# Ausreißeridentifikation für kategoriale und funktionale Daten im generalisierten linearen Modell

## Dissertation

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

Der Fakultät Statistik
der Technischen Universität Dortmund

vorgelegt von
**André Rehage**

Dortmund 2016

Erstgutachterin: Prof. Dr. Sonja Kuhnt
Zweitgutachter: Prof. Dr. Roland Fried
Tag der mündlichen Prüfung: 09.03.2017

# INHALTSVERZEICHNIS

# 1. EINLEITUNG

In der vorliegenden Arbeit werden Verfahren zur Identifikation von Ausreißern in generalisierten linearen Modellen entwickelt und anhand statistischer Gütekriterien, Simulationsstudien und realer Datensätze überprüft. Der Fokus liegt dabei auf kategorialen und funktionalen Zielgrößen.

Das generalisierte lineare Modell (GLM, Nelder und Wedderburn, 1972) ermöglicht eine Quantifizierung des Einflusses unabhängiger Variablen auf eine abhängige Variable anhand einer Stichprobe. Ziel ist es, die Linearkombination von Prädiktorvariablen zu schätzen, die die abhängige Variable am besten beschreibt. Es wird dabei angenommen, dass alle Realisationen der abhängigen Variable demselben datengenerierenden Prozess entstammen. Ist diese Annahme verletzt, werden die aus abweichenden Prozessen stammenden Beobachtungen Ausreißer genannt. Die Identifikation solcher Ausreißer ist aus zwei unterschiedlichen Gründen in jeder statistischen Analyse essenziell:

1. Ausreißer können störende Beobachtungen sein, wenn sie durch Messfehler verursacht werden. Ziel ist es, sie aus der Stichprobe zu entfernen, damit sie nicht-robuste Parameterschätzer des GLM nicht verzerren.

2. Ausreißer können Hinweise auf bisher nicht beachtete Prädiktorvariablen und Phänomene sowie falsche Modellspezifikationen geben und damit die wichtigsten Beobachtungen der Stichprobe sein.

Eine Anwendung für Ausreißeridentifikation in GLMs mit kategorialen und funktionalen Zielgrößen ist das Hochgeschwindigkeitsflammspritzen – ein thermisches Beschichtungsverfahren, um Substratoberflächen mit gewünschten Eigenschaften zu produzieren. Die Modellierung und Kontrolle dieses Verfahrens steht im Mit-

telpunkt des Teilprojekts B1 des Sonderforschungsbereichs 823 der Deutschen Forschungsgemeinschaft. Da zur Analyse der Qualität der beschichteten Oberflächen zerstörende, zeitaufwändige und kostenintensive Prüfungen notwendig sind, sollen diese Qualitätsmerkmale durch andere, leichter und zerstörungsfrei zu messende Eigenschaften ersetzt werden. Die Pulverpartikel, die auf das Substrat gespritzt werden, scheiden sich bei sehr kurzer Beschichtungszeit in Form einzelner sogenannter Splats ab (Tillmann *et al.*, 2013). Die Merkmalskombinationen dieser Splats können in Form von Kontingenztafeln erhoben werden und auf Basis generalisierter linearer Modelle mit geeigneten Methoden auf Ausreißer analysiert werden. Weitere Merkmale des Hochgeschwindigkeitsflammspritzens umfassen die Temperatur und Geschwindigkeit der Partikel im Flug. Mit Hilfe generalisierter linearer Modelle mit funktionalen Zielgrößen (Scheipl *et al.*, 2016) lässt sich der Einfluss der unabhängigen Variablen auf die Temperatur und Geschwindigkeit schätzen und eine tagesgenaue Adjustierung vornehmen (Tillmann *et al.*, 2012). Die Identifikation möglicher Ausreißer in den Kontingenztafeln oder funktionalen Zielgrößen kann dabei helfen, einen tieferen Einblick in den datengenerierenden Prozess zu erhalten und die Schätzung der Parameter zu verbessern.

Teilergebnisse der vorliegenden Arbeit wurden in folgenden Aufsätzen und Buchkapiteln vorab publiziert: Kuhnt *et al.* (2016); Kuhnt und Rehage (2016); Rehage und Kuhnt (2014); Kuhnt *et al.* (2014); Kuhnt und Rehage (2013); Tillmann *et al.* (2012). Für die statistische Programmiersprache R (R Core Team, 2016) sind im Rahmen der Dissertation die Pakete `alphaOutlier` (Rehage und Kuhnt, 2016) und `FUNTA` (Rehage, 2016) entstanden.

Bevor in der Praxis Ausreißer identifiziert werden können, ist die Vereinbarung einer möglichst objektiven Definition des Ausreißers wünschenswert. In Kapitel 2 wird dazu das Konzept der $\alpha$-Ausreißer vorgestellt. Anhand dieses objektiven Konzepts können $\alpha$-Ausreißerregionen für Zufallsvariablen mit beliebiger, aber bekannter Wahrscheinlichkeitsdichte hergeleitet werden. Realisationen dieser Zufallsvariable werden als $\alpha$-Ausreißer bezeichnet, falls sie sich in der $\alpha$-Ausreißerregion befinden. Für gewisse Verteilungen ist die Bestimmung ihrer $\alpha$-Ausreißerregion sehr aufwändig. Für Multinomialverteilungen mit Parametern $n \in \mathbb{N}, \boldsymbol{\theta} \in \mathbb{R}^k$ wird daher

ein Algorithmus vorgestellt, der die exakte Bestimmung auch für große $n, k$ ermöglicht. Die Güte der Chi-Quadrat-Approximation standardisierter multinomialverteilter Zufallsvariablen wird in Bezug auf die Elemente in der $\alpha$-Ausreißerregion ebenfalls überprüft. Des Weiteren wird das Konzept der $\alpha$-Ausreißer mit Hilfe robuster Kerndichteschätzer (Kim und Scott, 2012) auch auf den Fall unbekannter Verteilungstypen übertragen. Außerdem wird gezeigt, dass der Ausreißeridentifizierer basierend auf der Populationsversion des modifizierten Boxplots (Tukey, 1977) bei passender Wahl von $\alpha$ mit der $\alpha$-Ausreißerregion diverser Verteilungen mit univariater, stetiger, streng steigend-fallender und symmetrischer Wahrscheinlichkeitsdichte übereinstimmt. Weiterhin kann ein $\alpha$ gefunden werden, so dass die $\alpha$-Ausreißerregion der bivariaten Standardnormalverteilung mit der Ausreißerregion der Populationsversion des Bagplots (Rousseeuw *et al.*, 1999) übereinstimmt.

Kapitel 3 behandelt Grundlagen generalisierter linearer Modelle. Durch die Annahme poissonverteilter Zellhäufigkeiten können auch Kontingenztafeln im Rahmen generalisierter linearer Modelle analysiert werden. In Verbindung mit dem Log-Link wird dieser Spezialfall auch als loglineares Poissonmodell bezeichnet. Hierfür werden Schätzer und ihre Eigenschaften vorgestellt. Zur Modellierung funktionaler Zielgrößen werden Grundlagen der funktionalen Datenanalyse (FDA, Ramsay und Silverman, 2005) präsentiert. Die Ansätze zur Behandlung funktionaler Zielgrößen im gewöhnlichen (Faraway, 1997; Reiss *et al.*, 2010) und generalisierten linearen Modell (Hall *et al.*, 2008; Goldsmith *et al.*, 2015; Scheipl *et al.*, 2016) werden eingeführt und besprochen. Dabei dient im Modell von Scheipl *et al.* (2016) eine pönalisierte Log-Likelihood der Schätzung der Regressionsfunktionen.

Die Identifikation einzelner Zellen in Kontingenztafeln als Ausreißer steht im Zentrum von Kapitel 4. Hierfür wird das Minimalmusterverfahren (Kuhnt, 2000) herangezogen. Ein Minimalmuster besteht aus einer Teilmenge der Zellen einer Kontingenztafel, deren Elemente als potenziell ausreißerfrei aufgefasst werden und daher zur Schätzung der Parameter des loglinearen Poissonmodells verwendet werden können. Zur Wahl von Zellen einer Kontingenztafel, die Teil eines Minimalmusters sind, wird die Relevanz bestimmter geometrischer Strukturen, der sogenannten ŧ-Schlingen, hervorgehoben. Es kann Situationen geben, in denen die Ausreißermen-

gen des Minimalmusterverfahrens nicht eindeutig sind. Basierend auf Minimalmustern wird der OMPC-Algorithmus vorgeschlagen, der die Eindeutigkeit der Ausreißermenge garantiert. Der OMPC-Identifizierer wird bezüglich seiner Performanz in Simulationsstudien und realen Datensätzen mit Ausreißeridentifizierern basierend auf dem $L_1$- (Hubert, 1997) sowie LTCS-Schätzer (Shane und Simonoff, 2001) verglichen. Außerdem wird ein Szenario behandelt, bei dem vollständige Kontingenztafeln im Multinomial-Unabhängigkeitsmodell in Bezug auf ihre *outlyingness* untersucht werden.

Die Identifikation funktionaler Beobachtungen als Ausreißer wird in Kapitel 5 thematisiert. Das Prinzip der $\alpha$-Ausreißer ist ohne Weiteres nicht übertragbar, da das Konzept der Wahrscheinlichkeitsdichte in unendlichdimensionalen Räumen nicht wohldefiniert ist (Delaigle und Hall, 2010). Als Ersatz können funktionale Datentiefen (Fraiman und Muniz, 2001; Cuevas *et al.*, 2006; Lopez-Pintado und Romo, 2009) betrachtet werden, die die Beobachtungen „von innen nach außen“ sortieren. Basierend auf dieser Sortierung wird ein Bootstrap-Verfahren (Febrero *et al.*, 2008) verwendet, um einen Schwellenwert zur Unterscheidung zwischen Ausreißern (kleinen Datentiefewerten) und Nicht-Ausreißern (großen Datentiefewerten) zu bestimmen. Diese Konzepte berücksichtigen jedoch nur die Lage, nicht die Form der Beobachtungen. Daher werden zwei Pseudo-Datentiefen (FUNTA, rFUNTA) basierend auf den Schnittwinkeln der zentrierten Beobachtungen entwickelt, um auch die *outlyingness* bezüglich der Form zu erfassen. Theoretische Gütekriterien für FUNTA und rFUNTA werden in Form von Implosions- und Explosionsbruchpunkten hergeleitet. Des Weiteren wird anhand von Simulationsstudien und realen Datensätzen die Performanz von FUNTA und rFUNTA im Vergleich mit weiteren Datentiefen verglichen. Falls die Verteilung der funktionalen Zufallsvariablen durch einen Gaußprozess mit bekannten Parametern bestimmt ist, werden neue Ansätze basierend auf dem $\alpha$-Ausreißerkonzept vorgestellt, um dieses Wissen in die Ausreißeridentifikation einfließen zu lassen.

Abschließend werden die Ergebnisse in Kapitel 6 zusammengefasst und Ansatzpunkte für weitere Forschung genannt.

# 2. AUSREISSERIDENTIFIKATION

Ausreißer und Methoden ihrer Identifikation sind in praktisch allen Datensituationen von Interesse. Eine einheitliche Definition des Begriffs „Ausreißer“ hat sich jedoch nicht durchgesetzt. In dieser Arbeit wird unter einem Ausreißer eine Beobachtung verstanden, die unter einer gewissen Modellannahme nicht mit dem Großteil der Beobachtungen vereinbar scheint. Formal gesprochen vermuten wir, dass dem Ausreißer ein abweichender datengenerierender Prozess zugrunde liegt.

Da in der Literatur vielfältige Möglichkeiten zur Ausreißeranalyse zu finden sind, soll hier ein kurzer Überblick geliefert werden. Oftmals werden in der Praxis keine (mehr oder weniger) formalen Kriterien angewendet, sondern etwaige Ausreißer mit der „Methode des scharfen Hinsehens“ identifiziert. Dies ist umso schwieriger, je mehr Dimensionen der interessierende Datensatz besitzt. In der deskriptiven Statistik ist der Boxplot ein populäres Werkzeug zur Datenvisualisierung und Ausreißeridentifikation: Beobachtungen werden als Ausreißer bezeichnet, falls sie nicht in einem bestimmten Abstand vom Median liegen. Dieses Verfahren kann für jede Variable separat durchgeführt werden, was allerdings mit einem Informationsverlust einhergeht. Für bivariate Daten kann der Bagplot (Rousseeuw *et al.*, 1999) eine auf dem Konzept der Halbraumtiefe (Tukey, 1975) basierende Darstellung liefern (vgl. Abschnitt 2.3). Diese und andere Möglichkeiten der Ausreißeridentifikation sind unabhängig von etwaigen Modellannahmen an die Daten. Ist es jedoch plausibel, dass eine bekannte Wahrscheinlichkeitsverteilung hinter dem datengenerierenden Prozess steht, so offenbaren sich weitere Methoden: Barnett und Lewis (1994, S. 43 ff.) besprechen diverse Kontaminationsmodelle, von denen zwei wichtige hier aufgeführt werden. Das laut Barnett und Lewis wichtigste Modell ist das *slippage model*. Darin werden alle bis auf eine kleine Anzahl von $k_{\text{out}}$ Beobachtungen aus einer Verteilung $P$ gezogen. Die übrigen $k_{\text{out}}$ Beobachtungen, die Ausreißer genannt

werden, stammen zwar aus der gleichen Verteilungsklasse, variieren aber in Bezug auf Lage *oder* Streuung. Des Weiteren existiert das flexiblere Mischungsmodell, das durch eine bekannte Anzahl $n_{\text{pop}}$ unterschiedlicher Populationen mit bekannter Wahrscheinlichkeit $\pi_i, i = 1, \ldots, n_{\text{pop}}$ charakterisiert ist. Die Ausreißer können hier auch aus unterschiedlichen Verteilungsklassen stammen sowie in Bezug auf Lage *und* Streuung abweichen. In Beispiel 2.1 wird dieses Modell veranschaulicht.

BEISPIEL 2.1. *Die Zufallsvariable $\mathcal{Y}_1$ folge der $P_1 = \aleph(0, 1)$-Verteilung, $\mathcal{Y}_2$ folge der $P_2 = \aleph(4, 0.5)$-Verteilung und $\mathcal{Y}_M$ folge der $P_M = \pi_1 P_1 + \pi_2 P_2$-Verteilung, wobei $\pi_1 = 1 - \pi_2 = 0.9$. In Abbildung 2.1 sind die mit $\pi_1$ resp. $\pi_2$ multiplizierten Wahrscheinlichkeitsdichten von $P_1$ blau und von $P_2$ orange eingezeichnet. Des Weiteren ist die Dichte von $P_M$ gestrichelt gekennzeichnet. $P_1$ wird als Population der Nicht-Ausreißer und $P_2$ als Population der Ausreißer aufgefasst, da $\pi_2 \ll \pi_1$. $P_M$ wird als Mischverteilung bezeichnet. Für $\mathcal{Y}_M$ wird mit einer Wahrscheinlichkeit von $0.1$ ein Ausreißer realisiert.*

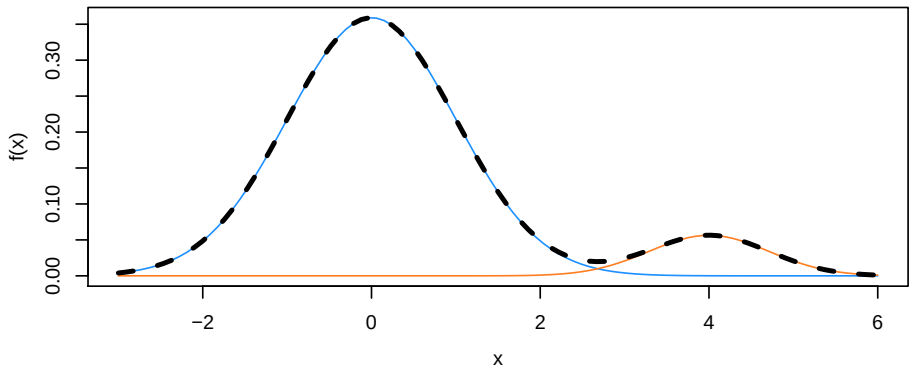



Abbildung 2.1: Mischverteilung (gestrichelt) zweier univariater Normalverteilungen

Eine solche Situation erinnert an das Konzept einer bekannten Anzahl unterschiedlicher Populationen, welches bei Klassifikationsverfahren wie der Diskriminanzanalyse verwendet wird. Ein Nachteil dieser Herangehensweise ist, dass die Anzahl bzw. Verteilung der Populationen *a priori* festgelegt werden muss. Einen anderen Ansatz verfolgt das Konzept der $\alpha$-Ausreißer (Davies und Gather, 1989, 1993), für das die Anzahl bzw. Verteilung der Ausreißerpopulationen nicht bekannt sein muss. Im folgenden Abschnitt wird dieses Konzept definiert und veranschaulicht.

## 2.1 $\alpha$-AUSREISSER BEI BEKANNTER VERTEILUNGSFAMILIE

Das Konzept der $\alpha$-Ausreißer, welches zur Identifikation von Ausreißern aus unterschiedlichen Populationen geeignet ist, wurde erstmals von Davies und Gather (1989, 1993) für normalverteilte Zufallsvariablen definiert. Um $\alpha$-Ausreißer zu bestimmen, muss lediglich die Wahrscheinlichkeitsdichte der Nicht-Ausreißer bekannt sein. Christmann (1992, 1998) verallgemeinerte das Konzept auf beliebige Verteilungsfamilien und einige Regressionstypen. Kuhnt (2000) betrachtete $\alpha$-Ausreißer in Kontingenztafeln. Gather *et al.* (2003) behandelten verschiedene Typen sogenannter $\alpha$-Inlierregionen im linearen Regressionsmodell. Definition 2.2 basiert auf der Darstellung in Gather *et al.* (2003).

DEFINITION 2.2. $\alpha$-Ausreißerregion
*Sei $\mathcal{P}$ eine Verteilungsfamilie auf einem messbaren Raum $(\Omega, \mathcal{A})$. Weiterhin werde $(\Omega, \mathcal{A})$ von einem $\sigma$-endlichen Maß $\nu$ dominiert, so dass $P \in \mathcal{P}$ eine $\nu$-Dichte $f$ besitzt. Der Träger von $P$ ist gegeben durch supp$(P)$ $\forall$ $P \in \mathcal{P}$. Außerdem gelte supp$(\mathcal{P}) = \bigcup_{P\in\mathcal{P}}$ supp$(P)$. Für festes $\alpha \in ]0,1[$ ist die* $\alpha$-Ausreißerregion *von $P \in \mathcal{P}$ definiert durch*

$$\begin{aligned} \text{out}(\alpha, P) &= \{x \in \textit{supp}(\mathcal{P}) : f(x) < B(\alpha)\}, \quad \textit{wobei} \\ B(\alpha) &= \sup\{K > 0 : P(\{y : f(y) < K\}) \leq \alpha\} \end{aligned} \tag{2.1}$$

*gelte. Das Komplement der $\alpha$-Ausreißerregion wird $\alpha$-Inlierregion genannt.*

Definition 2.2 kann im Fall stetiger Verteilungen so interpretiert werden, dass jene Beobachtungen als $\alpha$-Ausreißer angesehen werden, die in einem oder mehreren Intervallen liegen, deren Wahrscheinlichkeitsmasse sehr klein – sprich $\alpha$ – ist. Gleichzeitig sollte die Gesamt-Intervalllänge der $\alpha$-Ausreißerregion maximiert sein. Bei diskreten Verteilungen können die Punktmassen der Größe nach geordnet werden und beginnend mit der kleinsten Punktmasse die Wahrscheinlichkeiten so lange kumuliert werden, bis eine weitere Punktmasse zu einer Überschreitung von $\alpha$ führen würde. Alternativ ist es umgekehrt möglich, beginnend mit der größten Punktmasse bis zu einer Wahrscheinlichkeitsmasse von $1-\alpha$ zu kumulieren und so den Träger der $\alpha$-Inlierregion zu erhalten.

Zur Klassifikation der Beobachtungen eines konkreten Datensatzes als $\alpha$-Ausreißer wird in Definition 2.3 ein einschrittiger $\alpha$-Ausreißeridentifizierer formuliert, der auf der angenommenen Verteilungsfamilie $\mathcal{P}$ und dem Parameter der Verteilung $\boldsymbol{\theta}$ beruht. In der Regel ist $\boldsymbol{\theta}$ unbekannt und muss entweder – z. B. durch Vorwissen – festgelegt oder geschätzt werden. Deswegen wird im Folgenden die Schreibweise $\widehat{\boldsymbol{\theta}}$ verwendet. Bei manchen Verteilungen hängt zudem der Träger von den Parametern der Verteilung ab, wie bei der vierparametrigen Beta-Verteilung. Es sei deshalb im Folgenden angenommen, dass der Träger der Verteilung bekannt ist.

DEFINITION 2.3. Einschrittiger $\alpha$-Ausreißeridentifizierer
*Es bezeichne $\mathbf{y} \in supp(\mathcal{P})$ eine Beobachtung, für die entschieden werden soll, ob sie bezüglich der Zufallsvariable $\boldsymbol{\mathcal{Y}}$ ein $\alpha$-Ausreißer ist. $\boldsymbol{\mathcal{Y}}$ habe die Wahrscheinlichkeitsdichte $f(\cdot, \boldsymbol{\theta})$, wobei $\boldsymbol{\theta}$ der unbekannte Parameter ist. Die $\alpha$-Ausreißerregion wird für den festgelegten oder geschätzten Parameter $\widehat{\boldsymbol{\theta}}$ durch $\mathrm{out}(\alpha, P_{\widehat{\boldsymbol{\theta}}})$ bestimmt. Dann wird die Abbildung $\mathrm{OI} : supp(\mathcal{P}) \to \{0, 1\}$ mit*

$$\mathrm{OI}(\mathbf{y}; \alpha, P_{\widehat{\boldsymbol{\theta}}}) = \begin{cases} 1, & \textit{falls } \mathbf{y} \in \mathrm{out}(\alpha, P_{\widehat{\boldsymbol{\theta}}}), \\ 0, & \textit{sonst}, \end{cases}$$

*als* einschrittiger $\alpha$-Ausreißeridentifizierer *bezeichnet. Falls $\mathrm{OI}(\mathbf{y}; \alpha, P_{\widehat{\boldsymbol{\theta}}}) = 1$, wird $\mathbf{y}$ als $\alpha$-Ausreißer klassifiziert, andernfalls wird $\mathbf{y}$ als $\alpha$-Inlier bzw. $\alpha$-Nicht-Ausreißer klassifiziert.*

Liegen mehrere Beobachtungen $\mathbf{y}_1, \ldots, \mathbf{y}_k \in \mathrm{supp}(\mathcal{P})$ vor, bezeichnet

$$\mathbf{OI}(\mathbf{y}_1, \ldots, \mathbf{y}_k; \alpha, P_{\widehat{\boldsymbol{\theta}}}) = (\mathrm{OI}(\mathbf{y}_1; \alpha, P_{\widehat{\boldsymbol{\theta}}}), \ldots, \mathrm{OI}(\mathbf{y}_k; \alpha, P_{\widehat{\boldsymbol{\theta}}}))^\top$$

den vektorwertigen $\alpha$-Ausreißeridentifizierer.

