0.00	0.00
1996	0.00	0.00	0.00	0.00	0.00
1997	0.00	0.00	0.00	0.00	0.00
1998	0.00	0.00	0.00	0.00	0.00
1999	0.00	0.50	0.00	0.00	0.00
2000	0.00	0.50	0.00	0.00	0.50

Tabelle B.3: Ex-Post-Ante Fehlerraten bei Projektion auf zwei Dimensionen

Literaturverzeichnis

Aha, D. W. (1992): Generalizing from case studies: A case study. In: *ML*, 1–10.

Barker, M. & Rayens, W. (2003): Partial least squares for discrimination. *Journal of Chemometrics*, 17, 166–173.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980): *Regression diagnostics*. Wiley.

Bohachevsky, I. O., Johnson, M. E., & Stein, M. L. (1986): Generalized simulated annealing for function optimization. *Technometrics*, 28 (3), 209–217.

Brazdil, P., Soares, C., & Pinto da Costa, J. (2003): Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. *Machine Learning*, 50 (3), 251–277.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984): *Classification and regression trees*. Wadsworth Publishing Co Inc.

Breiman, L. & Ihaka, I. (1984): Nonlinear discriminant analysis via scaling & ace. *Technical Report 40*, Department of Statistics, University of California, Berkeley.

Büning, H. (1991): *Robust and adaptive tests*. Walter de Gruyter.

Büning, H. & Trenkler, G. (1994): *Nichtparametrische statistische Methoden*. de Gruyter, 2. Auflage.

Burges, C. J. C. (1998): A Tutorial on Support Vector Machines for Pattern Recognition. *Knowledge Discovery and Data Mining*, 2 (2), 121–167.

Chang, W.-C. (1983): On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, 32, 267–275.

Cook, R. D. & Yin, X. (2001): Dimension reduction and visualization in discriminant analysis (Pkg: p147-199). *The Australian and New Zealand Journal of Statistics*, 43 (2), 147–177.

Czogiel, I., Luebke, K., Zentgraf, M., & Weihs, C. (2006): Localized linear discriminant analysis. *Technical Report 10*, Sonderforschungsbereich 475, Universität Dortmund.

De Leeuw, J., Young, F. W., & Takane, Y. (1976): Additive structure in qualitative data: An alternating least squares method with optimal scaling features. *Psychometrika*, 41, 471–504.

Devroye, L. (1988): Automatic pattern recognition: A study of the probability of error. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10 (4), 530–543.

Devroye, L., Györfi, L., & Lugosi, G. (1996): *A probabilistic theory of pattern recognition*. Springer.

Diaconis, P. & Freedman, D. (1984): Asymptotics of graphical projection pursuit. *The Annals of Statistics*, 12, 793–815.

Dietterich, T. G. (1998): Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, 10 (7), 1895–1923.

Dryden, I. L. & Mardia, K. V. (1998): *Statistical shape analysis*. John Wiley & Sons.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001): *Pattern classification*. Wiley, 2. Auflage.

Edgington, E. S. (1995): *Randomization tests*. Marcel Dekker Inc, 3. Auflage.

Fisher, R. A. (1936): The use of multiple measurements in taxonomic problems. *Annal of Eugenics*, 7, 179–188.

Fleishman, A. I. (1978): A method for simulating non-normal distributions. *Psychometrika*, 43, 521–532.

Friedman, J. H. (1987): Exploratory projection pursuit. *Journal of the American Statistical Association*, 82, 249–266.

Friedman, J. H. (1989): Regularized discriminant analysis. *Journal of the American Statistical Association*, 84, 165–175.

Fukunaga, K. (1990): *Statistical pattern recognition (Second edition).* Academic Press.

Garczarek, U. M. (2002): *Classification rules in standardized partition spaces.* Dissertation, Universität Dortmund, Fachbereich Statistik.

Garthwaite, P. H. (1994): An interpretation of partial least squares. *Journal of the American Statistical Assosiation*, 89 (425), 122–127.

Gifi, A. (1990): *Nonlinear multivariate analysis.* Wiley.

Giri, N. C. (1996): *Multivariate statistical analysis.* Marcel Dekker Inc.

Golub, G. H., Heath, M., & Wahba, G. (1979): Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21 (2), 215–223.

Guseman, J., L. F., Peters, J., B. Charles, & Walker, H. F. (1975): On minimizing the probability of misclassification for linear feature selection. *The Annals of Statistics*, 3, 661–668.

Hamamoto, Y., Kanaoka, T., & Tomita, S. (1993): On theoretical comparison between the orthonormal discriminant vectors and discriminant analysis. *The Journal of the Pattern Recognition Society*, 26, 1863–1867.