BEMERKUNG 2.4. *(Gather* et al.*, 2003)*
*Bezüglich einer Stichprobe vom Umfang $k$ kann von Interesse sein, die Wahrscheinlichkeit mindestens eine reguläre Beobachtung als $\alpha$-Ausreißer zu identifizieren, nach oben zu begrenzen. Hierfür ist die bekannte Adjustierung $\alpha_k = 1-(1-\alpha)^{1/k}$ (Šidák, 1967) anwendbar.*

Im Folgenden werden einige Beispiele von $\alpha$-Ausreißerregionen in konkreten Ver-

teilungen beschrieben.

BEISPIEL 2.5. *Sei $\mathcal{Y}_1$ eine poissonverteilte Zufallsvariable mit Parameter $\theta = 6$ sowie $\mathcal{Y}_2$ eine binomialverteilte Zufallsvariable mit Parametern $n = 6, \pi = 0.6$. Abbildung 2.2 zeigt die $0.1$-Ausreißerregionen (orange) für die zu $\mathcal{Y}_1$ und $\mathcal{Y}_2$ gehörenden diskreten Dichten. Die Grenze für $\alpha$ in (2.1) kann hier nicht voll ausgeschöpft werden. Die $0.1$-Ausreißerregion der $Poi(6)$-Verteilung lautet $\mathbb{N}_0 \backslash \{2, \ldots, 10\}$, wobei $P(\mathcal{Y}_1 \in \mathbb{N}_0 \backslash \{2, \ldots, 10\}) = 0.06$. Falls beispielsweise die $0.06$-Ausreißerregion von Interesse ist, wäre sie hier identisch mit der $0.1$-Ausreißerregion. Die $0.1$-Ausreißerregion der $Bin(6, 0.6)$-Verteilung lautet $\{0, 1, 6\}$, tatsächlich ist $P(\mathcal{Y}_2 \in \{0, 1, 6\}) = 0.088$. Ein Konzept ähnlich der Randomisierung zur Ausschöpfung eines Testniveaus bei diskreten Verteilungen wurde für $\alpha$-Ausreißerregionen noch nicht formuliert, seine Übertragbarkeit ist jedoch offensichtlich.*

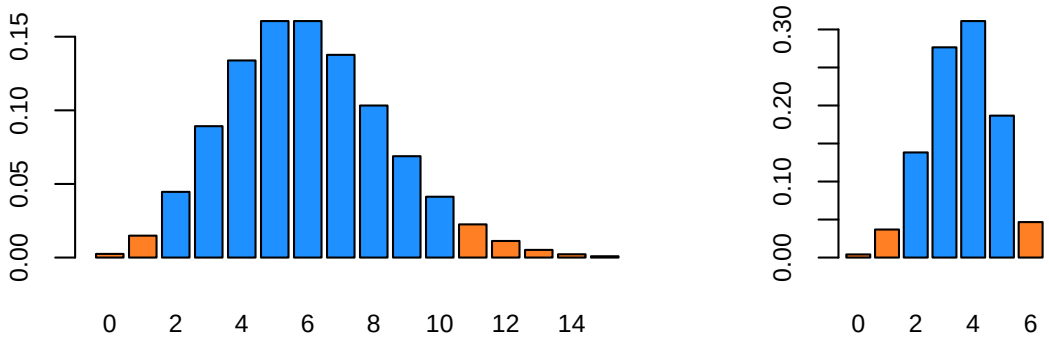



Abbildung 2.2: 0.1-Ausreißerregionen (orange) verschiedener Verteilungen. Links: Poissonverteilung mit $\lambda = 6$, rechts: Binomialverteilung mit $n = 6, \pi = 0.6$.

Stammt ein gewisser Anteil der Realisationen tatsächlich aus einer anderen Verteilung, können diese Realisationen mit hoher Wahrscheinlichkeit als $\alpha$-Ausreißer klassifiziert werden. Das nächste Beispiel veranschaulicht solch einen Fall.

BEISPIEL 2.6. Fortsetzung von Beispiel 2.1

*Unter der Voraussetzung eines bekannten wahren Ausreißer-Anteils von $\alpha = 0.1$ lässt sich herleiten: Wird fälschlicherweise davon ausgegangen, dass für die Zufallsvariable $\mathcal{Y}_M \sim \aleph(0, 1)$ gilt, folgt: $\text{out}(\alpha, P_1) = \mathbb{R} \backslash [-1.645, 1.645]$. Die Wahrscheinlichkeit, dass $\mathcal{Y}_1$ als $\alpha$-Ausreißer klassifiziert wird, beträgt $P(\mathcal{Y}_1 \in \text{out}(\alpha, P_1)) = 0.1$. Für die kleinere Population*

*gilt* $P(\mathcal{Y}_2 \in \text{out}(\alpha, P_1)) = 1$. *Für die Mischungsvariable gilt* $P(\mathcal{Y}_M \in \text{out}(\alpha, P_1)) = 0.9 \cdot 0.1 + 0.1 \cdot 1 = 0.19$. *Die Wahrscheinlichkeit, eine Realisation von* $\mathcal{Y}_M$ *als* $\alpha$*-Ausreißer zu identifizieren, ist also deutlich größer als* $\alpha = 0.1$.

Während bei multimodalen stetigen Verteilungen häufig (numerische) Integration nötig ist, um die Ausreißerregion zu bestimmen, existieren bei Dichten mit besonders einfacher Form Zusammenhänge mit den Quantilen ihrer Verteilung: Bei stetigen, streng monoton fallenden Wahrscheinlichkeitsdichten stimmen die Grenzen der $\alpha$-Ausreißerregionen mit dem $(1-\alpha)$-Quantil der Verteilung $q_{1-\alpha}$ und der oberen Grenze des Trägers überein und bei stetigen, streng monoton steigenden Wahrscheinlichkeitsdichten mit der unteren Grenze des Trägers und dem $\alpha$-Quantil $q_\alpha$. Weitere Dichtetypen lassen sich wie folgt definieren:

DEFINITION 2.7. *Sei* $f(\cdot, \boldsymbol{\theta})$ *eine Wahrscheinlichkeitsdichte auf dem Träger* $supp(P_{\boldsymbol{\theta}})$. $f(\cdot, \boldsymbol{\theta})$ *wird* steigend-fallend *genannt, falls mindestens ein* $y_{\text{mod}} \in supp(P_{\boldsymbol{\theta}})$ *existiert, so dass für alle* $y_1, y_2, y_3, y_4 \in supp(P_{\boldsymbol{\theta}})$ *mit* $y_1 < y_2 < y_{\text{mod}} < y_3 < y_4$ *gilt, dass*

$$f(y_1) \leq f(y_2) \leq f(y_{\text{mod}}) \geq f(y_3) \geq f(y_4). \tag{2.2}$$

*Weiterhin wird* $f(\cdot, \boldsymbol{\theta})$ streng steigend-fallend *genannt, falls genau ein* $y_{\text{mod}}$ *existiert und in* (2.2) *„<“ statt „≤“ und „>“ statt „≥“ gilt.* $f(\cdot, \boldsymbol{\theta})$ *wird* fallend-steigend *genannt, falls mindestens ein* $y_{\text{min}} \in supp(P_{\boldsymbol{\theta}})$ *existiert, so dass für alle* $y_1, y_2, y_3, y_4 \in supp(P_{\boldsymbol{\theta}})$ *mit* $y_1 < y_2 < y_{\text{min}} < y_3 < y_4$ *gilt, dass*

$$f(y_1) \geq f(y_2) \geq f(y_{\text{min}}) \leq f(y_3) \leq f(y_4). \tag{2.3}$$

*Weiterhin wird* $f(\cdot, \boldsymbol{\theta})$ streng fallend-steigend *genannt, falls genau ein* $y_{\text{min}}$ *existiert und in* (2.3) *„>“ statt „≥“ und „<“ statt „≤“ gilt.*

Als Beispiel für streng fallend-steigende Wahrscheinlichkeitsdichten kann die Beta$(a, b)$-Verteilung genannt werden, wenn $0 < a < 1$ und $0 < b < 1$ gilt. Bei stetigen, symmetrischen, steigend-fallenden Wahrscheinlichkeitsdichten stimmen die Grenzen der $\alpha$-Inlierregion mit den $\alpha/2$- und $(1-\alpha/2)$-Quantilen der jeweiligen Verteilung überein (siehe Gather *et al.*, 2003, Lemma 2 und dessen Erweiterung in

Kuhnt und Rehage, 2013, Lemma 6.1). Bei asymmetrischer Dichte geben Gather *et al.* (2003) ein Kriterium für die Grenzen der $\alpha$-Ausreißerregion an, ohne es formal zu beweisen. Dies wird mit Lemma 2.8 nachgeholt.

LEMMA 2.8. *Sei $P_{\boldsymbol{\theta}}$ eine stetige Verteilung mit zugehöriger Verteilungsfunktion $F(\cdot, \boldsymbol{\theta})$ und streng steigend-fallender Dichte $f(\cdot, \boldsymbol{\theta})$ auf dem Träger supp$(P_{\boldsymbol{\theta}}) = ]a_{P_{\boldsymbol{\theta}}}, b_{P_{\boldsymbol{\theta}}}[$. Weiterhin gelte $\lim_{y \searrow a_{P_{\boldsymbol{\theta}}}} f(y, \boldsymbol{\theta}) = 0 = \lim_{y \nearrow b_{P_{\boldsymbol{\theta}}}} f(y, \boldsymbol{\theta})$. Dann existiert für jedes $\alpha \in ]0, 1[$ eine $\alpha$-Ausreißerregion bezüglich $P_{\boldsymbol{\theta}}$, die definiert ist durch*

$$\begin{aligned} \text{out}(\alpha, P_{\boldsymbol{\theta}}) &= \{y \in \textit{supp}(P_{\boldsymbol{\theta}}) \setminus [y_1, y_2]\}, \qquad \textit{mit} \\ 1 - \alpha &= F(y_2, \boldsymbol{\theta}) - F(y_1, \boldsymbol{\theta}) =: h(y_1, y_2; \boldsymbol{\theta}) \qquad \textit{und} \qquad (2.4) \\ f(y_1, \boldsymbol{\theta}) &= f(y_2, \boldsymbol{\theta}). \qquad (2.5) \end{aligned}$$

*Beweis:* Sei $y_{\text{mod}} := \text{argmax}_y f(y, \boldsymbol{\theta})$. Da $f(\cdot, \boldsymbol{\theta})$ streng steigend-fallend ist, ist $y_{\text{mod}}$ eindeutig und die Struktur der möglichen $\alpha$-Ausreißerregionen ist gegeben durch $]a_{P_{\boldsymbol{\theta}}}, \tilde{y}_1(\alpha)[ \cup ]\tilde{y}_2(\alpha), b_{P_{\boldsymbol{\theta}}}[$, wobei $a_{P_{\boldsymbol{\theta}}} < \tilde{y}_1(\alpha) < y_{\text{mod}} < \tilde{y}_2(\alpha) < b_{P_{\boldsymbol{\theta}}}$ gilt. Es ist zu zeigen, dass für jedes $\alpha \in ]0, 1[$ ein Tupel $(\tilde{y}_1(\alpha), \tilde{y}_2(\alpha))$ existiert, für das (2.4) und (2.5) mit $y_1 = \tilde{y}_1$ und $y_2 = \tilde{y}_2$ gelten.

Da $f(\cdot, \boldsymbol{\theta})$ streng steigend-fallend ist und $\lim_{y \searrow a_{P_{\boldsymbol{\theta}}}} f(y, \boldsymbol{\theta}) = 0 = \lim_{y \nearrow b_{P_{\boldsymbol{\theta}}}} f(y, \boldsymbol{\theta})$ gilt, existiert für jedes $\tilde{y}_1(\alpha)$ genau ein $\tilde{y}_2(\alpha)$ für das $f(\tilde{y}_1(\alpha), \boldsymbol{\theta}) = f(\tilde{y}_2(\alpha), \boldsymbol{\theta})$ folgt. Des Weiteren bezeichne $g$ eine Hilfsfunktion, mit deren Hilfe (2.4) und (2.5) zusammengefasst werden sollen. Definiere $g : ]a_{P_{\boldsymbol{\theta}}}, y_{\text{mod}}[ \to ]y_{\text{mod}}, b_{P_{\boldsymbol{\theta}}}[$ mit $g(\tilde{y}_1, \boldsymbol{\theta}) = \tilde{y}_2$, wobei $f(\tilde{y}_1, \boldsymbol{\theta}) = f(\tilde{y}_2, \boldsymbol{\theta})$ und $\tilde{y}_2 > y_{\text{mod}}$. Je größer $\tilde{y}_1$ ist, desto kleiner muss $\tilde{y}_2$ sein, damit $f(\tilde{y}_1, \boldsymbol{\theta}) = f(\tilde{y}_2, \boldsymbol{\theta})$ gilt. Damit ist $g$ monoton fallend und wegen der streng steigend-fallenden Dichte auch streng monoton fallend.

Alternativ kann $g$ wie folgt geschrieben werden: $g(\tilde{y}_1, \boldsymbol{\theta}) = f_{\text{fall}}^{-1}(f(\tilde{y}_1, \boldsymbol{\theta}), \boldsymbol{\theta})$, wobei $f_{\text{fall}}^{-1}$ die stetige Umkehrfunktion von $f_{\text{fall}} : ]y_{\text{mod}}, b_{P_{\boldsymbol{\theta}}}[ \to ]0, f(y_{\text{mod}})[$ mit $f_{\text{fall}}(y, \boldsymbol{\theta}) = f(y, \boldsymbol{\theta})$ ist. Die Funktion $f_{\text{fall}}$ ist invertierbar, da sie auf dem gewählten Träger streng monoton fallend und damit bijektiv ist. Als Komposition stetiger Funktionen ist $g$ ebenfalls stetig.

Der Beweis von Lemma 2.8 folgt, wenn für alle $\alpha \in ]0, 1[$ genau ein $\tilde{y}_1 \in ]a_{P_{\boldsymbol{\theta}}}, y_{\text{mod}}[$ existiert, so dass $1-\alpha = F(g(\tilde{y}_1, \boldsymbol{\theta}), \boldsymbol{\theta}) - F(\tilde{y}_1, \boldsymbol{\theta}) =: h(\tilde{y}_1, \boldsymbol{\theta})$. Dies entspricht der Bijek-

tivität von $h : ]a_{P_{\boldsymbol{\theta}}}, y_{\text{mod}}[ \to ]0, 1[$, gewährleistet durch die folgenden Eigenschaften:

1. Stetigkeit: $F(g(\cdot, \boldsymbol{\theta}))$ ist eine Komposition stetiger Funktionen. Daher ist $h$ als Differenz stetiger Funktionen ebenfalls stetig.

2. Strenge Monotonie: $h$ ist streng monoton fallend. Für alle $\tilde{y}_1 \in ]a_{P_{\boldsymbol{\theta}}}, y_{\text{mod}}[$ gilt:

$$\begin{aligned}\left.\frac{\partial h}{\partial \tilde{y}}\right|_{\tilde{y}=\tilde{y}_1} &= \left.\frac{\partial (F \circ g)}{\partial \tilde{y}}\right|_{\tilde{y}=\tilde{y}_1} - \left.\frac{\partial F}{\partial \tilde{y}}\right|_{\tilde{y}=\tilde{y}_1} \\ &= \underbrace{\left.\frac{\partial F}{\partial \tilde{y}}\right|_{\tilde{y}=g(\tilde{y}_1,\boldsymbol{\theta})}}_{>0} \cdot \underbrace{\left.\frac{\partial g}{\partial \tilde{y}}\right|_{\tilde{y}=\tilde{y}_1}}_{<0} - \underbrace{\left.\frac{\partial F}{\partial \tilde{y}}\right|_{\tilde{y}=\tilde{y}_1}}_{>0} < 0.\end{aligned}$$

Somit folgt die Behauptung. □

Lemma 2.8 gilt analog für kompakte Träger. Eine zu Lemma 2.8 ähnliche Aussage ist auch für streng fallend-steigende Dichten möglich.

LEMMA 2.9. *Sei $P_{\boldsymbol{\theta}}$ eine stetige Verteilung mit zugehöriger Verteilungsfunktion $F(\cdot, \boldsymbol{\theta})$ und streng fallend-steigender Dichte $f(\cdot, \boldsymbol{\theta})$ auf einem beschränkten Träger $supp(P_{\boldsymbol{\theta}}) = [a_{P_{\boldsymbol{\theta}}}, b_{P_{\boldsymbol{\theta}}}]$, wobei $f(a_{P_{\boldsymbol{\theta}}}, \boldsymbol{\theta}) \leq f(b_{P_{\boldsymbol{\theta}}}, \boldsymbol{\theta})$. Sei $y_{\min} := \operatorname{argmin}_{y \in ]a_{P_{\boldsymbol{\theta}}}, b_{P_{\boldsymbol{\theta}}}[} f(y, \boldsymbol{\theta})$ der Antimodus der Verteilung $P_{\boldsymbol{\theta}}$. Des Weiteren sei $\tilde{y}_3$ der Wert in $]y_{\min}, b_{P_{\boldsymbol{\theta}}}]$, für den $f(\tilde{y}_3) = f(a_{P_{\boldsymbol{\theta}}})$ gilt. Außerdem bezeichne $\alpha_{\max} := P(\mathcal{Y} \in [a_{P_{\boldsymbol{\theta}}}, \tilde{y}_3])$. Dann existiert für jedes $\alpha \in ]0, \alpha_{\max}]$ eine $\alpha$-Ausreißerregion bezüglich $P_{\boldsymbol{\theta}}$, die definiert ist durch*

$$\begin{aligned}\text{out}(\alpha, P_{\boldsymbol{\theta}}) &= \{y \in supp(P_{\boldsymbol{\theta}}) \cap ]y_1, y_2[\,\}, \qquad \textit{mit} \\ \alpha &= F(y_2, \boldsymbol{\theta}) - F(y_1, \boldsymbol{\theta}) \qquad \textit{und} \\ f(y_1, \boldsymbol{\theta}) &= f(y_2, \boldsymbol{\theta}).\end{aligned}$$

*Beweis:* Der Beweis ist in großen Teilen analog zum Beweis von Lemma 2.8. Wegen Stetigkeit und streng fallend-steigender Form von $f$ gilt: Für jedes $\tilde{y}_1 \in [a_{P_{\boldsymbol{\theta}}}, y_{\min}[$ existiert genau ein $\tilde{y}_2 \in ]y_{\min}, b_{P_{\boldsymbol{\theta}}}]$, so dass $f(\tilde{y}_1, \boldsymbol{\theta}) = f(\tilde{y}_2, \boldsymbol{\theta})$. Wie im Beweis von Lemma 2.8 kann eine Funktion $g : [a_{P_{\boldsymbol{\theta}}}, y_{\min}[ \to ]y_{\min}, \tilde{y}_3]$ definiert werden: $g(\tilde{y}_1, \boldsymbol{\theta}) = f_{\text{stei}}^{-1}(f(\tilde{y}_1, \boldsymbol{\theta}), \boldsymbol{\theta})$, wobei $f_{\text{stei}}^{-1}$ die stetige Umkehrfunktion der invertierbaren Funktion $f_{\text{stei}} : ]y_{\min}, \tilde{y}_3] \to ]f(y_{\min}), f(\tilde{y}_3)]$ mit $f_{\text{stei}}(y, \boldsymbol{\theta}) = f(y, \boldsymbol{\theta})$ ist. Es

kann analog argumentiert werden, dass $g$ stetig ist. Weiterhin ist die Hilfsfunktion $h(\tilde{y}_1, \boldsymbol{\theta}) := F(g(\tilde{y}_1, \boldsymbol{\theta}), \boldsymbol{\theta}) - F(\tilde{y}_1, \boldsymbol{\theta}) = \alpha$ analog zum Beweis von Lemma 2.8 bijektiv. Somit folgt die Behauptung. □

Das obige Lemma gilt analog für $f(a_{P_{\boldsymbol{\theta}}}, \boldsymbol{\theta}) \geq f(b_{P_{\boldsymbol{\theta}}}, \boldsymbol{\theta})$. Wird in der Situation von Lemma 2.9 $\alpha > \alpha_{\max}$ gewählt, ist die $\alpha$-Ausreißerregion für $f(a_{P_{\boldsymbol{\theta}}}) \leq f(b_{P_{\boldsymbol{\theta}}})$ identisch zu $\{y \in \text{supp}(P_{\boldsymbol{\theta}}) \cap [a_{P_{\boldsymbol{\theta}}}, q_\alpha]\}$. Gilt andererseits $f(a_{P_{\boldsymbol{\theta}}}) \geq f(b_{P_{\boldsymbol{\theta}}})$ und $\alpha > \alpha_{\max}$, so folgt $\text{out}(\alpha, P_{\boldsymbol{\theta}}) = \{y \in \text{supp}(P_{\boldsymbol{\theta}}) \cap [q_{1-\alpha}, b_{P_{\boldsymbol{\theta}}}]\}$. Wenn die Wahrscheinlichkeitsdichte nicht nur fallend-steigend und stetig sondern auch symmetrisch ist, kann ebenfalls auf Quantile für die Bestimmung der $\alpha$-Ausreißerregion zurückgegriffen werden.

KOROLLAR 2.10. *Sei $f(\cdot, \boldsymbol{\theta})$ die symmetrische, streng fallend-steigende Dichte von $P_{\boldsymbol{\theta}}$ mit Träger $[a_{P_{\boldsymbol{\theta}}}, b_{P_{\boldsymbol{\theta}}}]$ und $q_\alpha$ das zugehörige $\alpha$-Quantil. Dann gilt:*

$$\text{out}(\alpha, P_{\boldsymbol{\theta}}) = \{y \in \mathit{supp}(P_{\boldsymbol{\theta}}) \cap ]q_{\frac{1-\alpha}{2}}, q_{\frac{1+\alpha}{2}}[\,\}.$$

*Beweis:* Folgt aus Lemma 2.9 mit $y_1 = q_{\frac{1-\alpha}{2}}, y_2 = q_{\frac{1+\alpha}{2}}, F(y_2, \boldsymbol{\theta}) = \frac{1+\alpha}{2}, F(y_1, \boldsymbol{\theta}) = \frac{1-\alpha}{2}$ und $f(y_1, \boldsymbol{\theta}) = f(y_2, \boldsymbol{\theta})$ wegen der Symmetrie um den Median. □

Mittels Korollar 2.10 lässt sich beispielsweise die $\alpha$-Ausreißerregion der Beta$(\frac{1}{2}, \frac{1}{2})$-Verteilung herleiten.

Wie die vorigen Aussagen zeigen, ist die $\alpha$-Ausreißerregion für viele univariate Verteilungen leicht herzuleiten. Dies trifft auch für die multivariate Normalverteilung zu. Bei einer anderen multivariaten Verteilung ist die Herleitung komplizierter: Die Multinomialverteilung hat einen extrem großen Träger. Das $n$-fache Ziehen $k$ verschiedener Kugeln mit Zurücklegen ergibt $\binom{n+k-1}{n}$ Möglichkeiten. So hat der Träger der $Mult(n, \boldsymbol{\theta})$-Verteilung mit $n = 100$ und $k = 10$ bereits näherungsweise $4.263 \cdot 10^{12}$ Elemente. Zur Bestimmung der $\alpha$-Ausreißerregion auf herkömmliche Art werden alle Punktwahrscheinlichkeiten des Trägers auf- oder absteigend sortiert und so lange kumuliert, bis die zugehörige Summe den Wert $\alpha$ bzw. $1-\alpha$ erreicht. Die zugehörigen Elemente des Trägers werden dann als $\alpha$-Ausreißerregion bzw. $\alpha$-Inlierregion bezeichnet. Für das obige Beispiel $n = 100, k = 10$ ist dieses Vorgehen nicht mehr handhabbar. Eine etwas weniger rechenaufwändige Prozedur ist die Verwendung eines Suchalgorithmus, der nicht alle Elemente des Trägers auswertet, sondern nur

die (sukzessiv hinzugefügten) Elemente der $\alpha$-Inlierregion. Dabei wird in einem ersten Schritt die Inlierregion mit dem Modus der Multinomialverteilung initialisiert. Es sei angemerkt, dass der Modus einer Multinomialverteilung nicht eindeutig sein muss, siehe Finucan (1964). Weiterhin beschreibt Finucan einen Algorithmus zur Bestimmung des Modus, siehe Bemerkung 2.11.

BEMERKUNG 2.11. *Der Modus eines $Mult(n, \boldsymbol{\theta})$-verteilten Zufallsvektors $\boldsymbol{\mathcal{Y}}$ ist nicht in geschlossener Form darstellbar. Ein naiver Versuch wie $\lfloor \boldsymbol{\theta} n + 0.5 \rfloor$ erfüllt häufig nicht die Voraussetzung $\sum_{i=1}^{k} \lfloor \theta_i n + 0.5 \rfloor = n$. Stattdessen gilt: Falls $\sum_{i=1}^{k} \lfloor \theta_i (n + k/2) \rfloor = n$, ist $\lfloor \boldsymbol{\theta}(n + k/2) \rfloor$ der Modus der $Mult(n, \boldsymbol{\theta})$-Verteilung. Ist nun $\sum_{i=1}^{k} \lfloor \theta_i (n + k/2) \rfloor = n - 1$, wird der Eintrag von $\boldsymbol{\theta}(n + k/2)$, für den vor der Anwendung der Gaußklammer der Abstand zur nächstgrößeren ganzen Zahl proportional am kleinsten ist, um 1 erhöht. Für $\sum_{i=1}^{k} \lfloor \theta_i (n + k/2) \rfloor = n + 1$ existiert ein ähnliches Vorgehen.*

Zum Modus wird sukzessive das Element mit der höchsten Wahrscheinlichkeit aus der Nachbarschaft der bisherigen Inlierregion zur aktualisierten Inlierregion hinzugefügt. Algorithmus 1 skizziert das vorgeschlagene Vorgehen in Pseudo-Code.

**Algorithmus 1:** Suche der exakten $\alpha$-Inlierregion bei Multinomialverteilung

1 $\mathbf{y}_{\text{mod}} \leftarrow$ Modus;
2 Inlier $\leftarrow \{\mathbf{y}_{\text{mod}}\}$;
3 $\boldsymbol{\mathcal{Y}} \leftarrow$ Mult($n, \boldsymbol{\theta}$)-verteilte Zufallsvariable, wobei $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$;
4 **while** $P(\boldsymbol{\mathcal{Y}} \in \text{Inlier}) < 1 - \alpha$ **do**
5 $\quad x_{j,i} \leftarrow i$-ter Eintrag des $j$-ten Elements von Inlier;
6 $\quad$ Nachbarschaft(Inlier) $\leftarrow \{\mathbf{y}^{\text{cand}} \in \mathbb{N}_0^k | \sum_{i=1}^{k} y_i^{\text{cand}} = n \wedge \exists j : \max_i |y_i^{\text{cand}} - x_{j,i}| = 1\}$;
7 $\quad \mathbf{y}^{\text{next}} \leftarrow \text{argmax}_{\mathbf{y}^{\text{cand}}} P(\boldsymbol{\mathcal{Y}} = \mathbf{y}^{\text{cand}} | \mathbf{y}^{\text{cand}} \in \text{Nachbarschaft(Inlier)} \backslash \text{Inlier})$;
8 $\quad$ Inlier $\leftarrow \{\text{Inlier}, \mathbf{y}^{\text{next}}\}$
9 **end**

Häufige Bindungen gleicher Punktwahrscheinlichkeiten sind typisch für die Multinomialverteilung. Falls in Zeile 7 von Alg. 1 Bindungen auftreten, vergrößert sich die Inlierregion um alle Elemente des Trägers, die an der Bindung beteiligt sind.

Als approximative Alternative zur Bestimmung einer $\alpha$-Ausreißerregion bei gegebener Multinomialverteilung für *beliebig* große $k$ wird das folgende Vorgehen vorgeschlagen: Für $\boldsymbol{\mathcal{Y}} \sim Mult(n, \boldsymbol{\theta})$ gilt: Die Chi-Quadrat-Teststatistik $X^2(\boldsymbol{\mathcal{Y}}) =$

$\sum_{j=1}^{k} \frac{(y_j - n\theta_j)^2}{n\theta_j}$ ist als Summe von $k-1$ unabhängigen, quadrierten asymptotisch standardnormalverteilten Zufallsvariablen asymptotisch $\chi^2_{k-1}$-verteilt. Die $\alpha$-Ausreißerregion dieser transformierten Zufallsvariable lässt sich bei großem $n$ wie folgt approximieren: $\text{out}(\alpha, P_{\chi^2_{k-1}}) \approx ]\chi^2_{k-1,1-\alpha}, \infty[$ , wobei $\chi^2_{k-1,1-\alpha}$ das $(1-\alpha)$-Quantil der $\chi^2_{k-1}$-Verteilung ist. Es existieren verschiedene Faustregeln, ab wann diese Approximation hinreichend gut ist, siehe Büning und Trenkler (1994, S. 79). Zumeist wird jedoch eine gewisse Mindestgröße für alle erwarteten Zellhäufigkeiten gefordert, etwa $n\theta_j \geq 5$ oder $n\theta_j \geq 10$ für alle $j$. In Tabelle 2.1 werden die Ergebnisse einiger Simulationen gezeigt, in denen alle Elemente des Trägers verschiedener Multinomialverteilungen mit der exakten Methode ($B(\alpha)$ basierend auf $Mult(n, \boldsymbol{\theta})$) und der approximativen Methode ($B(\alpha)$ basierend auf $\chi^2_{k-1}$) klassifiziert wurden. Hier gelte $\alpha = 0.05$.

Tabelle 2.1: Approximation der $\alpha$-Ausreißerregionen multinomialverteilter Zufallsvariablen über $\chi^2$-Verteilung. RPR (Richtig-Positiv-Rate): Anteil der von der approximierenden Methode richtig klassifizierter Ausreißer, RNR (Richtig-Negativ-Rate): Anteil der von der approximierenden Methode richtig klassifizierter Nicht-Ausreißer.

| | | *Ausreißer* | | *Nicht-Ausreißer* | |
|---|---|---|---|---|---|
| $n$ | $\boldsymbol{\theta}^\top$ | Anzahl | RPR | Anzahl | RNR |
| 10 | $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ | 30 | 0.800 | 36 | 0.833 |
| 20 | $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ | 159 | 1.000 | 72 | 0.958 |
| 30 | $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ | 387 | 0.984 | 109 | 0.945 |
| 60 | $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ | 1674 | 1.000 | 217 | 0.972 |
| 90 | $\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ | 3861 | 1.000 | 325 | 0.963 |
| 90 | $\left(\frac{1}{6}, \frac{1}{3}, \frac{1}{2}\right)$ | 3905 | 0.999 | 281 | 0.979 |
| 90 | $\left(\frac{1}{9}, \frac{1}{9}, \frac{7}{9}\right)$ | 4023 | 0.998 | 163 | 0.975 |
| 40 | $\left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ | 10932 | 0.990 | 1409 | 0.974 |
| 50 | $(0.3, 0.2, 0.23, 0.27)$ | 21511 | 0.994 | 1915 | 0.963 |
| 50 | $(0.37, 0.2, 0.21, 0.22)$ | 21595 | 0.994 | 1831 | 0.962 |
| 60 | $\left(\frac{1}{6}, \frac{1}{6}, \frac{1}{3}, \frac{1}{3}\right)$ | 37416 | 0.996 | 2295 | 0.966 |
| 50 | $\left(\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right)$ | 297355 | 0.993 | 18896 | 0.961 |
| 60 | $(0.17, 0.17, 0.17, 0.17, 0.32)$ | 610517 | 0.996 | 24859 | 0.957 |

In Tabelle 2.1 zeigt sich, dass die $\alpha$-Ausreißerregionen der Multinomialverteilung und der zugehörigen $\chi^2_{k-1}$-Verteilung bei $n\theta_j \geq 10$ sehr ähnlich sind. Für die erste Zeile sind beide Faustregeln nicht erfüllt, die Approximation liefert entsprechend

schlechtere Werte. Die zweite Situation erfüllt nur $n\theta_j \geq 5$, liefert aber eine akzeptable Approximation. Die Anzahl der Beobachtungen $n$ und Klassen $k$ sowie die Klassenwahrscheinlichkeiten scheinen keine große Rolle zu spielen, die Ergebnisse sind recht einheitlich. Ausreißer werden sehr genau klassifiziert, Nicht-Ausreißer etwas weniger genau. Der hohe Speicherbedarf der Multinomialverteilungen erschwert einen Vergleich für $k > 5$. Für die Bestimmung der exakten Ausreißerregion in der letzten Zeile der Tabelle sind über 30 Gigabyte Speicherbedarf notwendig.

Die bisher behandelten Verteilungen, für die das Konzept der $\alpha$-Ausreißer definiert wurde, gehören zu den bekanntesten Verteilungen und zeichnen sich durch wenige Parameter aus. Eine flexiblere stetige Verteilung wird durch die $g$ und $h$-Verteilung (Tukey, 1977, $GH$-Verteilung) definiert. Sie eignet sich insbesondere, falls eine Normalverteilung oder andere klassische Verteilung nicht gerechtfertigt werden kann oder eine Online-Ausreißeranalyse für eine große Klasse stetiger Verteilungen vorbereitet werden soll. Eine $GH$-verteilte Zufallsvariable $\mathcal{Y}$ entsteht durch Transformation einer standardnormalverteilten Zufallsvariable $\mathcal{Z}$:

$$\mathcal{Y} = \mu + \sigma(G_g(\mathcal{Z})H_h(\mathcal{Z})\mathcal{Z}),$$

mit $G_g(\mathcal{Z}) = \frac{\exp(g\mathcal{Z})-1}{g\mathcal{Z}}$ und $H_h(\mathcal{Z}) = \exp(h\mathcal{Z}^2/2)$, wobei $g \in \mathbb{R}$ die Schiefe und $h \in \mathbb{R}_0^+$ die Schwere der Ränder parametrisiert (Hoaglin, 2006). Dabei ist $G_g(\mathcal{Z})$ wegen $\exp(g\mathcal{Z}) = 1 + \sum_{i=1}^{\infty} \frac{(g\mathcal{Z})^i}{i!}$ auch für $g = 0$ definiert und es folgt $G_0(\mathcal{Z}) = 1$. Weiterhin ist $\mu \in \mathbb{R}$ der Median der Verteilung und $\sigma \in \mathbb{R}^+$ ein Skalenparameter. Die streng steigend-fallende Dichte dieser Verteilung ist weder notwendigerweise symmetrisch, noch in geschlossener Form darstellbar. Xu *et al.* (2014) geben einen Überblick über diverse Verfahren zur Schätzung der Parameter der $GH(\mu, \sigma, g, h)$-Verteilung, die jedoch nicht robust und damit zur Ausreißeranalyse aufgrund von *Masking* und *Swamping* (siehe z. B. Def. 4.2) ungeeignet sind. Xu *et al.* entwickeln auf Basis des ursprünglichen Schätzverfahrens einen sequentiellen Algorithmus, der über die Verwendung von getrimmten Ordnungsstatistiken einen robustifizierten Ansatz zur Schätzung der vier Parameter liefert. Ein Spezialfall dieser Verteilung ist die Lognormalverteilung (für $h = 0$). Für $g = 0$ erhält man eine Verteilung mit gaußförmiger Dichte. Die $t$-Verteilung lässt sich nicht über die $GH$-Verteilung darstellen.

Um die $\alpha$-Ausreißerregion einer $GH$-Verteilung zu bestimmen, wird eine Alternative zu Definition 2.2 formuliert. Dazu schreiben wir zunächst Lemma 2.8 in eine von der Dichte unabhängige Version um, die stattdessen auf der Quantilfunktion einer Verteilung beruht: Seien $p \in ]0,1[$ und $q_p \in \text{supp}(P_{\boldsymbol{\theta}})$. Dann bezeichne $q_p$ das $p$-Quantil der Verteilung $P_{\boldsymbol{\theta}}$, wenn $F^{-1}(p, \boldsymbol{\theta}) = q_p$ gilt. Das folgende Lemma wird für die Herleitung der $\alpha$-Ausreißerregion der $GH$-Verteilung benötigt.

LEMMA 2.12. *Sei $\mathfrak{T}$ der Träger der streng steigend-fallenden Dichte $f(\cdot, \boldsymbol{\theta})$ mit zugehöriger Verteilungsfunktion $F(\cdot, \boldsymbol{\theta})$ und eindeutigem Modus $x_{\text{mod}}$. Dann ist die Ableitung der zugehörigen Quantilfunktion $F^{-1}(\cdot, \boldsymbol{\theta})$ streng fallend-steigend.*

*Beweis:* Sei $\mathfrak{T} = \mathfrak{T}_1 \cup \mathfrak{T}_2 := (\mathfrak{T} \cap ]-\infty, x_{\text{mod}}]) \cup (\mathfrak{T} \cap ]x_{\text{mod}}, \infty[)$ eine Zerlegung des Trägers. Wir betrachten die Teilfunktionen $f_1 : \mathfrak{T}_1 \to \mathbb{R}^+$ und $f_2 : \mathfrak{T}_2 \to \mathbb{R}^+$, so dass $f(x, \boldsymbol{\theta}) =: f_1(x) \; \forall \; x \in \mathfrak{T}_1$ und $f(x, \boldsymbol{\theta}) =: f_2(x) \; \forall \; x \in \mathfrak{T}_2$. Analog seien $F_1$ und $F_2$ als Teilfunktionen der Verteilungsfunktion $F(\cdot, \boldsymbol{\theta})$ sowie $F_1^{-1}, F_2^{-1}$ als Teilfunktionen der Quantilfunktion $F^{-1}(\cdot, \boldsymbol{\theta})$ definiert. Somit ergibt sich die Fallunterscheidung:

1. Seien $x_a, x_b \in \mathfrak{T}_1$. Die Funktion $f_1$ ist streng monoton steigend. Da $F_1$ als Stammfunktion von $f_1$ differenzierbar ist, gilt nach Kaballo (2000, Folgerung 21.3), dass $F_1$ streng konvex ist. Es gilt also für $x_a < x_b$: $F_1(tx_a + (1-t)x_b) < tF_1(x_a) + (1-t)F_1(x_b)$ für alle $t \in [0,1]$. Da $F_1^{-1}$ streng monoton steigend ist, folgt aus der strengen Konvexität

$$\begin{aligned} F_1^{-1}(F_1(tx_a + (1-t)x_b)) &< F_1^{-1}(tF_1(x_a) + (1-t)F_1(x_b)) \\ \Leftrightarrow tx_a + (1-t)x_b &< F_1^{-1}(tF_1(x_a) + (1-t)F_1(x_b)). \end{aligned}$$

   Mit $F_1(x_a) = q_a \Leftrightarrow x_a = F_1^{-1}(q_a)$ und äquivalent definiertem $q_b$ folgt

$$\begin{aligned} & tF_1^{-1}(q_a) + (1-t)F_1^{-1}(q_b) < F_1^{-1}(tF_1(F_1^{-1}(q_a)) + (1-t)F_1(F_1^{-1}(q_b))) \\ \Leftrightarrow\; & tF_1^{-1}(q_a) + (1-t)F_1^{-1}(q_b) < F_1^{-1}(tq_a + (1-t)q_b). \end{aligned}$$

   Somit ist $F_1^{-1}$ streng konkav. Mit Hilfe von Kaballo (2000, Folgerung 21.3) lässt sich schließen, dass die Ableitung von $F_1^{-1}$ streng monoton fallend ist.

2. Für $x \in \mathfrak{T}_2$ lässt sich analog zum ersten Fall zeigen, dass die Ableitung von $F_2^{-1}$ streng monoton steigend ist. Hierfür müssen lediglich die Rollen von „steigend“ und „fallend“ sowie „konkav“ und „konvex“ vertauscht werden.