Hand, D. J. (1997): *Construction and assessment of classification rules.* John Wiley & Sons.

Harville, D. A. (1997): *Matrix Algebra From a Statisticians's Perspective.* Springer.

Hastie, T., Buja, A., & Tibshirani, R. (1995): Penalized discriminant analysis. *The Annals of Statistics*, 23 (1), 73–102.

Hastie, T., Tibshirani, R., & Buja, A. (1994): Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89 (428), 1255–1270.

Hastie, T., Tibshirani, R., & Friedman, J. (2001): *The Elements of Statistical Learning*. Springer.

Heilemann, U. & Münch, H. (1996): West german business cycles 1963-1994: A multivariate discriminant analysis. In: *CIRET-Conference in Singapore*, CIRET-Studien 50.

Helland, I. S. & Almøy, T. (1994): Comparison of prediction methods when only a few components are relevant. *Journal of the American Statistical Association*, 89 (426), 583–591.

Hinkelmann, K. & Kempthorne, O. (1994): *Design and analysis of experiments: Volume I: introduction to experimental design*. Wiley.

Hochberg, Y. & Tamhane, A. C. (1987): *Multiple comparison procedures*. John Wiley & Sons.

Holm, S. (1979): A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005): The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14 (3), 675–699.

Hu, M.-K. (1962): Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory*, IT-8, 179–187.

Johnson, M. E. & Lowe, J., Victor W. (1979): Bounds on the sample skewness and kurtosis. *Technometrics*, 21, 377–378.

Johnson, N. (1949): Systems of frequency curves generated by methods of tranlation. *Biometrika*, 36, 149–176.

Kent, J. T. (1991): Comments on "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association*, 86, 336–337.

Kshirsagar, A. M. & Arseven, E. (1975): A note on the equivalency of two discrimination procedures. *The American Statistician*, 29, 38–39.

Lele, S. & Richtsmeier, J. T. (2001): *An invariant approach to statistical analysis of shapes*. CRC Press Inc.

Li, K.-C. (1991): Sliced inverse regression for dimension reduction (C/R: p328-342). *Journal of the American Statistical Association*, 86, 316–327.

Loog, M. (1999): Approximate pairwise accuracy criteria for multiclass linear dimension reduction: Generalisations of the fisher criterion. *Technical Report 44*, WBBM Report Series, TU Delft.

Loog, M., Duin, R. P. W., & Haeb-Umbach, R. (2001): Multiclass linear dimension reduction by weighted pairwise fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23 (7), 762–766.

Loog, M., Duin, R. P. W., & Viergever, M. A. (2004): The mdf discrimination measure: fisher in disguise. *Neural Networks*, 17 (4), 563–566.

Luebke, K., Czogiel, I., & Weihs, C. (2004): A computer intensive method for choosing the ridge parameter. *Technical Report 11*, Sonderforschungsbereich 475, Universität Dortmund.

Luebke, K. & Weihs, C. (2003): Prediction optimal data analysis by means of stochastic search. In: M. Schader, W. Gaul, & M. Vichi (Hrsg.) *Between Data Science and Applied Data Analysis*, 305–312. Springer.

Luebke, K. & Weihs, C. (2004a): Generation of prediction optimal projection on latent factors by a stochastic search algorithm. *Computational Statistics & Data Analysis*, 47 (2), 297–310.

Luebke, K. & Weihs, C. (2004b): Optimal separation projection. In: J. Antoch (Hrsg.) *COMPSTAT 2004 - Proceedings in Computational Statistics*, 1429 –1437. Physica.

Luebke, K. & Weihs, C. (2005a): Improving feature extraction by replacing the fisher criterion by an upper error bound. *Pattern Recognition*, 38 (11), 2220–2223.

Luebke, K. & Weihs, C. (2005b): Prediction optimal classification of business phases. *Technical Report 41*, Sonderforschungsbereich 475, Universität Dortmund.

Mardia, K., Kent, J., & Bibby, J. M. (1979): *Multivariate Analysis*. Academic Press.

McCulloch, R. E. (1986): Some remarks on allocatory and separatory linear discrimination. *Journal of Statistical Planning and Inference*, 14, 323–330.

McLachlan, G. J. (1992): *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons.

Michie, D. e., Spiegelhalter, D. J. e., & Taylor, C. C. e. (1994): *Machine learning, neural and statistical classification*. Prentice-Hall Inc.

Miller, R. G. (1986): *Beyond ANOVA, basics of applied statistics*. John Wiley & Sons.