Zusammen folgt, dass die Ableitung von $F^{-1}(\cdot, \boldsymbol{\theta})$ streng fallend-steigend ist. □

KOROLLAR 2.13. *Sei $P_{\boldsymbol{\theta}}$ eine Verteilung mit streng steigend-fallender Dichte auf $\mathbb{R}$ und Quantilfunktion $F^{-1}(\cdot, \boldsymbol{\theta})$. Daher folgt aus Lemma 2.8 für die $\alpha$-Ausreißerregion bezüglich $P_{\boldsymbol{\theta}}$ wegen $F^{-1}(p, \boldsymbol{\theta}) = q_p$:*

$$\operatorname{out}(\alpha, P_{\boldsymbol{\theta}}) = \{y \in \mathit{supp}(P_{\boldsymbol{\theta}}) \setminus [F^{-1}(p_1, \boldsymbol{\theta}), F^{-1}(p_2, \boldsymbol{\theta})]\} \tag{2.6}$$

$$\mathit{mit} \quad 1 - \alpha = F(F^{-1}(p_2, \boldsymbol{\theta}), \boldsymbol{\theta}) - F(F^{-1}(p_1, \boldsymbol{\theta}), \boldsymbol{\theta}) = p_2 - p_1 \tag{2.7}$$

$$\begin{aligned} \mathit{und} \quad & f(F^{-1}(p_1, \boldsymbol{\theta}), \boldsymbol{\theta}) = f(F^{-1}(p_2, \boldsymbol{\theta}), \boldsymbol{\theta}) \\ \Leftrightarrow \quad & \frac{1}{f(F^{-1}(p_1, \boldsymbol{\theta}), \boldsymbol{\theta})} = \frac{1}{f(F^{-1}(p_2, \boldsymbol{\theta}), \boldsymbol{\theta})} \; \forall \, p_1, p_2 \in ]0, 1[. \end{aligned} \tag{2.8}$$

*Wegen der Umkehrregel (Königsberger, 2004, S. 143) ist (2.8) äquivalent zu*

$$\begin{aligned} \left.\frac{\partial F^{-1}(\cdot, \boldsymbol{\theta})}{\partial p}\right|_{p=p_1} &= \left.\frac{\partial F^{-1}(\cdot, \boldsymbol{\theta})}{\partial p}\right|_{p=p_2} \\ &\overset{(2.7)}{=} \left.\frac{\partial F^{-1}(\cdot, \boldsymbol{\theta})}{\partial p}\right|_{p=1-\alpha+p_1}. \end{aligned} \tag{2.9}$$

*Da die Ableitung der Quantilfunktion stetig (Kaballo, 2000, Satz 19.8) und nach Lemma 2.12 streng fallend-steigend ist, existiert genau ein $p_1 \in ]0, \alpha[$, so dass* (2.9) *gilt. $p_1$ kann äquivalent als Minimalstelle der Funktion*

$$\rho(\tilde{p}; \alpha, \boldsymbol{\theta}) := \left( \left.\frac{\partial F^{-1}(\cdot, \boldsymbol{\theta})}{\partial p}\right|_{p=\tilde{p}} - \left.\frac{\partial F^{-1}(\cdot, \boldsymbol{\theta})}{\partial p}\right|_{p=1-\alpha+\tilde{p}} \right)^2$$

*aufgefasst werden, also $p_1 = \operatorname{argmin}_{\tilde{p}} \rho(\tilde{p}; \alpha, \boldsymbol{\theta})$.*

Diese Minimierung ist leicht numerisch durchzuführen, etwa mit dem Brent-Verfahren (Brent, 1973, Kapitel 4). Mit Hilfe des R-Pakets `alphaOutlier` (Rehage und Kuhnt, 2016) kann die $\alpha$-Ausreißerregion einer bekannten $GH$-Verteilung bestimmt werden. Abbildung 2.3 zeigt das Zusammenwirken von Quantilfunktion, er-

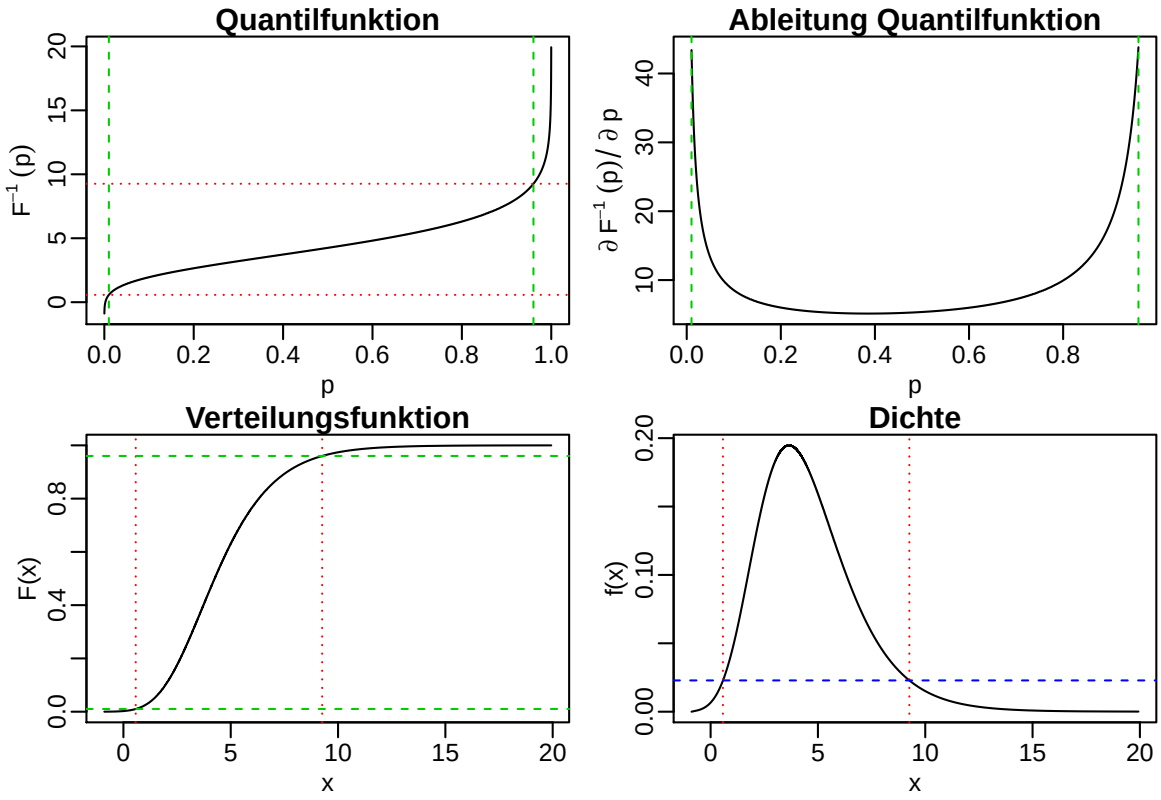



Abbildung 2.3: Grenzen der $0.05$-Ausreißerregion (rot) der $GH$-Verteilung mit $\mu = 4.25, \sigma = 2.14, g = 0.3, h = 0.01$. Grün: $p_1, p_2$. Blau: $B(0.05)$.

ster Ableitung der Quantilfunktion, Verteilungsfunktion und Wahrscheinlichkeitsdichte einer $GH$-verteilten Zufallsvariable. Die Stellen, in denen die Quantilfunktion (oben links) die Grenzen der $0.05$-Inlierregion (rot punktiert) schneidet, sind die Werte für $p_1$ und $p_2$ (grün gestrichelt) aus (2.6). An den Schnittpunkten mit der Ableitung der Quantilfunktion (oben rechts) lässt sich erahnen, dass die Ableitungen in $p_1$ und $p_2$ tatsächlich gleich sind. Des Weiteren sind die Verteilungsfunktion (unten links) und die Dichte (unten rechts) mit den Grenzen der $0.05$-Inlierregion abgebildet. Für die Dichte wird zudem veranschaulicht, dass $f(F^{-1}(p_1)) = f(F^{-1}(p_2)) = B(0.05)$ (blau gestrichelt) gilt.

Das Konzept der $\alpha$-Ausreißer wird im folgenden Abschnitt 2.2 auf unbekannte Verteilungssituationen verallgemeinert.

## 2.2 $\alpha$-AUSREISSER BEI VERWENDUNG VON KERNDICHTESCHÄTZERN

In der Situation von Definition 2.2 wird zur Bestimmung einer $\alpha$-Ausreißerregion eine bekannte Verteilung $P$ mit zugehöriger Dichte $f$ vorausgesetzt. Genau wie die Schätzung der Parameter von $P$ nichts an der Anwendbarkeit des $\alpha$-Ausreißerprinzips ändert, kann auch die Verteilung $P$ und damit ihre Dichte $f$ vollständig unbekannt sein. In solchen Fällen ist eine Schätzung von $f$ notwendig, weswegen der Begriff des Kerndichteschätzers hier kurz wiederholt wird.

DEFINITION 2.14. Kerndichteschätzer *(Büning und Trenkler, 1994, S. 260)*
*Seien $y_1, \ldots, y_k$ univariate Beobachtungen, die als Realisationen von $\mathcal{Y} \sim P$ mit stetiger Dichte $f$ aufgefasst werden. Dann ist $\hat{f}_k(y;h) = \frac{1}{kh}\sum_{i=1}^{k} K\left(\frac{y-y_i}{h}\right)$ ein* Kerndichteschätzer *für $f(y)$ mit Kern $K : \mathbb{R} \to \mathbb{R}$ und Bandbreite $h > 0$.*

In der Literatur beliebte Kerne sind etwa der Gaußkern

$$K_G(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right), \tag{2.10}$$

der Epanechnikov-Kern $K_E(y) = \frac{3}{4}(1-y^2)\mathbf{1}_{[-1,1]}(y)$ oder der Rechteckkern $K_R(y) = \frac{1}{2}\mathbf{1}_{[-1,1]}(y)$, jeweils mit $y \in \mathbb{R}$. Sie sind Beispiele für Kerne, deren Werte immer nichtnegativ sind und für die $\int_{\mathbb{R}} K(y)dy = 1$ gilt. Auf ihnen basierende Kerndichteschätzer erfüllen damit auch $\hat{f}_k(y;h) \geq 0 \; \forall \; y \in \mathbb{R}$ und $\int_{\mathbb{R}} \hat{f}_k(y;h)dy = 1$, also die Voraussetzungen an eine Wahrscheinlichkeitsdichte. Unter gewissen Annahmen ist ein Kerndichteschätzer asymptotisch erwartungstreu.

SATZ 2.15. *Gegeben sei ein Kerndichteschätzer $\hat{f}_k(y;h)$ mit Bandbreite $h = h(k) > 0$. Falls $\lim_{k\to\infty} h(k) = 0$, $\sup_y K(y) < \infty$, $\int_{\mathbb{R}} |K(y)|dy < \infty$, $\lim_{y\to\infty} |yK(y)| = 0$ und $\int_{\mathbb{R}} K(y)dy = 1$, gilt: Für alle $y \in \mathbb{R}$ ist $\hat{f}_k(y)$ ein asymptotisch erwartungstreuer Schätzer für $f(y)$.*

*Beweis:* Vergleiche Parzen (1962), Theorem 1A und Corollary 1A. □
Satz 2.15 rechtfertigt die Anwendung des $\alpha$-Ausreißerprinzips für Kerndichteschätzer in hinreichend großen Stichproben. Im R-Paket `alphaOutlier` kann die Funktion `aout.kernel` verwendet werden, um Kerndichteschätzer und dazu korre-

spondierende $\alpha$-Ausreißerregionen in einem Schritt zu erhalten, siehe auch Beispiel 2.16. Zur Wahl der Bandbreite empfehlen Venables und Ripley (2002) auf Basis einer Reihe von Vergleichsstudien die SJ-Schätzung $\hat{h}_{SJ}$ von $h$ (Sheather und Jones, 1991).

BEISPIEL 2.16. *Gegeben seien je $k_1 = 10$, $k_2 = 40$ und $k_3 = 1000$ Realisationen einer $P = \aleph(0,1)$-verteilten Zufallsvariable. Zusätzlich wurde in jedem Datensatz der erste Eintrag durch den $0.1$-Ausreißer $y_{out} = -2.5$ ersetzt. Abbildung 2.4 zeigt die drei daraus resultierenden Kerndichteschätzungen mit einem Gaußkern und $\hat{h}_{SJ}$. Die $0.1$-Ausreißerregionen sind schraffiert und die darin befindlichen Ausreißer rot markiert.*

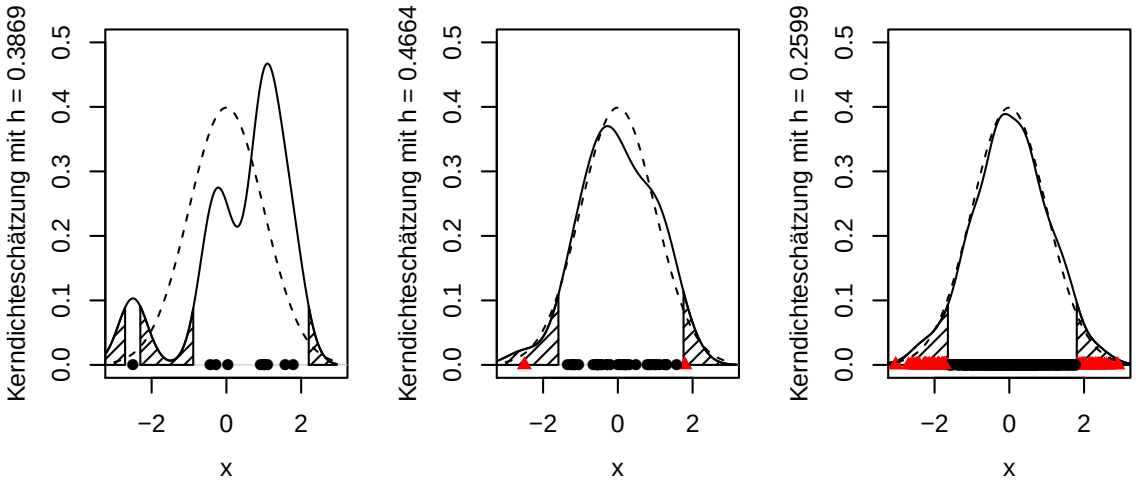



Abbildung 2.4: Kerndichteschätzer (durchgezogen) mit Dichte der $\aleph(0,1)$-Verteilung (gestrichelt) sowie schraffierter $0.1$-Ausreißerregion basierend auf 10 (links), 40 (mittig) und 1000 (rechts) Beobachtungen.

*Im linken Plot ist zu sehen, dass $y_{out}$ falsch – also als Nicht-Ausreißer – klassifiziert wurde. In den übrigen Plots wurde $y_{out}$ korrekt klassifiziert. Es ergeben sich die geschätzten $0.1$-Ausreißerregionen $\mathrm{out}(0.1, \hat{P}_{k_1}) = \mathbb{R}\backslash([-2.704, -2.303] \cup [-0.880, 2.204])$ (kein Ausreißer identifiziert), $\mathrm{out}(0.1, \hat{P}_{k_2}) = \mathbb{R}\backslash[-1.584, 1.758]$ (zwei Ausreißer identifiziert) und $\mathrm{out}(0.1, \hat{P}_{k_3}) = \mathbb{R}\backslash[-1.636, 1.805]$ (81 Ausreißer identifiziert). Die wahre $0.1$-Ausreißerregion lautet $\mathrm{out}(0.1, P) = \mathbb{R}\backslash[-1.645, 1.645]$. Für $k_2$ ist das Ergebnis deutlich näher an der wahren Ausreißerregion als für $k_1$. Der Unterschied zwischen der Ausreißerregion basierend auf 1000 Beobachtungen und der auf 40 Beobachtungen ist hingegen klein.*

In kleinen Stichproben kann die Verknüpfung von Kerndichteschätzern und $\alpha$-

Ausreißerregionen zu unerwünschten Ergebnissen führen: Für jedes $k \in \mathbb{N}$ und $h > 0$ existiert ein Kern und ein $\alpha \in ]0,1[$, so dass ein wahrer Ausreißer nicht als solcher identifiziert werden kann. Dies wird in Satz 2.17 gezeigt.

SATZ 2.17. *Sei* $\mathrm{out}(\alpha, P)$ *für* $\alpha \in ]0,1[$ *die* $\alpha$*-Ausreißerregion der wahren Verteilung* $P$ *und* $y_{out} \in \mathrm{out}(\alpha, P)$ *ein* $\alpha$*-Ausreißer. Sei* $\mathbf{y} = (y_1, \ldots, y_{k-1}, y_{out})^\top$ *eine Stichprobe aus der Verteilung* $P$ *mit* $y_i \notin \mathrm{out}(\alpha, P)\ \forall\ i \in \{1, \ldots, k-1\}$. *Sei* $\hat{f}_k(y; h)$ *die geschätzte Dichte von* $P$ *basierend auf* $\mathbf{y}$ *und einem Kern* $K_\delta$ *mit*

$$K_\delta(y) = \begin{cases} \tilde{K}(y), & y \in [-\delta, \delta], \\ 0, & \textit{sonst}, \end{cases}$$

*mit* $\tilde{K}(-y) = \tilde{K}(y) > 0\ \forall\ y \in [-\delta, \delta]$, $\tilde{K}(y_\ell) \geq \tilde{K}(y_r)\ \forall\ 0 \leq y_\ell < y_r$ *und* $\int_{-\delta}^{\delta} \tilde{K}(y)dy = 1$. *Weiterhin sei* $\hat{P}_k$ *die zu* $\hat{f}_k(y; h)$ *gehörende geschätzte Verteilung und* $\mathrm{out}(\alpha, \hat{P}_k)$ *die korrespondierende* $\alpha$*-Ausreißerregion. Dann gilt: Falls* $\alpha < \frac{1}{k}$ *und* $\delta^* := \min_{i=1,\ldots,k-1} \frac{|y_{out} - y_i|}{2h} > \delta$, *folgt* $y_{out} \notin \mathrm{out}(\alpha, \hat{P}_k)$.

*Beweis:* Um die Notation zu vereinfachen, bezeichne $y_k := y_{out}$. Es ist festzuhalten, dass $\hat{f}_k(\cdot; h)$ im Intervall $[y_{out} - \delta, y_{out} + \delta]$ einen symmetrischen, steigend-fallenden Verlauf nimmt. Dies liegt wegen $\min \frac{|y_{out} - y_i|}{h} > 2\delta$ an der Nicht-Überlappung der paarweisen Summanden des Kerndichteschätzers $K_\delta(\frac{y - y_{out}}{h})$ und $K_\delta(\frac{y - y_i}{h}), i = 1, \ldots, k-1$: $0 \leq K_\delta\left(\frac{y_{out} - y_i}{h}\right) = K_\delta\left(\frac{|y_{out} - y_i|}{h}\right) \leq K_\delta(2\delta^*) = 0$. Der Kerndichteschätzer erreicht innerhalb von $[y_{out} - \delta, y_{out} + \delta]$ sein Maximum in $y_{out}$:

$$\hat{f}_k(y_{out}; h) = \frac{1}{kh} \sum_{i=1}^{k} K_\delta\left(\frac{y_{out} - y_i}{h}\right) = \frac{1}{kh}\left(\tilde{K}(0) + \sum_{i=1}^{k-1} K_\delta\left(\frac{y_{out} - y_i}{h}\right)\right),$$

woraus wegen der Nicht-Überlappung $\hat{f}_k(y_{out}; h) = \frac{\tilde{K}(0)}{kh}$ folgt. Weiterhin gilt: $\int_{y_{out} - h\delta}^{y_{out} + h\delta} \hat{f}_k(y; h)dy = \frac{1}{k}$, da kein $y_i \in [y_{out} - 2h\delta, y_{out} + 2h\delta]$ für $i \in \{1, \ldots, k-1\}$. Dann ist $\alpha \geq \frac{1}{k}$ eine notwendige Bedingung dafür, dass $y_{out}$ als Ausreißer identifiziert werden kann. □

Im folgenden – pathologischen – Beispiel wird gezeigt, dass der Grenzfall $\alpha = \frac{1}{k}$ ausreichen kann, um einen Ausreißer zu identifizieren.

BEISPIEL 2.18. *Sei eine Stichprobe $y_1 = \ldots = y_{19} = 0, y_{out} = 3$ vom Umfang $k = 20$ gegeben. Mit Hilfe des Rechteckkerns und der Bandbreite $h = 1$ ergibt sich*

$$\hat{f}_{20}(y;1) = \begin{cases} \frac{19}{40}, & y \in [-1,1], \\ \frac{1}{40}, & y \in [2,4], \\ 0, & \textit{sonst.} \end{cases}$$

*Damit $y_{out}$ als $\alpha$-Ausreißer identifiziert wird, muss $B(\alpha) > \frac{1}{40}$ sein. Setze $B(\alpha) := \frac{1}{39}$, so folgt, dass der Anteil der als Ausreißer klassifizierten Beobachtungen gleich $\frac{1}{k}$ ist. Somit kann gezeigt werden, dass $\mathrm{out}(\alpha, \hat{P}) = \mathbb{R} \setminus [-1,1]$ die für dieses Beispiel geschätzte $\alpha$-Ausreißerregion ist. Da das Integral der geschätzten Dichte über die Ausreißerregion gleich $\frac{1}{k}$ ist, folgt $\alpha = \frac{1}{k}$. Andererseits wird klar, dass ein $\alpha < \frac{1}{k}$ nicht zur Identifikation eines $\alpha$-Ausreißers führen kann, da daraus $B(\alpha) \leq \frac{1}{40}$ folgt.*

Für kleine Stichproben ist die Verwendung eines robusteren Kerndichteschätzers erstrebenswert, der nicht alle Beobachtungen mit dem gleichen Gewicht in die Kerndichteschätzung einfließen lässt und somit die in Satz 2.17 beschriebene Situation verhindern soll. Für den Kerndichteschätzer von Kim und Scott (2012) liegen Ergebnisse bezüglich seiner Influenzfunktion vor, die von einer hohen Robustheit zeugen. Der Schätzer (siehe Def. 2.19) nutzt dabei Maximum-Likelihood-artige M-Schätzer. Diese Schätzerklasse unterliegt einem ähnlichen Optimierungskalkül wie ML-Schätzer, ist aber robuster bei Vorliegen von Ausreißern. Ein weiteres Konzept, das in der folgenden Definition genutzt wird, ist der Hilbertraum mit reproduzierendem Kern, vgl. Steinwart und Christmann (2008, Kap. 4.2).

DEFINITION 2.19. KS-Kerndichteschätzer
*Sei $\mathbf{y} = (y_1, \ldots, y_k)^\top$ eine Stichprobe, $\rho$ die robuste Verlustfunktion eines M-Schätzers mit $\rho(y) > 0$ und $\mathcal{H}$ ein Hilbertraum mit reproduzierendem Kern. Dann bezeichnet*

$$\hat{f}_k^{\mathrm{KS}}(y;h) = \operatorname*{argmin}_{g \in \mathcal{H}} \sum_{i=1}^{k} \rho\left( \left\| K\left( \frac{y - y_i}{h} \right) - g \right\| \right) \tag{2.11}$$

*den* KS-Kerndichteschätzer *zur Stichprobe* $\mathbf{y}$.

Nach Kim und Scott ist eine alternative Darstellung von (2.11) mit Hilfe passend zu wählender nichtnegativer Gewichte $w_1, \ldots, w_k$ mit $\sum_{i=1}^k w_i = 1$ möglich: $\hat{f}_k^{\mathrm{KS}}(y;h) = \frac{1}{kh}\sum_{i=1}^k w_i K\left(\frac{y-y_i}{h}\right)$. Des Weiteren nennen die Autoren als bekannte Beispiele für $\rho$ die Huber-Verlustfunktion

$$\rho_{\mathrm{Hu}}(y;a) = \begin{cases} \frac{y^2}{2}, & |y| \in [0,a] \\ a, & |y| \in ]a,\infty[ \end{cases}$$

und die Hampel-Verlustfunktion

$$\rho_{\mathrm{Ha}}(y;a,b,c) = \begin{cases} \frac{y^2}{2}, & |y| \in [0,a[ \\ ay - \frac{a^2}{2}, & |y| \in [a,b[ \\ \frac{a(y-c)^2}{2(b-c)} + \frac{a(b+c-a)}{2}, & |y| \in [b,c[ \\ \frac{a(b+c-a)}{2}, & |y| \in [c,\infty[. \end{cases}$$

Die Implementierung des robusten Kerndichteschätzers in R erfolgte auf Basis von MATLAB-Code von Kim (2011), vergleiche Anhang A.1.

BEISPIEL 2.20. Fortsetzung von Beispiel 2.16

*In der Situation von Beispiel 2.16 wird nun der KS-Kerndichteschätzer verwendet. Hierfür wurde vorab die Bandbreite über eine Log-Likelihood Kreuzvalidierung passend gewählt, siehe R-Funktion* `bandwidth_select` *in Anhang A.1. Als robuste Verlustfunktion wurde* $\rho_{\mathrm{Hu}}$ *verwendet. Die Ergebnisse sind in Abbildung 2.5 visualisiert.*

*Im linken Plot ist zu sehen, dass* $x_{out}$ *nun korrekt klassifiziert wurde. Dies gilt auch für die übrigen Plots. Es ergeben sich die geschätzten* $0.1$*-Ausreißerregionen* $\mathrm{out}(0.1, \hat{P}_{k_1}) = \mathbb{R}\backslash[-0.822, 2.191]$ *(ein Ausreißer identifiziert),* $\mathrm{out}(0.1, \hat{P}_{k_2}) = \mathbb{R}\backslash[-1.514, 1.654]$ *(zwei Ausreißer identifiziert) und* $\mathrm{out}(0.1, \hat{P}_{k_3}) = \mathbb{R}\backslash[-1.585, 1.768]$ *(98 Ausreißer identifiziert). Die wahre* $0.1$*-Ausreißerregion beträgt* $\mathrm{out}(0.1, P) = \mathbb{R}\backslash[-1.645, 1.645]$*. Für* $k = k_2$ *ist das Ergebnis deutlich näher an der wahren Ausreißerregion als für* $k = k_1$ *und sogar etwas genauer als für* $k = k_3$*. Dies liegt nicht daran, dass der KS-Kerndichteschätzer für sehr großes* $k$ *generell weniger geeignet ist als für moderates* $k$*, sondern an der spezifischen, tatsächlich leicht rechtsschiefen Stichprobe.*

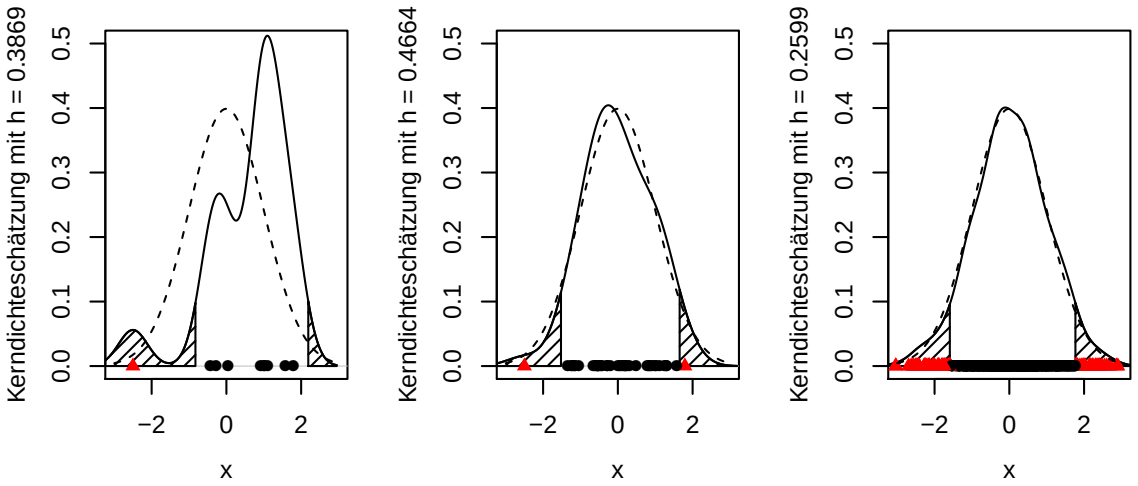



Abbildung 2.5: Robuster Kerndichteschätzer (durchgezogen) mit Dichte der $\aleph(0,1)$-Verteilung (gestrichelt) sowie schraffierter $0.1$-Ausreißerregion basierend auf 10 (links), 40 (mittig) und 1000 (rechts) Beobachtungen.

Beispiel 2.20 veranschaulicht die Nützlichkeit aus der Kombination von robuster Kerndichteschätzung und $\alpha$-Ausreißern für kleine und mittlere Stichproben. Andere Ausreißeridentifikationsverfahren und ihre Verknüpfungen zum Konzept der $\alpha$-Ausreißer werden in Abschnitt 2.3 thematisiert.

## 2.3 $\alpha$-AUSREISSER IM VERGLEICH MIT BOXPLOTS UND BAGPLOTS

Eine einfache, deskriptive Methode zur Ausreißererkennung in univariaten Daten ist der Boxplot von Tukey (1977). Der Boxplot besteht im einfachen Fall aus fünf Kennzahlen: dem Minimum $y_{(1)}$, dem unteren Quartil $y_{0.25}$, dem Median $y_{0.5}$, dem oberen Quartil $y_{0.75}$ und dem Maximum $y_{(k)}$. Als modifizierte Version des Boxplots sei diejenige bezeichnet, bei der die *whiskers* der Box statt bis $y_{(1)}$ und $y_{(k)}$ nur bis $\ell_{\text{Tukey}} = \min\{y_i | y_i \geq y_{0.25} - 1.5\,\text{IQR}\}$ bzw. $u_{\text{Tukey}} = \max\{y_i | y_i \leq y_{0.75} + 1.5\,\text{IQR}\}$ eingezeichnet werden, wobei $\text{IQR} = y_{0.75} - y_{0.25}$ der Interquartilsabstand ist. Alle Beobachtungen außerhalb von $[\ell_{\text{Tukey}}, u_{\text{Tukey}}]$ werden als Ausreißer bezeichnet. Tukeys Boxplot ist ein deskriptives Verfahren zur Visualisierung von univariaten, metrisch skalierten Merkmalen. Für Merkmale mit sehr wenigen möglichen Ausprägungen ist der Boxplot überdimensioniert; ein gestapeltes Balkendiagramm liefert jedoch

eine optisch ähnlich leicht verständliche Visualisierung. Für die Bestimmung von sinnvollen Ausreißern mit Hilfe des Boxplots sollte die zugrunde liegende Verteilung symmetrisch und die Dichte steigend-fallend sein. Für asymmetrisch verteilte Daten ist der Boxplot weniger gut geeignet, hier sollte der adjustierte Boxplot (Hubert und Vandervieren, 2008) präferiert werden. Weiterhin offenbart Tukeys Boxplot für schwerrändrige Verteilungen ein unerwünschtes Verhalten, indem er zu viele Ausreißer anzeigt (siehe Tabelle 2.2, Standard-Cauchyverteilung). In solchen Fällen ist der generalisierte Boxplot von Bruffaerts *et al.* (2014) eine Alternative. Die Beobachtungen werden zunächst standardisiert und in das Intervall $]0, 1[$ transformiert. Über eine Probit-Transformation wird erreicht, dass die so erhaltenen Beobachtungen als Realisierungen einer transformierten Normalverteilung aufgefasst werden können. Mittels der $GH$-Verteilung und den in Bruffaerts *et al.* vorgestellten Schätzern können Quantile geschätzt werden, die die Grenzen der *whiskers* bestimmen.

Im Unterschied zu Tukeys Boxplot beruht das Konzept der $\alpha$-Ausreißer auf einer konkret parametrisierten Verteilungsannahme. Allerdings kann die Ausreißerregion eines modifizierten Boxplots auch bezüglich einer Verteilung formuliert werden:

LEMMA 2.21. Populationsversion des modifizierten Boxplots
*Sei $P_{\boldsymbol{\theta}}$ eine stetige Verteilung mit zugehöriger Quantilfunktion $F^{-1}(\cdot, \boldsymbol{\theta})$ und streng steigend-fallender, symmetrischer Dichte $f(\cdot, \boldsymbol{\theta})$ auf dem offenen oder abgeschlossenen Träger $supp(P_{\boldsymbol{\theta}})$. Dann ist die Ausreißerregion des modifizierten Boxplots gegeben durch*

$$\begin{aligned}\mathrm{outBoxplot}(1.5, P_{\boldsymbol{\theta}}) = supp(P_{\boldsymbol{\theta}}) \setminus [&2.5F^{-1}(0.25, \boldsymbol{\theta}) - 1.5F^{-1}(0.75, \boldsymbol{\theta}),\\ &2.5F^{-1}(0.75, \boldsymbol{\theta}) - 1.5F^{-1}(0.25, \boldsymbol{\theta})],\end{aligned} \tag{2.12}$$

*wobei* $1.5$ *den Faktor bezeichnet, der mit dem Interquartilsabstand multipliziert wird um die Länge der* whiskers *zu erhalten. Als Grenzen der Box fungieren $F^{-1}(0.25, \boldsymbol{\theta})$ und $F^{-1}(0.75, \boldsymbol{\theta})$.*

*Beweis:* Die Populationsversion des Interquartilsabstand lautet $\mathrm{IQR} = F^{-1}(0.75, \boldsymbol{\theta}) - F^{-1}(0.25, \boldsymbol{\theta})$. Dann folgen für die Grenzen der *whiskers* $\ell_{\text{Tukey}} = F^{-1}(0.25, \boldsymbol{\theta}) - 1.5\,\mathrm{IQR} = 2.5F^{-1}(0.25, \boldsymbol{\theta}) - 1.5F^{-1}(0.75, \boldsymbol{\theta})$ und $u_{\text{Tukey}} = F^{-1}(0.75, \boldsymbol{\theta}) + 1.5\,\mathrm{IQR} = 2.5F^{-1}(0.75, \boldsymbol{\theta}) - 1.5F^{-1}(0.25, \boldsymbol{\theta})$. □

Mit Hilfe von (2.12) lässt sich für jede beliebige Verteilung mit streng steigend-fallender, symmetrischer Dichte ein $\alpha$ herleiten, so dass $\mathrm{out}(\alpha, P_{\boldsymbol{\theta}}) = \mathrm{outBoxplot}(1.5, P_{\boldsymbol{\theta}})$ gilt: Die untere Grenze der $\alpha$-Inlierregion lautet $2.5F^{-1}(0.25, \boldsymbol{\theta}) - 1.5F^{-1}(0.75, \boldsymbol{\theta})$. Da die Dichte symmetrisch ist, soll genau $\frac{\alpha}{2}$ der Wahrscheinlichkeitsmasse unterhalb dieser Grenze liegen. Dies ist äquivalent zu

$$\begin{aligned} \frac{\alpha}{2} &= F(2.5F^{-1}(0.25, \boldsymbol{\theta}) - 1.5F^{-1}(0.75, \boldsymbol{\theta}), \boldsymbol{\theta}) \\ \Leftrightarrow \alpha &= 2F(2.5F^{-1}(0.25, \boldsymbol{\theta}) - 1.5F^{-1}(0.75, \boldsymbol{\theta}), \boldsymbol{\theta}). \end{aligned}$$

Alternativ kann $\alpha$ über die obere Grenze der *whiskers* hergeleitet werden und liefert dann $\alpha = 2 - 2F(2.5F^{-1}(0.75, \boldsymbol{\theta}) - 1.5F^{-1}(0.25, \boldsymbol{\theta}), \boldsymbol{\theta})$. Tabelle 2.2 zeigt einige Verteilungen mitsamt zugehöriger Werte von $\alpha$, bei denen die Ausreißerregionen übereinstimmen. Bei schwerrändrigen Verteilungen wie der $t$-Verteilung mit zwei Freiheitsgraden oder der Standard-Cauchyverteilung liegt der erwartete Anteil falsch klassifizierter Nicht-Ausreißer im Boxplot bei 8 bzw. 16%. Anders als Bruffaerts *et al.* (2014, Kapitel 2.1) behaupten, liegt $\alpha$ für die $t_2$-Verteilung nicht bei etwa $5\%$, sondern bei etwa $8\%$.

Tabelle 2.2: Kombinationen von $P_{\boldsymbol{\theta}}$ und $\alpha$, für die $\alpha$-Ausreißerregion und Boxplot-Ausreißerregion übereinstimmen.

| $P_{\boldsymbol{\theta}}$ | $\aleph(0,1)$ | $t_{10}$ | $t_2$ | Beta$(5,5)$ | Cauchy$(0,1)$ | Logistisch$(0,1)$ |
|---|---|---|---|---|---|---|
| $\alpha$ | $0.69766\%$ | $1.88188\%$ | $8.23371\%$ | $0.02874\%$ | $15.59583\%$ | $2.43902\%$ |

Für multivariate Datensituationen ist der Boxplot nicht geeignet. Bivariate Daten können jedoch mit dem verwandten Konzept des Bagplots (Rousseeuw *et al.*, 1999) visualisiert werden. In Abb. 2.6 ist ein mit der R-Funktion `bagplot` aus dem Paket `aplpack` (Wolf, 2014) erstellter Beispielplot zu sehen. Der rote Stern bezeichnet den Halbraummedian, die dunkelblaue Fläche den *bag* und die hellblaue Fläche wird vom *fence* begrenzt. Außerhalb liegende Datenpunkte (rot) werden als Ausreißer bezeichnet. Sinnvolle Einblicke in die Datenstruktur sind unter anderem davon abhängig, ob die zugrundeliegende Verteilung unimodal ist und ihre Dichte konvexe Konturen hat.

Der Bagplot ist eine bivariate grafische Methode zur Veranschaulichung gewisser

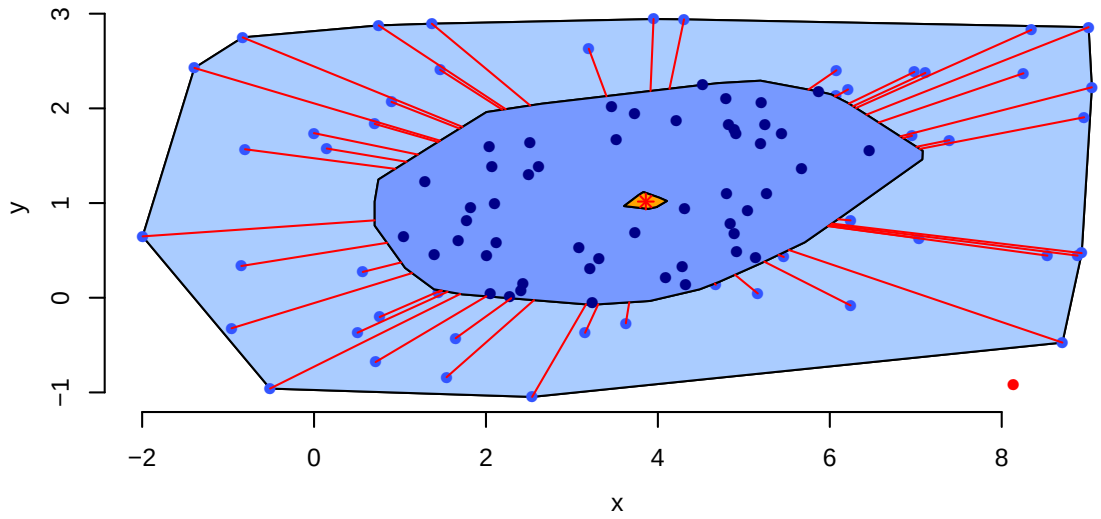



Abbildung 2.6: Schematische Darstellung eines Bagplots.

Lagemaße. Er basiert auf der von Tukey (1975) für univariate Daten formal und für bivariate Daten verbal eingeführten Halbraumtiefe (*halfspace depth*). Die Halbraumtiefe kann als multivariate Verallgemeinerung des Rangkonzepts aufgefasst werden. Formal ist die Stichprobenversion der $p$-dimensionalen Halbraumtiefe nach Donoho und Gasko (1992) für $\mathbf{y} \in \mathbb{R}^p$ bezüglich der Beobachtungen $\mathbf{y}_1, \ldots, \mathbf{y}_k \in \mathbb{R}^p$ definiert durch

$$d^{HS}(\mathbf{y}; \mathbf{y}_1, \ldots, \mathbf{y}_k) = k^{-1} \min_{\boldsymbol{u} \in \mathbb{R}^p, \|\boldsymbol{u}=1\|} \#\{i : \boldsymbol{u}^\top \mathbf{y}_i \geq \boldsymbol{u}^\top \mathbf{y}\}. \tag{2.13}$$

Anschaulich lässt sich dies an Abbildung 2.7 erklären. Die Halbraumtiefe soll – bezüglich der Beobachtungen in der linken Grafik – an der Stelle $\mathbf{y}^Q$ bestimmt werden, die durch ein schwarzes Quadrat gekennzeichnet ist. Dann werden alle Geraden betrachtet, die durch $\mathbf{y}^Q$ verlaufen. Ein Beispiel ist in der mittleren Grafik zu sehen, die Gerade wird mit $g_1^Q$ bezeichnet. Die Beobachtungen werden orthogonal auf $g_1^Q$ projiziert (rechte Grafik). Die Gerade kann in zwei durch $\mathbf{y}^Q$ getrennte Halbgeraden zerlegt werden. Auf eine Halbgerade werden sechs Beobachtungen projiziert, auf die andere Halbgerade eine. Das Minimum dieser Werte ist $1$. Falls keine andere Gerade $g_2^Q$ durch $\mathbf{y}^Q$ existiert, so dass auf eine Halbgerade von $g_2^Q$ weniger als eine Beobachtung projiziert wird, ist $d^{HS}(\mathbf{y}^Q; \mathbf{y}_1, \ldots, \mathbf{y}_k) = 1/k$, ansonsten $0$.

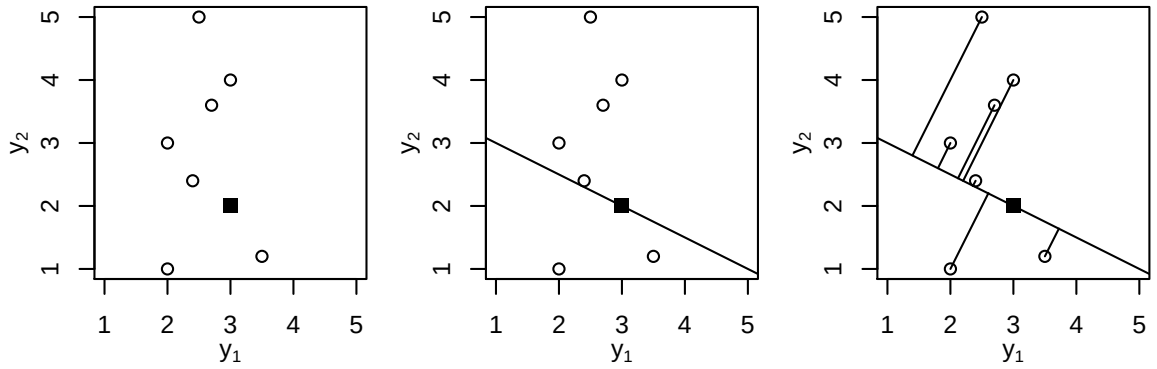



Abbildung 2.7: Visualisierung der Halbraumtiefe. Links: Beobachtungen (Kreise) und Punkt $\mathbf{y}^Q$ (Quadrat). Mitte: Die Gerade $g_1^Q$ verläuft durch den Punkt $\mathbf{y}^Q$. Rechts: Die Beobachtungen werden orthogonal auf $g_1^Q$ projiziert.

Bezogen auf den Bagplot wird für jede Beobachtung $\mathbf{y} \in \mathbb{R}^2$ die Halbraumtiefe $d^{HS}(\mathbf{y}; \cdot)$ bestimmt. Die konvexe Hülle der $50\%$ der Beobachtungen mit der größten Halbraumtiefe bildet den *bag* $B$, welcher vergleichbar ist mit der Box des Boxplots. Aufgrund von Bindungen enthält der *bag* tatsächlich in der Regel mehr als $50\%$ der Beobachtungen. Der *fence* kann als bivariate Version der *whiskers* des Boxplots verstanden werden. Sei $d_{0.5}^{HS} = \max_{\mathbf{y}} d^{HS}(\mathbf{y}; \cdot)$ der Halbraummedian und $\partial B = \bar{B} \backslash B^\circ$ der Rand des *bags*. Es werden im Folgenden die Beobachtungen außerhalb des *bags* betrachtet. Sei weiter $\partial B_{\mathbf{y}} := \{\mathrm{b} \in \partial B : \exists t \in [0,1] \text{ mit } t d_{0.5}^{HS} + (1-t)\mathbf{y} = \mathrm{b}\}$. Der *fence* ist die konvexe Hülle aller Beobachtungen $\mathbf{y}$, für die

$$\|d_{0.5}^{HS} - \mathbf{y}\| \leq 3\|d_{0.5}^{HS} - \partial B_{\mathbf{y}}\| \tag{2.14}$$

gilt. Die Beobachtungen außerhalb des *fence* werden als Ausreißer bezeichnet.