Mood, A. M., Graybill, F. A., & Boes, D. C. (1974): *Introduction to the Theory of Statistics*. McGraw-Hill, 3. Auflage.

Næs, T. & Indahl, U. (1998): A unified description of classical classification methods for multicollinear data. *Journal of Chemometrics*, 12, 205–220.

Næs, T. & Mevik, B.-H. (2001): Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, 15 (4), 413–426.

Nelder, J. & Mead, R. (1965): A simplex method for functional minimization. *Computer Journal*, 7, 308–313.

Nocairi, H., Qannari, E. M., Vigneau, E., & Bertrand, D. (2005): Discrimination on latent components with respect to patterns. application to multicollinear data. *Computational Statistics & Data Analysis*, 48, 139–147.

Okada, T. & Tomita, S. (1985): An optimal orthonormal system for discriminant analysis. *The Journal of the Pattern Recognition Society*, 18, 139–144.

Pfahringer, B., Bensusan, H., & Giraud-Carrier, C. (2000): Meta-learning by landmarking various learning algorithms. In: *Proceedings of the Seventeenth International Conference on Machine Learning, ICML'2000*, 743–750. San Francisco, California: Morgan Kaufmann.

Press, W., Flannery, B., Teukolsky, S., & Vetterling, W. (1992): *Numerical Recipes in C*. Cambridge University Press, 2. Auflage.

Prokop, R. J. & Reeves, A. P. (1992): A survey of moment-based techniques for unoccluded object representation and recognition. *CVGIP: Graphical Models and Image Processing*, 54 (5), 438–460.

Rao, C. R. (1948): The utilization of multiple measurements in problems of biological classification (with discussion). *Journal of the Royal Statistical Society series B*, 10, 159–203.

Raudys, S. & Duin, R. (1998): On expected classification error of the fisher linear classifier with pseudo-inverse covariance matrix. *Pattern Recognition Letters*, 19, 385–392.

Reinsel, G. C. & Velu, R. P. (1998): *Multivariate Reduced-Rank Regression, Theory and Applications*. Springer.

Rendell, L. A. & Cho, H. (1990): Empirical learning as a function of concept character. *Machine Learning*, 5, 267–298.

Ripley, B. D. (1996): *Pattern recognition and neural networks*. Cambridge University Press.

Röhl, M. C., Weihs, C., & Theis, W. (2002): Direct minimization of error rates in multivariate classification. *Computational Statistics*, 17, 29–46.

Salamon, P., Sibani, P., & Frost, R. (2002): *Facts, Conjectures and Improvement for Simulated Annealing*. Monographs on Mathematical Modeling and Computation. SIAM.

Santner, T. J., Williams, B., & Notz, W. (2003): *The Design and Analysis of Computer Experiments*. Springer-Verlag.

Schervish, M. J. (1984): Linear discrimination for three known normal populations. *Journal of Statistical Planning and Inference*, 10, 167–175.

Schmidli, H. (1995): *Reduced Rank Regression*. Physica Verlag.

Seber, G. A. F. (1984): *Multivariate observations*. Wiley.

Shao, J. (1999): *Mathematical statistics*. Springer-Verlag Inc.

Stapleton, J. H. (1995): *Linear Statistical Models*. Wiley.

Tadikamalla, P. R. (1980): On simulating non-normal distributions. *Psychometrika*, 45, 273–279.

Weihs, C. & Jessenberger, J. (1999): *Statistische Methoden zur Qualitätssicherung und -optimierung in der Industrie*. Wiley VCH.

Welch, W. J., Buck, R. J., Sacks, J., Wynn, H. P., Mitchell, T. J., & Morris, M. D. (1992): Screening, predicting, and computer experiments. *Technometrics*, 34, 15–25.

Wettschereck, D. & Dietterich, T. G. (1995): An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms. *Machine Learning*, 19 (1), 5–27.

Wilks, S. S. (1963): *Mathematical Statistics*. Wiley, 2. Auflage.

Wolpert, D. H. (2001): The supervised no-free-lunch theorems. In: *In Proceedings of the 6th On-line World Conference on Soft Computing in Industrial Applications*, Springer Engineering Series, 25–42.

Yan, J., Zhang, B., Yan, S., Yang, Q., Li, H., Chen, Z., Xi, W., Fan, W., Ma, W.-Y., & Cheng, Q. (2004): Immc: incremental maximum margin criterion. In: *KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, 725–730.

Zhu, M. & Hastie, T. (2003): Feature extraction for non-parametric discriminant analysis. *Journal of Computational and Graphical Statistics*, 12 (1), 101–120.