BEISPIEL 2.22. *Sei $\boldsymbol{\mathcal{Y}} = (\mathcal{Y}_1, \mathcal{Y}_2)^\top$ eine bivariat standardnormalverteilte Zufallsvariable, also $P_{\boldsymbol{\theta}} = \aleph(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ mit $\boldsymbol{\mu} = (0,0)^\top, \boldsymbol{\Sigma} = \mathbf{I}_2$. Weiterhin sei $\chi^2_{0.5,2}$ das $0.5$-Quantil der $\chi^2$-Verteilung mit zwei Freiheitsgraden. Wegen $\boldsymbol{\mu} = (0,0)^\top, \sigma_{11} = \sigma_{22}$ und $\sigma_{12} = 0$ hängt die Populationshalbraumtiefe $D^{HS}((y_1, y_2)^\top, P_{\boldsymbol{\theta}})$ nur von der euklidischen Distanz zum Ursprung $\sqrt{y_1^2 + y_2^2}$ ab. Wegen der Unabhängigkeit von $\mathcal{Y}_1$ und $\mathcal{Y}_2$ folgt mit*

$$\sqrt{\mathcal{Y}_1^2 + \mathcal{Y}_2^2} \sim \sqrt{\chi_2^2}$$

*die Bagplot-Ausreißerregion*

$$\begin{aligned}\text{outBagplot}(3, \aleph(\boldsymbol{\mu}, \boldsymbol{\Sigma})) &= \left\{(y_1, y_2)^\top \in \mathbb{R}^2 : \sqrt{y_1^2 + y_2^2} > 3\sqrt{\chi^2_{0.5,2}}\right\} \\ &= \left\{(y_1, y_2)^\top \in \mathbb{R}^2 : y_1^2 + y_2^2 > 9\chi^2_{0.5,2}\right\}.\end{aligned}$$

*Sei nun $F(\cdot, 2)$ die Verteilungsfunktion der $\chi^2$-Verteilung mit zwei Freiheitsgraden und $F^{-1}(\cdot, 2)$ die zugehörige Quantilfunktion. Gesucht ist $\alpha \in ]0, 1[$, so dass die Grenze der $\alpha$-Ausreißerregion der bivariaten Standardnormalverteilung mit der Bagplot-Ausreißerregion übereinstimmt:*

$$\begin{aligned}F^{-1}(1-\alpha, 2) &= 9F^{-1}(0.5, 2) \\ &\Leftrightarrow \alpha = 1 - F[9F^{-1}(0.5, 2), 2] = 0.19531\%.\end{aligned}$$

**Diskussion.** Ein wesentlicher Unterschied der hier diskutierten Ausreißerdefinitionen liegt darin, welche Eigenschaften der zugrunde liegenden Verteilung (a) der Nicht-Ausreißer und (b) der Ausreißer sie als bekannt voraussetzen. Während über die Verteilung der Nicht-Ausreißer zumeist Aussagen über den Verteilungstyp oder Symmetrie getroffen werden können, ist das für Ausreißer schwieriger. Beeinflussen Kovariablen die interessierenden, potenziell Ausreißer beinhaltenden Beobachtungen, kann das Aufstellen eines parametrischen oder nicht-parametrischen Modells eine Abweichung zwischen erwartetem und tatsächlich beobachteten Wert quantifizieren. Insbesondere wenn über den datengenerierenden Prozess der abhängigen Variable (Zielgröße) Informationen bekannt sind, können Verteilungsannahmen getroffen werden. Diese Verteilungsannahmen können sowohl in Bezug auf parametrische Modelle, genauer generalisierte lineare Modelle (GLM), als auch auf Ausreißeridentifikationsverfahren ($\alpha$-Ausreißer) als zusätzliche Informationsquelle genutzt werden. In Bezug auf die Identifikation von $\alpha$-Ausreißern im GLM besteht noch Forschungsbedarf, der in den nächsten Kapiteln aufgearbeitet werden soll. Dies betrifft zum Beispiel Kontingenztafeln, deren Einträge durch loglineare Poissonmodelle beschrieben werden können. Wenn die Hälfte der Einträge einer Zeile oder Spalte der Tafel Ausreißer sind, existieren zwei konkurrierende Modelle, von

denen eines besser zur Beschreibung der Ausreißer und das andere besser zur Beschreibung der Nicht-Ausreißer geeignet ist. Die Reaktion bestehender und neuer Ausreißeridentifizierer wird in solchen Fällen untersucht. Robuste Schätzer nehmen dabei eine besondere Stellung ein, da sie durch etwaige Ausreißer weniger beeinflusst werden.

# 3. GENERALISIERTE LINEARE MODELLE

In der Datenanalyse ist die Quantifizierung des Einflusses unabhängiger Variablen (Kovariablen) auf eine abhängige Variable (Zielgröße) von zentraler Bedeutung. Unter der Annahme, dass der Einfluss der Kovariablen linear in den unbekannten Parametern ist, existiert mit dem linearen Modell eine leicht zu interpretierende Möglichkeit, den Einfluss der Kovariablen auf die Zielgröße zu schätzen. In dieser Arbeit wird von nicht-stochastischen Kovariablen $\mathbf{z}_1, \ldots, \mathbf{z}_q \in \mathbb{R}^k$ ausgegangen, die in $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_q) \in \mathbb{R}^{k \times q}$ zusammengefasst werden. Die Designmatrix $\mathbf{X} \in \mathbb{R}^{k \times p}$ kann als Spalten nicht nur die beobachteten Werte der Kovariablen selbst, sondern auch Transformationen dieser Vektoren enthalten. Falls nicht anderweitig definiert, wird $\mathbf{X} = (\mathbb{1}_k, \mathbf{z}_1, \ldots, \mathbf{z}_q)$ verwendet. Allgemein bezeichnen $\mathbf{X}_{1\bullet}, \ldots, \mathbf{X}_{k\bullet}$ die Zeilen und $\mathbf{X}_{\bullet 1}, \ldots, \mathbf{X}_{\bullet p}$ die Spalten der Designmatrix. Für $\mathbf{Z}$ gilt eine analoge Notation.

Eine zentrale Annahme des linearen Modells liegt in der Normalverteilung der Zielgrößen $\mathcal{Y}_1, \ldots, \mathcal{Y}_k$ bedingt auf die Kovariablen. Im Rahmen generalisierter linearer Modelle findet eine Erweiterung auf Verteilungen aus der Klasse von Exponentialfamilien (vgl. Nelder und Wedderburn, 1972 sowie Wood, 2006, S. 62) statt.

DEFINITION 3.1. Einfache einparametrige Exponentialfamilie
*Eine Familie $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$ auf einem Messraum $(\mathfrak{T}, \mathcal{A}), \mathfrak{T} \subseteq \mathbb{R}$ heißt einfache einparametrige Exponentialfamilie, falls ihre Dichte in der Form*

$$f(y; \theta) = c(y) \exp\left(y\theta - b(\theta)\right) \tag{3.1}$$

*dargestellt werden kann, mit reellwertigen Funktionen $c : \mathfrak{T} \to \mathbb{R}_0^+$ und $b : \Theta \to \mathbb{R}$. Eine Erweiterung ist durch den positiven Dispersionsparameter $\phi$ möglich:*

$$f(y; \theta, \phi) = c(y, \phi) \exp\left(\frac{y\theta - b(\theta)}{a(\phi)}\right), \tag{3.2}$$

*mit einer reellwertigen Funktion* $a : \mathbb{R}^+ \to \mathbb{R}$.

Tabelle 3.1 fasst die Charakteristika einiger Exponentialfamilien zusammen. Neben sehr bekannten Verteilungen wird auch die Waldverteilung mit der Wahrscheinlichkeitsdichte $f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right) \mathbf{1}_{]0,\infty[}(y)$ angegeben. Wenn kein Dispersionsparameter benötigt wird, wird $\phi = 1$ gesetzt.

Tabelle 3.1: Charakteristika einiger Exponentialfamilien, vgl. Wood (2006, S. 61)

| Verteilung | $\theta$ | $b(\theta)$ | $\phi$ | $\mathsf{E}(\mathcal{Y})$ | $\mathsf{Var}(\mathcal{Y})$ |
|---|---|---|---|---|---|
| Normal, $\aleph(\mu, \sigma^2)$ | $\mu$ | $\frac{\theta^2}{2}$ | $\sigma^2$ | $\theta$ | $\sigma^2$ |
| Gamma, $\Gamma(\lambda, \nu)$ | $-\lambda$ | $-\ln(-\theta)$ | $\frac{1}{\nu}$ | $\frac{1}{\theta}$ | $\frac{1}{\theta^2 \nu}$ |
| Wald, $W(\mu, \lambda)$ | $-\frac{1}{2\mu^2}$ | $-\sqrt{-2\theta}$ | $\frac{1}{\lambda}$ | $\frac{1}{\sqrt{-2\theta}}$ | $\frac{1}{\lambda\sqrt{-2\theta^3}}$ |
| Bernoulli, $B(\pi)$ | $\mathrm{logit}(\pi)$ | $\ln(1+\exp(\theta))$ | 1 | $\mathrm{expit}(\theta)$ | $\frac{\exp(\theta)}{(1+\exp(\theta))^2}$ |
| Poisson, $Poi(\lambda)$ | $\ln(\lambda)$ | $\exp(\theta)$ | 1 | $\exp(\theta)$ | $\exp(\theta)$ |

Generalisierte lineare Modelle (GLM) erweitern das gewöhnliche lineare Regressionsmodell in mehreren Aspekten.

DEFINITION 3.2. Generalisiertes lineares Modell (Fahrmeir und Tutz, 1994, S. 18f.) *Es sei* $(\mathcal{Y}_1, \mathbf{Z}_{1\bullet}), \ldots, (\mathcal{Y}_k, \mathbf{Z}_{k\bullet})$ *die Stichprobe einer Zielgröße* $\mathcal{Y}$ *jeweils zusammen mit Beobachtungen von* $q$ *Kovariablen* $\mathbf{z}_1, \ldots, \mathbf{z}_q$. *Ein zugehöriges Modell, das den Zusammenhang zwischen der Zielgröße und den Kovariablen erklärt, wird* generalisiertes lineares Modell (GLM) *genannt, falls es aus den zwei folgenden Komponenten besteht:*

*(a) Verteilungskomponente: Die Dichte der Verteilung von* $\mathcal{Y}_i$ *gegeben* $\mathbf{Z}_{i\bullet}$ *lässt sich gemäß* (3.1) *oder* (3.2) *darstellen.*

*(b) Strukturkomponente: Der bedingte Erwartungswert* $\mathsf{E}(\mathcal{Y}_i \mid \mathbf{Z}_{i\bullet}) = \mu_i$ *hängt über eine injektive, hinreichend glatte sogenannte Responsefunktion* $h$ *vom linearen Prädiktor* $\eta_i := \mathbf{X}_{i\bullet}\boldsymbol{\beta}$ *in der Form* $\mathsf{E}(\mathcal{Y}_i \mid \mathbf{Z}_{i\bullet}) = h(\mathbf{X}_{i\bullet}\boldsymbol{\beta})$ *ab. Dabei ist* $\boldsymbol{\beta} \in \mathbb{R}^p$ *der unbekannte Regressionskoeffizientenvektor und* $\mathbf{X}_{i\bullet} \in \mathbb{R}^p$ *ein Designvektor, gegeben durch eine geeignete Funktion* $X(\mathbf{Z}_{i\bullet})$.

Die Inverse von $h$ wird als Linkfunktion $g$ bezeichnet. Häufige Wahlen sind $g(\mu) = \mu$ (Identitätslink), $g(\mu) = \ln(\mu)$ (Log-Link) und $g(\mu) = \mu^{-1}$ (inverser Link).

Wenn $\mathcal{Y}$ bedingt normalverteilt ist und $g(\mu) = \mu$ angenommen wird, ist das lineare Modell ein Spezialfall des GLM. Auch diskret verteilte $\mathcal{Y}$ mit einer Verteilung aus einer einfachen einparametrigen Exponentialfamilie können im Rahmen generalisierter linearer Modelle sinnvoll beschrieben werden.

## 3.1 LOGLINEARE MODELLE FÜR KATEGORIALE DATEN

Ein Merkmal wird als kategorial bezeichnet, falls es endlich viele Werte („Kategorien“) annehmen kann. Es können auch mehrere kategoriale Merkmale gleichzeitig beobachtet werden. Das Merkmal $I_d$ habe $k_d$ disjunkte Ausprägungen, $d = 1, \ldots, D$. Die daraus folgenden $\prod_{d=1}^{D} k_d = k$ Merkmalskombinationen können in einen $D$-dimensionalen Array $\mathbf{N}$ geschrieben werden. Die Kontingenztafel $\mathbf{N}$ dient der Darstellung der Häufigkeit der Merkmalsausprägung zweier oder mehrerer kategorialer Merkmale. Allgemein können zweidimensionale Kontingenztafeln wie in Tabelle 3.2 notiert werden. Die $k_1$ Zeilen in der Tafel können als $k_1$ Ausprägungen des Merkmals $I_1$ aufgefasst werden, die $k_2$ Spalten als $k_2$ Ausprägungen des Merkmals $I_2$. Seien $r \in \{1, \ldots, k_1\}$ und $c \in \{1, \ldots, k_2\}$, dann ist die Häufigkeit der Kombination $(r, c)$ in einer gegebenen Stichprobe durch $n_{rc}$ angegeben.

Tabelle 3.2: Kontingenztafel mit $k_1$ Zeilen, $k_2$ Spalten und $k$ Einträgen $n_{11}, \ldots, n_{k_1 k_2}$

| | 1 | 2 | $\cdots$ | $k_2$ |
|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | | $n_{1k_2}$ |
| 2 | $n_{21}$ | $n_{22}$ | | $n_{2k_2}$ |
| $\vdots$ | | | | |
| $k_1$ | $n_{k_1 1}$ | $n_{k_1 2}$ | | $n_{k_1 k_2}$ |

Die Randhäufigkeiten der Kontingenztafel sind durch $n_{r+} = \sum_{i=1}^{k_2} n_{ri}$ sowie $n_{+c} = \sum_{i=1}^{k_1} n_{ic}$ definiert. Jede Merkmalskombination $n_{rc}$ tritt mit einer gewissen Wahrscheinlichkeit $P((I_1, I_2) = (r, c)) =: \pi_{rc}$ auf. Die häufig getroffene Unabhängigkeitsannahme lautet:

$$\pi_{rc} = P(I_1 = r)P(I_2 = c) \quad \forall\, (r, c) \in \{1, \ldots, k_1\} \times \{1, \ldots, k_2\}. \tag{3.3}$$

BEISPIEL 3.3. *Das Hochgeschwindigkeitsflammspritzen ist ein Spritzprozess, um Materialien mit einer Beschichtung zu versehen und somit vor Verschleiß zu schützen. Im Sonderforschungsbereich 823 der TU Dortmund werden im Teilprojekt B1 solche Prozesse untersucht. Dafür wird ein Wolframcarbid-Cobalt-Pulver (WC-Co-Pulver) erhitzt und mit Hilfe einer Flamme auf ein Substrat gespritzt. Wird das Substrat nur kurz mit dem WC-Co-Pulver beschichtet, können sich darauf abscheidende Pulverpartikel identifiziert werden, vgl. Tillmann* et al. *(2013). Abbildung 3.1 zeigt einen Ausschnitt des Substrats nach dem Spritzprozess. Die darin zu erkennenden einzelnen hellgrauen und mehr oder weniger kreisförmigen Abscheidungen werden Splats genannt. Jeder Splat kann nach Größe und Form kategorisiert werden, wodurch sich eine Kontingenztafel ergibt.*

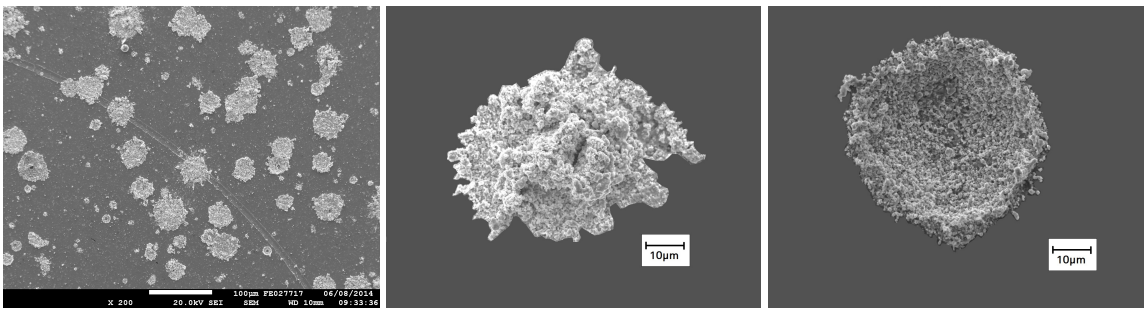



Abbildung 3.1: Rasterelektronenmikroskopaufnahmen von Splats. Links: Substrat (dunkelgrau) mit hellgrauen Splats. Mitte: Bergförmiger Splat. Rechts: Talförmiger Splat.

*In Tabelle 3.3 wird die Häufigkeit der Kombinationen aus Größe (groß, klein) und Querschnittform (bergförmig, flach, talförmig) dargestellt. Dabei ist zu erkennen, dass große Splats eher bergförmig oder flach sind und kleine Splats eher talförmig. Die beiden Merkmale scheinen nicht unabhängig zu sein. Besonders auffällig ist der Wert für $n_{23}$. Methoden zur Ausreißeridentifikation von Zellen in Kontingenztafeln oder ganzen Kontingenztafeln werden in Kapitel 4 vorgestellt.*

Tabelle 3.3: Kontingenztafel der Splats, kategorisiert von T. Priggemeier und B. Hussong (Mitarbeiter am Lehrstuhl für Werkstofftechnologie, TU Dortmund)

| | bergförmig | flach | talförmig |
|---|---|---|---|
| groß | $n_{11} = 45$ | $n_{12} = 48$ | $n_{13} = 30$ |
| klein | $n_{21} = 14$ | $n_{22} = 23$ | $n_{23} = 79$ |

Die im GLM übliche Unterscheidung zwischen unabhängigen und abhängigen Variablen ist hier folgendermaßen gegeben: Jede Zellhäufigkeit (Anzahl der Kombinationen aus den jeweiligen Merkmalsausprägungen) wird als abhängige Variable gesehen. Die Kovariablen sind durch die jeweilige betrachtete Zelle der Kontingenztafel festgelegt. Die Designmatrix $\mathbf{X}$ wird nicht beobachtet, sondern der Dimension der Kontingenztafel und der angenommenen Zusammenhangsstruktur der Merkmale entsprechend konstruiert.

Es werde eine $\times_{d=1}^{D} k_d$-Kontingenztafel mit $\mathcal{I} := \times_{d=1}^{D} I_d$, und $I_d = \{1, \ldots, k_d\}$ betrachtet. Die Einträge dieser Tafel werden in den Vektor $\mathbf{y} = (y_1, \ldots, y_k)^\top$ geschrieben. Dies geschieht durch eine Abbildung $\wp$. Für zweidimensionale Kontingenztafeln ist der $\mathrm{vec}$-Operator ein einfacher Spezialfall dieser Abbildung. Die Designmatrix $\mathbf{X}$ kann wie folgt partitioniert werden: $\mathbf{X} = (\mathbb{1}_k, \mathbf{X}^1, \ldots, \mathbf{X}^D)$, wobei $\mathbf{X}^d \in \{0,1\}^{k \times k_d - 1}$ mit Einträgen

$$\mathbf{X}_{ij}^d = \begin{cases} 1, & \text{wenn } y_i \text{ zu Kategorie } j \in I_d \text{ gehört} \\ 0, & \text{sonst} \end{cases} \tag{3.4}$$

für $d = 1, \ldots, D$ gilt. Interaktionen zweier Merkmale $I_{d_1}, I_{d_2}$ können ins Modell aufgenommen werden, wenn $\mathbf{X}$ um die Partition $\mathbf{X}^{d_1 \cap d_2}$ erweitert wird, für die in (3.4) die obere Bedingung zu „wenn $y_i$ zu Kategorie $j_1 \in I_{d_1}$ und $j_2 \in I_{d_2}$ gehört“ modifiziert wird. Für höherdimensionale Interaktionen ist $\mathbf{X}$ analog zu erweitern.

BEISPIEL 3.4. Designmatrizen einer $(3 \times 3)$-Kontingenztafel

*Die Designmatrix einer $(3 \times 3)$-Kontingenztafel im Unabhängigkeitsmodell kann wie folgt gewählt werden:*

$$\mathbf{X}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{3.5}$$

*Die einzelnen Spalten der transponierten Matrix korrespondieren hier zu den (zeilenweise*

*abgelesenen) Einträgen der Kontingenztafel. In Zeile 2 steht in Spalte $i$ eine 1, falls der $i$-te Eintrag der Tafel in der ersten Zeile der Tafel steht. In Zeile 4 steht in Spalte $i$ eine 1, falls der $i$-te Eintrag der Tafel in der ersten Spalte steht, und so weiter. Falls eine Interaktion zwischen den beiden Merkmalen angenommen wird, lautet die Designmatrix*

$$\mathbf{X}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Eine alternative Kodierung enthält an einigen Stellen der Designmatrix den Eintrag $-1$, wenn $y_i$ zur Referenzkategorie $k_d, d = 1, \ldots, D$ gehört, siehe (A.2) und (A.3) im Anhang A.5.

Zur Modellierung von Zähldaten aus Kontingenztafeln wird im Kontext generalisierter linearer Modelle eine Verteilungsannahme benötigt. Die Poissonverteilung ist eine mögliche Wahl, da sie eine einparametrige Exponentialfamilie bildet.

BEMERKUNG 3.5. *Sei $f(y;\theta) = \frac{\lambda^y \exp(-\lambda)}{y!}\mathbf{1}_{\mathbb{N}_0}(y)$ die Dichte der Poissonverteilung mit Parameter $\lambda > 0$. Sie bildet mit $c(y) = y!^{-1}\mathbf{1}_{\mathbb{N}_0}(y), \theta = \ln(\lambda), b(\theta) = \exp(\theta)$ in* (3.1) *eine einparametrige Exponentialfamilie.*

Außerdem ist die Poissonverteilung eine sinnvolle Wahl, wenn die Zahl $n$ der Merkmalsträger *a priori* unbekannt ist. Im Kontext von Kontingenztafeln ist das der Fall, wenn die Messung zeitlich oder räumlich begrenzt ist und nicht abbricht, wenn eine vorher festgelegte Zahl an Merkmalsträgern erreicht wurde. In Beispiel 3.3 ist aus technischen Gründen nur eine zeitliche Beschränkung möglich. Poissonverteilte Zielgrößen $\mathcal{Y}$ sind somit im Kontext generalisierter linearer Modelle handhabbar. In Kombination mit dem Log-Link stimmt dieser Spezialfall mit dem schon

vor Formulierung des GLM (Nelder und Wedderburn, 1972) bekannten loglinearen Poissonmodell überein (siehe z. B. Bishop, 1969). Besondere Relevanz haben loglineare Poissonmodelle im GLM-Kontext unter anderem, da sie sich zur Beschreibung von Kontingenztafeln eignen. Weitere Anwendungsmöglichkeiten betreffen die Modellierung von Zählzeitreihen, siehe allgemein die Monografie von Kedem und Fokianos (2002) sowie Fried *et al.* (2014) für die Verknüpfung von Zählzeitreihen und Ausreißeranalyse. Anschließend wird das in dieser Arbeit betrachtete loglineare Poissonmodell direkt in der Notation generalisierter linearer Modelle definiert:

DEFINITION 3.6. Loglineares Poissonmodell
*Gegeben seien Zufallsvariablen $\mathcal{Y}_i$ und Kovariablen $\mathbf{Z}_{i\bullet}$ mit $\mathcal{Y}_i|\mathbf{Z}_{i\bullet} \sim Poi(\lambda_i), i = 1, \ldots, k$. Sei weiterhin $\mathbf{X}_{i\bullet} = X(\mathbf{Z}_{i\bullet})$ mit passend gewählter Funktion $X$ der $i$-te Zeilenvektor der Designmatrix $\mathbf{X}$. Dann sei das* loglineare Poissonmodell *definiert durch*

$$\mathsf{E}(\mathcal{Y}_i|\mathbf{Z}_{i\bullet}) = \exp(\mathbf{X}_{i\bullet}\boldsymbol{\beta}), \qquad i = 1, \ldots, k, \tag{3.6}$$

*wobei $\boldsymbol{\beta} \in \mathbb{R}^p$ ein unbekannter Parametervektor ist.*

Zur Modellierung der Zellhäufigkeiten einer Kontingenztafel wird $\mathbf{X}$ wie bereits beschrieben gebildet. Wenn das loglineare Poissonmodell keine Interaktionsterme beinhaltet, wird es loglineares Poisson-Unabhängigkeitsmodell genannt. Falls es alle möglichen Interaktionsterme beinhaltet, wird es saturiertes loglineares Poissonmodell genannt. Im zweidimensionalen Fall sind das alle Zweifach-Interaktionen, im dreidimensionalen Fall alle Zweifach- und Dreifach-Interaktionen und so weiter.

Wird eine Kontingenztafel mit bekanntem Stichprobenumfang $n$ erhoben, liegt ein anderer datengenerierender Prozess vor als bisher angenommen (Agresti, 2002, S. 6 f.). Sei $\mathbf{N} \in \mathbb{N}_0^{\times_{d=1}^{D} k_d}$ die $D$-dimensionale Kontingenztafel und $\mathbf{y} = \wp(\mathbf{N}) \in \mathbb{R}^k$ enthalte die Einträge von $\mathbf{N}$ in vektorisierter Form. Dann ist die $\mathbf{y}$ zugrunde liegende Zufallsvariable $\boldsymbol{\mathcal{Y}}$ multinomialverteilt mit Parametern $n$ und $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$, wobei $\sum_{i=1}^k \theta_i = 1$ und $\theta_i \geq 0 \; \forall \; i$. Die Wahrscheinlichkeitsdichte von $\boldsymbol{\mathcal{Y}}$ ist durch

$$f_{\boldsymbol{\mathcal{Y}}}(y_1, \ldots, y_k; n, \boldsymbol{\theta}) = \begin{cases} n! \prod_{i=1}^k \frac{\theta_i^{y_i}}{y_i!}, & \text{falls } \sum_{i=1}^k y_i = n \text{ und } y_i \in \mathbb{N}_0 \; \forall \; i, \\ 0, & \text{sonst,} \end{cases}$$

gegeben. Da die Multinomialverteilung eine $(k-1)$-parametrige Exponentialfamilie bildet, ist sie im Rahmen multivariater GLMs handhabbar (Fahrmeir und Tutz, 1994, Kap. 3.1).

### *3.1.1 Regressionsschätzer und ihre Eigenschaften*

Die Strukturkomponente $\mathsf{E}(\mathcal{Y}_i \mid \mathbf{Z}_{i\bullet}) = h(\mathbf{X}_{i\bullet}\boldsymbol{\beta})$ im generalisierten linearen Modell enthält den unbekannten Regressionskoeffizientenvektor $\boldsymbol{\beta} \in \mathbb{R}^p$, welcher basierend auf einer Stichprobe $\mathcal{Y}_1, \ldots, \mathcal{Y}_k$ und der Designmatrix $\mathbf{X} \in \mathbb{R}^{k\times p}$ durch einen Schätzer $\widehat{\boldsymbol{\beta}} = T(\{(\mathbf{X}_{i\bullet}, \mathcal{Y}_i), i = 1, \ldots, k\}, g) \in \mathbb{R}^p$ zu schätzen ist. Im Folgenden wird $T$ als Schätzer bezeichnet. Spezielle Methoden zur Schätzung von $\boldsymbol{\beta}$ werden in Abschnitt 3.1.2 vorgestellt. Zunächst werden jedoch in diesem Abschnitt wünschenswerte Eigenschaften von Regressionsschätzern thematisiert, vgl. Rousseeuw und Leroy (1987, S. 116 ff.).

DEFINITION 3.7. Exact-Fit-Eigenschaft
*Seien $k$ Realisationen $\{(\mathbf{Z}_{i\bullet}, y_i);\ i = 1, \ldots, k\}$ einer Stichprobe und ein Schätzer $T$ im GLM mit Designmatrix $\mathbf{X}$ und Linkfunktion $g$ gegeben, wobei $T(\{(\mathbf{Z}_{i\bullet}, y_i);\ i = 1, \ldots, k\}, g) = \widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$. Wenn ein $\boldsymbol{\theta} \in \mathbb{R}^p$ existiert, so dass mindestens $k - \lfloor k/2 \rfloor + p - 1$ der Realisationen $g(y_i) = \mathbf{X}_{i\bullet}\boldsymbol{\theta}$ exakt erfüllen und $\widehat{\boldsymbol{\beta}} = \boldsymbol{\theta}$ gilt, hat $T$ die* Exact-Fit-Eigenschaft.

Basierend auf der Exact-Fit-Eigenschaft definieren Rousseeuw und Leroy (1987, S. 123) den Exact-Fit-Punkt. Er basiert auf einem kontaminierten Vektor $\widetilde{\mathbf{y}} = (\tilde{y}_1, \ldots, \tilde{y}_k)^\top$, der in allen bis auf $m$ Einträgen mit y übereinstimmt und ansonsten beliebig extreme Werte annehmen kann. In einer Situation, in der in Definition 3.7 alle Realisationen der Stichprobe $g(y_i) = \mathbf{X}_{i\bullet}\boldsymbol{\theta}$ exakt erfüllen, ist der Exact-Fit-Punkt der kleinstmögliche Anteil kontaminierter Beobachtungen $\tilde{y}_i$, der dafür sorgt, dass $T(\{(\mathbf{Z}_{i\bullet}, \tilde{y}_i);\ i = 1, \ldots, k\}, g) \neq \boldsymbol{\beta}$. Ein mit der Exact-Fit-Eigenschaft verwandtes Konzept ist der Bruchpunkt. Eine häufig geforderte Eigenschaft ist, dass bei einigen beliebig extremen Ausreißern der Schätzer nicht beliebig verfälscht werden darf. Dies ist die Idee hinter dem Bruchpunkt.

DEFINITION 3.8. Finite-Sample-Bruchpunkt
*Seien Realisationen* $\{(\mathbf{Z}_{i\bullet}, y_i);\ i = 1, \ldots, k\}$ *einer Stichprobe und ein Schätzer* $T$ *gegeben, wobei* $T(\{(\mathbf{Z}_{i\bullet}, y_i);\ i = 1, \ldots, k\}) = \widehat{\boldsymbol{\beta}}$. *Sei* $\widetilde{\mathbf{y}}(m) = (\tilde{y}_1, \ldots, \tilde{y}_k)^\top$ *ein Vektor, der in* $k - m$ *Elementen mit* $\mathbf{y}$ *übereinstimmt und dessen übrige* $m$ *Elemente beliebig extrem kontaminiert wurden. Die so verursachte schlimmstmögliche* Verzerrung (bias) *des Schätzers im Vergleich zum unkontaminierten Vektor* $\mathbf{y}$ *ist durch*

$$\begin{aligned} &\textit{bias}(m; T, \{(\mathbf{Z}_{i\bullet}, y_i);\ i = 1, \ldots, k\}) := \\ &\sup_{\widetilde{\mathbf{y}}(m) \in \mathbb{R}^k} (\|T(\{(\mathbf{Z}_{i\bullet}, \tilde{y}_i);\ i = 1, \ldots, k\}) - T(\{(\mathbf{Z}_{i\bullet}, y_i);\ i = 1, \ldots, k\})\|) \end{aligned}$$

*gegeben. Dann ist der* Finite-Sample-Bruchpunkt (finite-sample breakdown point) $\varepsilon^*$ *des Schätzers bezüglich der Beobachtungen definiert durch*

$$\begin{aligned} \varepsilon^*(T, \{(\mathbf{Z}_{i\bullet}, y_i);\ i = 1, \ldots, k\}) = &\min\{m/k : \\ &\textit{bias}(m; T, \{(\mathbf{Z}_{i\bullet}, y_i); i = 1, \ldots, k\}) = \infty\}. \end{aligned}$$

In Definition 3.8 kann $T$ zum Beispiel ein beliebiger Schätzer im GLM mit gewählter Verteilung der Zielgröße und Linkfunktion sein. Sinnvolle Schätzer haben einen Finite-Sample-Bruchpunkt von höchstens $0.5$. Insbesondere bei einem möglichen Vorliegen von Ausreißern sind Schätzverfahren mit hohem Finite-Sample-Bruchpunkt zu bevorzugen. Allerdings ist der Finite-Sample-Bruchpunkt ein Kriterium, das nur den schlimmstmöglichen Fall eines Ausreißers betrachtet. Andere Robustheitskriterien wie die Influenzkurve (vgl. Rousseeuw und Leroy, 1987, S. 186 f.) können den Einfluss unterschiedlich extremer Ausreißer visualisieren.

### *3.1.2 Spezielle Schätzer im loglinearen Poissonmodell*

Wegen der bekannten bedingten Verteilung im loglinearen Poissonmodell kann wie für das gewöhnliche lineare Modell ein ML-Schätzer $\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}$ von $\boldsymbol{\beta}$ bestimmt werden. Bei ungruppierten Beobachtungen lautet der Beitrag von $\mathcal{Y}_i$ zur Log-Likelihood $l_i(\theta_i) = \frac{\mathcal{Y}_i\theta_i - b(\theta_i)}{\phi}$ (Fahrmeir und Tutz, 1994, S. 38). Mit Tabelle 3.1 folgt der ML-

Schätzer

$$\widehat{\boldsymbol{\beta}}^{\mathrm{ML}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmax}} \sum_{i=1}^{k} \left(\mathcal{Y}_i \, \mathbf{X}_{i\bullet}\boldsymbol{\beta} - \exp(\mathbf{X}_{i\bullet}\boldsymbol{\beta})\right). \tag{3.7}$$

Der ML-Schätzer hat einen Finite-Sample-Bruchpunkt von $1/k$ und ist damit nicht robust gegenüber Ausreißern. Im Gegensatz zum gewöhnlichen linearen Modell existiert für das Maximierungsproblem (3.7) zumeist keine Lösung in geschlossener Form. Zur numerischen Lösung kann die Methode der iterativ gewichteten kleinsten Quadrate (Agresti, 2002, S. 343) verwendet werden. Unter der Unabhängigkeitsannahme (3.3), die sich auf beliebig-dimensionale Kontingenztafeln verallgemeinern lässt, sind die ML-Schätzwerte basierend auf Poisson-Likelihood und Multinomial-Likelihood identisch (Agresti, 2002, S. 340).

Eine Alternative zum ML-Schätzer ist der $L_1$- oder LAD-Schätzer, kurz $\widehat{\boldsymbol{\beta}}^{L_1}$. Hubert (1997) schlägt seine Verwendung im Kontext von Kontingenztafeln vor. Unter der Bedingung, dass alle $\mathcal{Y}_i > 0$ sind, minimiert $\widehat{\boldsymbol{\beta}}^{L_1}$ die Summe der absoluten Distanzen zwischen logarithmierten Beobachtungen und linearen Prädiktoren:

$$\widehat{\boldsymbol{\beta}}^{L_1} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{k} |\ln \mathcal{Y}_i - \mathbf{X}_{i\bullet}\boldsymbol{\beta}| \tag{3.8}$$

Hubert zeigt, dass der Bruchpunkt des $L_1$-Schätzers für zweidimensionale $k_1 \times k_2$ Kontingenztafeln unter dem loglinearen Unabhängigkeitsmodell durch

$$\varepsilon^*\left(\widehat{\boldsymbol{\beta}}^{L_1}, \{(\mathbf{X}_{i\bullet}, y_i);\ i = 1, \ldots, k\}, k_1, k_2\right) = \frac{1}{k_1 k_2}\left\lfloor \frac{\min(k_1, k_2) + 1}{2} \right\rfloor$$

gegeben ist. Zur Lösung des Minimierungsproblems (3.8) wird in Kapitel 4 der L-BFGS-B-Algorithmus von Byrd *et al.* (1995) verwendet.

Ein weiterer Ansatz zur Konstruktion eines robusten Schätzers basiert auf der Chi-Quadrat-Statistik.

DEFINITION 3.9. LTCS-Schätzer *(Shane und Simonoff, 2001)*
*In der Situation des loglinearen Poissonmodells* (3.6) *ist der* LTCS-Schätzer (least trim-

med chi-squared residuals estimator) *mit Tuningparameter* $\mathfrak{h}$ *gegeben durch*

$$\widehat{\boldsymbol{\beta}}_{\mathfrak{h}}^{LTCS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{k} \mathbf{1}_{[0,\chi^2_{(\mathfrak{h})}]}(\chi^2_{(i)})\chi^2_{(i)}, \tag{3.9}$$

*wobei* $\chi^2_{(i)}$ *die aufsteigend sortierten Summanden der Chi-Quadrat-Statistik* $\chi^2 = \sum_{i=1}^{k} \chi_i^2 = \sum_{i=1}^{k} \frac{(\mathcal{Y}_i - \exp(\mathbf{X}_{i\bullet}\boldsymbol{\beta}))^2}{\exp(\mathbf{X}_{i\bullet}\boldsymbol{\beta})}$ *bezeichnen.*

Shane und Simonoff (2001) stellen fest, dass das Minimierungsproblem (3.9) durch viele lokale Minima charakterisiert ist. Eine Möglichkeit, zu einer Lösung von (3.9) zu kommen, wird im späteren Verlauf dieser Arbeit in Bemerkung 4.8 beschrieben. Der Bruchpunkt dieses Schätzers hängt vom Tuningparameter $\mathfrak{h}$ ab. Shane und Simonoff finden für

$$\mathfrak{h} = \mathfrak{h}_{op} \in [\lfloor (k+p+1)/2 \rfloor, \lfloor (k+p+2)/2 \rfloor], \tag{3.10}$$

den optimalen Bruchpunkt von $\widehat{\boldsymbol{\beta}}_{\mathfrak{h}}^{\text{LTCS}}$ in

$$\varepsilon^* \left( \widehat{\boldsymbol{\beta}}_{\mathfrak{h}_{op}}^{\text{LTCS}}, \{(\mathbf{X}_{i\bullet}, y_i);\ i = 1, \ldots, k\} \right) \in [\lfloor (k-p+1)/2 \rfloor / k, \lfloor (k+1)/2 \rfloor / k].$$

Cantoni und Ronchetti (2001) geben eine Übersicht über weitere, hier nicht thematisierte robuste Schätzer.

## 3.2 MODELLIERUNG FUNKTIONALER ZIELGRÖSSEN

Bislang wurden für generalisierte lineare Modelle ausschließlich endlichdimensionale Zielgrößen (Skalare, Vektoren) in Betracht gezogen. Um die Modellierung unendlichdimensionaler Zielgrößen (Funktionen) zu ermöglichen ist es sinnvoll mit Werkzeugen der funktionalen Datenanalyse (*functional data analysis*, FDA) zu arbeiten. Sie erfährt als relativ junges Forschungsgebiet der Statistik seit den 1990er Jahren zunehmende Aufmerksamkeit. Eine verbale Definition der FDA geben Hsing und Eubank (2015, S. 1):

> „Functional data analysis [...] is concerned with the development of methodology for statistical analysis of data that represent sample paths of [stochastic; A. R.] processes for which the index set is some (closed) interval of the real line [...]."

Funktionale Daten können in realen Anwendungsproblemen vorkommen, wenn Beobachtungen $y$ über ein gewisses Kontinuum $\mathcal{T} = [a, b] \subset \mathbb{R}$ (z. B. Ort, Zeit, Wellenlänge) – hinweg beobachtet werden und davon auszugehen ist, dass bei beliebig feiner Messgenauigkeit die Differenzen zwischen beliebig nah benachbarten Kontinuumselementen beliebig klein werden. Aus der Analysis ist diese Eigenschaft als Stetigkeit bekannt. Im Falle solcher Daten kann die formale Darstellung wie folgt gewählt werden.

NOTATION 3.10. Modellierungsansatz für funktionale Daten nach Cuevas (2014)
*Eine funktionale Beobachtung $y = \{y(t) \mid t \in \mathcal{T}\}$ wird als Realisation eines zeitstetigen stochastischen Prozesses $\mathcal{Y} = \{\mathcal{Y}(t) \mid t \in \mathcal{T}\}$ auf der kompakten Indexmenge $\mathcal{T} \subset \mathbb{R}$ aufgefasst. Wegen der Stetigkeit gilt $y \in \mathcal{C}(\mathcal{T})$, wobei $\mathcal{C}(\mathcal{T})$ die Menge der reellen stetigen Funktionen auf $\mathcal{T}$ bezeichne. Der auf $\mathcal{C}(\mathcal{T})$ basierende Banachraum mit der üblichen Supremumsnorm wird durch $\mathbb{B}$ symbolisiert. Zusammen folgt: $\mathcal{Y} \in \mathbb{B}$.*

Eine sehr flexible und weit verbreitete Klasse zeitstetiger stochastischer Prozesse bilden die Gaußprozesse.

DEFINITION 3.11. Gaußprozess (Cuevas, 2014)
*Sei $\mathcal{Y} = \{\mathcal{Y}(t) \mid t \in \mathcal{T}\}$ ein zeitstetiger stochastischer Prozess auf der kompakten Menge $\mathcal{T} \subset \mathbb{R}$ und $\mathcal{Y} \in \mathbb{B}$. Weiterhin sei $y^*$ ein Element aus dem topologischen Dualraum $\mathbb{B}^*$ von $\mathbb{B}$, also ein stetiges lineares Funktional. $\mathcal{Y}$ ist ein* Gaußprozess, *falls $y^*(\mathcal{Y})$ für alle $y^* \in \mathbb{B}^*$ eine reellwertige normalverteilte Zufallsvariable ist, wobei $\mu = \{\mu(t) \mid t \in \mathcal{T}\} = \{\mathsf{E}(\mathcal{Y}(t)) \mid t \in \mathcal{T}\}$ die Erwartungsfunktion und $\sigma = \{\sigma(s,t) \mid s,t \in \mathcal{T}\} = \{\mathsf{Cov}(\mathcal{Y}(s), \mathcal{Y}(t)) \mid s,t \in \mathcal{T}\}$ die Kovarianzfunktion ist; kurz $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$.*

In der Praxis können Daten nicht kontinuierlich aufgezeichnet werden, sondern nur in einem bestimmten Abstand zueinander (z. B. eine Messung pro Sekunde bei Zeitreihen oder eine Messung pro Quadratmeter bei Höhenprofilen). Sei $y_{t_j}$ der

beobachtete Wert zum Zeitpunkt $t_j \in \mathcal{T}$ mit $j \in \{1, \ldots, T\}$ und $t_j < t_{j+1} \; \forall \; j = 1, \ldots, T-1$. Die Werte für $y_{t_j}$ werden zum Vektor $\mathbf{y} = (y_{t_1}, \ldots, y_{t_T})^\top$ zusammengefasst. Falls mehrere funktionale Beobachtungen aufgezeichnet werden, können sowohl die Zeitpunkte $\tilde{t}_1, \ldots, \tilde{t}_T$ selbst, als auch deren Anzahl variieren.

Es werden nun einige Möglichkeiten vorgestellt, um zu einer reellen stetigen Funktion auf $\mathcal{T}$ zu gelangen, die $\mathbf{y}$ approximiert oder interpoliert. Lineare Interpolation ist als naive Variante i. d. R. ungeeignet, da sie zum einen keinen wirklichen Erkenntnisgewinn generiert und zum anderen für alle $t \notin [t_1, t_T]$ nicht definiert ist. Andere, z. B. polynomielle Interpolationsverfahren sind auf ganz $\mathcal{T}$ definiert, aber modellieren auch potenzielle Messfehler mit. Falls von Messfehlern auszugehen ist, sollte $\mathbf{y}$ approximiert werden. Man unterscheidet hier zwischen parametrischen Ansätzen wie einer Fourierreihenentwicklung (vor allem bei periodischen Daten) oder Splines und nichtparametrischen Ansätzen wie *localized least squares*. Die Grundlage für diese Glättungsverfahren liefert die an Guo (2004) angelehnte Definition 3.12.

DEFINITION 3.12. „Signal plus Rauschen"-Darstellung funktionaler Daten
*Sei $\boldsymbol{\mathcal{Y}} = (\mathcal{Y}_{t_1}, \ldots, \mathcal{Y}_{t_T})^\top$ der Zufallsvektor der Zielgröße, das Signal $\mathcal{F} \in \mathcal{C}(\mathcal{T})$ ein unbekannter zeitstetiger stochastischer Prozess und das Rauschen $\boldsymbol{\epsilon} = (\epsilon_{t_1}, \ldots, \epsilon_{t_T})^\top$ ein latenter Störvektor, für den $\boldsymbol{\epsilon} \sim \aleph(\mathbf{0}, \mathsf{Cov}(\boldsymbol{\epsilon}))$ gilt. Dann definiert*

$$\mathcal{Y}_{t_j} = \mathcal{F}(t_j) + \epsilon_{t_j}, \qquad j \in \{1, \ldots, T\}$$

*das* datengenerierende Modell in der „Signal plus Rauschen"-Darstellung.

BEMERKUNG 3.13. *Im Gegensatz zu typischen multivariaten Datensituationen kann in der „Signal plus Rauschen"-Darstellung nicht davon ausgegangen werden, dass $\mathsf{Cov}(\boldsymbol{\epsilon})$ eine Diagonalmatrix ist, da insbesondere autokorreliertes Rauschen in funktionalen Daten vorliegen oder sich die Varianz über $\mathcal{T}$ hinweg ändern kann.*

Die Schätzung von $\mathcal{F}$ ist nichttrivial, da $\mathcal{F}$ im unendlichdimensionalen Raum $\mathcal{C}(\mathcal{T})$ lebt. Häufig wird dazu übergegangen, $\mathcal{F}$ durch eine Funktion mit endlichdimensionaler Basis zu approximieren. Es werden nun zwei parametrische Verfahren aufgeführt, auf die in diesem Kapitel wiederholt eingegangen wird. Bei der Fourierreihenentwicklung wird angenommen, dass $\mathcal{F}$ einer gewissen Periodizität $\omega \in \mathbb{R}$

unterliegt und sich für hinreichend großes $K \in \{2n \mid n \in \mathbb{N}\}$ wie folgt darstellen lässt:

$$\begin{aligned}\mathcal{F}^F(t) &= \delta_0^F + \sum_{k=1}^{K/2}[\delta_{2k-1}^F \sin(k\omega t) + \delta_{2k}^F \cos(k\omega t)] + \mathcal{U}(t), \qquad t \in \mathcal{T}, \\ &= \sum_{k=0}^{K} \delta_k^F \phi_k^F(t) + \mathcal{U}(t) =: \left(\boldsymbol{\delta}^F\right)^\top \boldsymbol{\phi}^F(t) + \mathcal{U}(t),\end{aligned}$$

wobei

$$\phi_k^F(t) = \begin{cases} \sin(k\omega t), & \text{falls } k \in \{2n-1 \mid n \in \mathbb{N}\}, \\ \cos(k\omega t), & \text{falls } k \in \{2n \mid n \in \mathbb{N}_0\}, \end{cases}$$

$\mathsf{E}(\mathcal{U}(t)) = 0 \; \forall \; t \in \mathcal{T}$ gilt und $\boldsymbol{\delta}^F = (\delta_0^F, \ldots, \delta_K^F)^\top$ mit der KQ-Methode

$$\widetilde{\boldsymbol{\delta}}^F = \underset{\boldsymbol{\delta}^F}{\operatorname{argmin}} \sum_{i=1}^{T} \left[ \mathcal{Y}_{t_i} - \left( \delta_0^F + \sum_{k=1}^{K/2} \left[\delta_{2k-1}^F \sin(k\omega t_i) + \delta_{2k}^F \cos(k\omega t_i)\right] \right) \right]^2$$

geschätzt werden kann. Somit ist ein Fourierreihen-Schätzer für $\mathcal{F}(t)$ durch

$$\widetilde{\mathcal{F}}^F(t) = \widetilde{\delta}_0^F + \sum_{k=1}^{K/2}[\widetilde{\delta}_{2k-1}^F \sin(k\omega t) + \widetilde{\delta}_{2k}^F \cos(k\omega t)] = \left(\widetilde{\boldsymbol{\delta}}^F\right)^\top \boldsymbol{\phi}^F(t)$$

definiert. Die Anzahl der Basisfunktionen $K$ sollte so gewählt werden, dass ein guter Kompromiss zwischen Bias- und Varianzminimierung erreicht wird. Die Fourierdarstellung hat den Vorteil, dass analytische Ableitungen sehr leicht bestimmt werden können und die Funktion in geschlossener Form darstellbar ist.

Eine flexiblere Alternative bieten *B-Splines*. Zunächst wird $\mathcal{T}$ durch innere Knoten $\tau_1 \leq \ldots \leq \tau_{L-1} \in \mathcal{T}$ in $L$ (zumeist gleichgroße) disjunkte Teilintervalle aufgeteilt. Ähnlich wie im Fall der Fourierbasis muss auch hier die Anzahl $K \in \mathbb{N}$ der Basisfunktionen $\phi^{BS}$ (oder deren Ordnung $m$) festgelegt werden, wobei $K = L - 1 + m$ gilt. Pro Teilintervall sind $m$ Basisfunktionen echt positiv. Mit Hilfe einer Rekursionsformel können die Basisfunktionen explizit dargestellt werden, siehe z. B. Wood (2006, S. 152). Abbildung 3.2 zeigt ein Beispiel für $K = 6$ und $L = 3$.

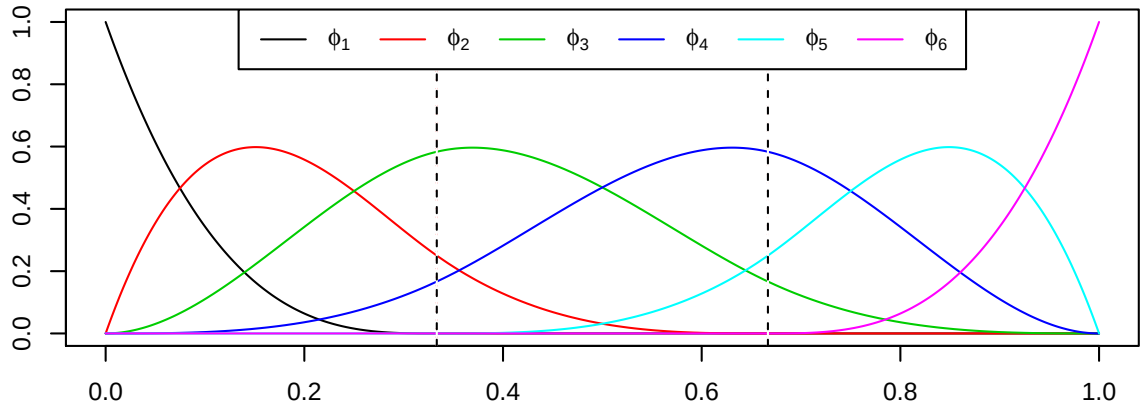



Abbildung 3.2: Sechs B-Spline-Basisfunktionen $\phi_1, \ldots, \phi_6$ mit inneren Knoten $\tau_1 = \frac{1}{3}, \tau_2 = \frac{2}{3}$

Nach der Festlegung des Tupels $(K, m)$ und des Vektors $(\tau_1, \ldots, \tau_{L-1})^\top$ wird das Modell für eine B-Spline-Basis durch

$$\mathcal{F}^{BS}(t) = \delta_0^{BS} + \sum_{k=1}^{K} \delta_k^{BS} \phi_k^{BS}(t) + \mathcal{U}(t) =: \left(\boldsymbol{\delta}^{BS}\right)^\top \boldsymbol{\phi}^{BS}(t) + \mathcal{U}(t), \qquad t \in \mathcal{T}$$

definiert, mit $\mathsf{E}(\mathcal{U}(t)) = 0 \; \forall \, t \in \mathcal{T}$, $\boldsymbol{\delta}^{BS} = (\delta_0^{BS}, \ldots, \delta_K^{BS})^\top$ und $\phi_0^{BS}(t) = 1 \; \forall \, t \in \mathcal{T}$. Die im $L_2$-Sinne optimale Linearkombination von Basisfunktionen erhält man mittels

$$\widetilde{\boldsymbol{\delta}}^{BS} = \underset{\boldsymbol{\delta}^{BS}}{\operatorname{argmin}} \sum_{i=1}^{T} \left[ \mathcal{Y}_{t_i} - \sum_{k=0}^{K} \delta_k^{BS} \phi_k^{BS}(t_i) \right]^2,$$

woraus folgt, dass

$$\widetilde{\mathcal{F}}^{BS}(t) = \sum_{k=0}^{K} \widetilde{\delta}_k^{BS} \phi_k^{BS}(t) = \left(\widetilde{\boldsymbol{\delta}}^{BS}\right)^\top \boldsymbol{\phi}^{BS}(t), \qquad t \in \mathcal{T}$$

eine B-Spline-Schätzung von $\mathcal{F}(t)$ ist. Alternativ kann die gewichtete KQ-Methode mit der Inversen der geschätzten Varianz-Kovarianz-Matrix als Gewichtsmatrix verwendet werden, siehe Ramsay und Silverman (2005, S. 61).

Bei diesen parametrischen Verfahren gehen in $\widetilde{\mathcal{F}}(t_i)$ nur implizit verstärkt Beobachtungen in der Nähe von $t_i$ ein. Bei Verfahren mit lokalen Ansätzen, den lokalisierten KQ-Methoden *(localized least squares)*, ist dies sogar explizit der Fall. Der zugehörige

Nadaraya-Watson-Schätzer (Nadaraya, 1964; Watson, 1964) ist gegeben durch

$$\widetilde{\mathcal{F}}^{NW}(t) = \frac{\sum_{i=1}^{T} y_{t_i} K_h(t - t_i)}{\sum_{i=1}^{T} K_h(t - t_i)}, \quad t \in \mathcal{T},$$

wobei $K_h(u) = \frac{1}{h} K(\frac{u}{h})$ ein reskalierter Kern ist, zum Beispiel der Gaußkern, siehe (2.10). Da $\widetilde{\mathcal{F}}^{NW}(t)$ für jedes $t \in \mathcal{T}$ separat bestimmt wird, ist der Rechenaufwand dieses nichtparametrischen Verfahrens deutlich höher als bei parametrischen Verfahren. Außerdem kann die Schätzung an den Rändern von $\mathcal{T}$ ungenau sein. Vorteilhaft sind nichtparametrische Schätzer vor allem, um lokale Besonderheiten der Kurven beizubehalten.

In der vorliegenden Arbeit liegt der Fokus jedoch auf parametrischen funktionalen Methoden. Ramsay und Silverman (2005) haben zu diesem Thema eine grundlegende Monografie verfasst, die von der Aufbereitung der Daten über funktionale Hauptkomponentenanalyse (vgl. Anhang A.2) bis hin zu linearen Modellen mit funktionaler Zielgröße ein anwendungsorientierter Wegweiser ist. Ausgehend von den Ausführungen von Ramsay und Silverman (2005) und Faraway (1997) wird in den folgenden beiden Unterkapiteln das GLM für funktionale Zielgrößen entwickelt und in Abschnitt 3.2.3 auf weitere Aspekte der funktionaler Regression eingegangen.

### *3.2.1 Function-on-Scalar Regression*

Das lineare Modell mit funktionaler Zielgröße wurde erstmals von Faraway (1997) beschrieben und später von Reiss *et al.* (2010) als *function-on-scalar regression* (FOSR) bezeichnet. Sei $\boldsymbol{\mathcal{F}}(t) = (\mathcal{F}_1(t), \ldots, \mathcal{F}_n(t))^\top$ der Vektor der funktionalen Zielgrößen, $\mathbf{X} \in \mathbb{R}^{n \times p}$ die Designmatrix der skalaren Kovariablen, $\boldsymbol{\beta}(t) = (\beta_0(t), \ldots, \beta_{p-1}(t))^\top$ der unbekannte funktionale Regressionskoeffizientenvektor, der Zufallsvektor $\boldsymbol{\mathcal{U}}(t) = (\mathcal{U}_1(t), \ldots, \mathcal{U}_n(t))^\top$ der latente funktionale Residuenvektor und $t \in \mathcal{T}$. Dann bezeichnet $\boldsymbol{\mathcal{F}}(t) = \mathbf{X}\boldsymbol{\beta}(t) + \boldsymbol{\mathcal{U}}(t), t \in \mathcal{T}$ das lineare Modell mit funktionaler Zielgröße, wobei $\mathcal{U}_1(t), \ldots, \mathcal{U}_n(t)$ u. i. v. sind mit $\mathsf{E}(\mathcal{U}_i(t)) = 0 \; \forall \; t \in \mathcal{T}$ und $\mathsf{E}(\mathcal{U}_i(s)\mathcal{U}_i(t)) = \sigma(s, t)$. Es sei angemerkt, dass gerade die Annahme der Unab-

hängigkeit der Residuen im funktionalen Kontext sorgfältig zu überprüfen ist. Da $\boldsymbol{\mathcal{F}}(t)$ unbekannt ist (Def. 3.12), wird $\boldsymbol{\mathcal{F}}(t)$ wie in Abschnitt 3.2 durch $\widetilde{\boldsymbol{\mathcal{F}}}(t)$ parametrisch geschätzt. Die Modellgleichung der *function-on-scalar regression* lautet dann

$$\widetilde{\boldsymbol{\mathcal{F}}}(t) = \mathbf{X}\boldsymbol{\beta}(t) + \boldsymbol{\mathcal{U}}(t), \qquad t \in \mathcal{T}.$$

Nicht nur $\boldsymbol{\mathcal{F}}(t)$ muss mit Hilfe von Basisfunktionen dargestellt werden, sondern auch $\boldsymbol{\beta}(t)$. Seien $\boldsymbol{\phi}(t) = (\phi_0(t), \ldots, \phi_{K_y}(t))^\top$ und $\boldsymbol{\varphi}(t) = (\varphi_0(t), \ldots, \varphi_{K_\beta}(t))^\top$ parametrische Basissysteme. Dann lassen sich die Zielgrößen und Regressionsfunktionen schreiben als $\widetilde{\boldsymbol{\mathcal{F}}}(t) = \widetilde{\boldsymbol{\Delta}}\boldsymbol{\phi}(t)$ bzw. $\boldsymbol{\beta}(t) = \boldsymbol{\Psi}\boldsymbol{\varphi}(t)$, wobei $\widetilde{\boldsymbol{\Delta}} \in \mathbb{R}^{n \times K_y+1}$ und $\boldsymbol{\Psi} \in \mathbb{R}^{p \times K_\beta+1}$. Die gängigen Annahmen $K_y = K_\beta =: K$ und $\boldsymbol{\phi}(t) = \boldsymbol{\varphi}(t)$ werden im Folgenden als gegeben angesehen. Zu schätzen ist also die Matrix $\boldsymbol{\Psi}$ in

$$\widetilde{\boldsymbol{\mathcal{F}}}(t) = \widetilde{\boldsymbol{\Delta}}\boldsymbol{\phi}(t) = \mathbf{X}\boldsymbol{\Psi}\boldsymbol{\phi}(t) + \boldsymbol{\mathcal{U}}(t), \qquad t \in \mathcal{T}.$$

Der KQ-Schätzer für funktionale Regressionskoeffizientenvektoren ergibt sich über

$$\widehat{\boldsymbol{\Psi}} = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \int_{\mathcal{T}} \|\widetilde{\boldsymbol{\mathcal{F}}}(t) - \mathbf{X}\boldsymbol{\Psi}\boldsymbol{\phi}(t)\|^2 dt,$$

siehe Ramsay und Silverman (2005, S. 236 f.), oder in geschlossener Form:

$$\operatorname{vec}\left(\widehat{\boldsymbol{\Psi}}\right) = \left[\left(\mathbf{X} \otimes \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{1/2}\right)^\top \left(\mathbf{X} \otimes \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{1/2}\right)\right]^{-1} \left(\mathbf{X} \otimes \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{1/2}\right)^\top \operatorname{vec}\left(\mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{1/2} \widetilde{\boldsymbol{\Delta}}^\top\right), \qquad (3.11)$$

wobei $\mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}} = \left[\int_{\mathcal{T}} \phi_i(t)\phi_j(t)dt\right]_{i,j} \in \mathbb{R}^{K \times K}$, siehe Reiss *et al.* (2010, Formel (13) mit $\lambda_i = 0 \; \forall \; i$) und $\mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{1/2}$ die Quadratwurzel der Matrix $\mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}$ ist. Somit folgt

$$\widehat{\widetilde{\boldsymbol{\mathcal{F}}}}(t) = \mathbf{X}\widehat{\boldsymbol{\Psi}}\boldsymbol{\phi}(t), \qquad t \in \mathcal{T}.$$

Wenn vermieden werden soll, dass stark schwankende Funktionen aus $\boldsymbol{\phi}(t)$ ein großes Gewicht zur Schätzung von $\widetilde{\boldsymbol{\mathcal{F}}}$ erhalten, bieten sich regularisierte Basisfunktionen an. Der pönalisierte KQ-Schätzer lautet dann

$$\widehat{\boldsymbol{\Psi}} = \underset{\boldsymbol{\Psi}}{\operatorname{argmin}} \int_{\mathcal{T}} \|\widetilde{\boldsymbol{\mathcal{F}}}(t) - \mathbf{X}\boldsymbol{\Psi}\boldsymbol{\phi}(t)\|^2 dt + \int_{\mathcal{T}} \|\boldsymbol{\Lambda} L \boldsymbol{\Psi}\boldsymbol{\phi}(t)\|^2 dt,$$

mit der Glättungsmatrix $\mathbf{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p), \lambda_i \geq 0$ und einem linearen Differentialoperator $L$. Die Darstellung in geschlossener Form (vgl. Reiss *et al.*, 2010, Formel (13)) ist ähnlich zu Formel (3.11) und kann als funktionale Erweiterung des Ridge-Regressionsschätzers (Hoerl und Kennard, 1970) angesehen werden:

$$\text{vec}\left(\widehat{\mathbf{\Psi}}\right) = \left[\left(\mathbf{X} \otimes \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{1/2}\right)^\top \left(\mathbf{X} \otimes \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{1/2}\right) + \mathbf{P}_{\mathbf{\Lambda}}\right]^{-1} \left(\mathbf{X} \otimes \mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{1/2}\right)^\top \text{vec}\left(\mathbf{J}_{\boldsymbol{\phi}\boldsymbol{\phi}}^{1/2} \widetilde{\mathbf{\Delta}}^\top\right),$$

wobei $\mathbf{P}_\Lambda = \mathbf{\Lambda} \otimes \mathbf{R}$ und $\mathbf{R} = \left[\int_{\mathcal{T}} (L\phi_i(t))(L\phi_j(t))dt\right]_{i,j} \in \mathbb{R}^{K \times K}$. In der Regel wird für $L$ die zweite Ableitung verwendet, so dass

$$L\mathbf{\Psi}\boldsymbol{\phi}(t) = \begin{pmatrix} \sum_{i=1}^K \mathbf{\Psi}_{1,i+1}\phi_i''(t) \\ \vdots \\ \sum_{i=1}^K \mathbf{\Psi}_{p,i+1}\phi_i''(t) \end{pmatrix}$$

gilt. Untersuchungen von Wood (2006, 2011) haben ergeben, dass die Bestimmung der Glättungsparameter mittels gewöhnlicher Kreuzvalidierung sehr rechenintensiv ist. Eine Alternative dazu ist die generalisierte Kreuzvalidierung (GCV), die weniger rechenintensiv und zudem invariant gegenüber orthogonalen Transformationen ist (Wood, 2006). Des Weiteren ist eine restringierte ML-Schätzung (REML) der Glättungsparameter möglich. Anhand von Simulationen zeigt Wood (2011), dass eine REML-Schätzung der Glättungsparameter einen kleineren mittleren quadratischen Fehler als die GCV-Schätzung ebendieser hat und weniger stark zur Überanpassung neigt. Folglich empfiehlt Wood (2011) den REML-Schätzer für dieses Problem. Ein hier bisher nicht diskutiertes Problem ist die u. i. v.-Annahme der Residuen. Bei Verletzung der Annahme ist nicht mehr der KQ-Schätzer bester linearer unverzerrter Schätzer, sondern der verallgemeinerte KQ-Schätzer (VKQ-Schätzer). Im nicht-funktionalen Kontext basiert dieser auf der Kenntnis der positiv definiten Kovarianzmatrix $\text{Cov}(\boldsymbol{\mathcal{U}}) = \mathbf{\Sigma}$ und löst unter Anwendung der Cholesky-Zerlegung $\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2} = \mathbf{\Sigma}^{-1}$ das Minimierungsproblem

$$\widehat{\boldsymbol{\beta}}^{VKQ} = \underset{\boldsymbol{\beta}}{\text{argmin}} \, \|\mathbf{\Sigma}^{-1/2}\mathbf{y} - \mathbf{\Sigma}^{-1/2}\mathbf{X}\boldsymbol{\beta}\|^2$$

durch

$$\widehat{\boldsymbol{\beta}}^{VKQ} = (\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}. \tag{3.12}$$

In der Praxis muss $\boldsymbol{\Sigma}$ allerdings geschätzt werden. Der *feasible* VKQ-Schätzer setzt die geschätzte Kovarianzmatrix $\widehat{\boldsymbol{\Sigma}}$ in (3.12) ein:

$$\widehat{\boldsymbol{\beta}}^{fVKQ} = (\mathbf{X}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{X}) \mathbf{X}^\top \widehat{\boldsymbol{\Sigma}}^{-1} \mathbf{y},$$

wobei noch sichergestellt werden muss, dass $\widehat{\boldsymbol{\Sigma}}$ positiv definit ist. In funktionalen Datensituationen wie der Varianzanalyse mit funktionalen Zielgrößen können die Untersuchungseinheiten einer Gruppe homogen strukturierte Residuenfunktionen aufweisen, was zur Verletzung der u. i. v.-Annahme führt. Reiss *et al.* (2010) schlagen deshalb einen *feasible* verallgemeinerten, pönalisierten KQ-Schätzer für FOSR vor. Die Kovarianzmatrix wird über ein iteratives Verfahren geschätzt. Dabei wird zunächst die Kovarianzmatrix der Residuen der pönalisierten FOSR geschätzt. Mit dieser Kovarianzmatrix wird der fVKQ-Schätzer

$$\begin{aligned} \operatorname{vec}\left(\widehat{\boldsymbol{\Psi}}\right) = &\left[\left(\mathbf{X} \otimes \left(\widehat{\boldsymbol{\Sigma}}^{-1/2} \mathbf{J}_{\phi\phi}^{1/2}\right)\right)^\top \left(\mathbf{X} \otimes \left(\widehat{\boldsymbol{\Sigma}}^{-1/2} \mathbf{J}_{\phi\phi}^{1/2}\right)\right) + \mathbf{P}_{\boldsymbol{\Lambda}}\right]^{-1} \\ &\times \left(\mathbf{X} \otimes \left(\widehat{\boldsymbol{\Sigma}}^{-1/2} \mathbf{J}_{\phi\phi}^{1/2}\right)\right)^\top \operatorname{vec}\left(\left(\widehat{\boldsymbol{\Sigma}}^{-1/2} \mathbf{J}_{\phi\phi}^{1/2}\right) \widetilde{\boldsymbol{\Delta}}^\top\right). \end{aligned}$$

bestimmt, woraus wiederum eine neue geschätzte Kovarianzmatrix hervorgeht. Diese Prozedur wird wiederholt, bis das Verfahren konvergiert. Der fVKQ-Schätzer weist ein im Gegensatz zum pönalisierten KQ-Schätzer leicht verbessertes Verhalten bei Verletzungen der u. i. v.-Annahme auf (Reiss *et al.*, 2010). Die Verletzung der Annahme normalverteilter Zielgrößen wird im nächsten Abschnitt behandelt.

### *3.2.2 Generalisierte Function-on-Scalar Regression*

Im Kontext der FOSR existieren unterschiedliche Ansätze, um nicht-normalverteilte Zielgrößen zu integrieren. Dabei bezieht sich der Verteilungstyp der funktionalen

Zielgröße auf die Verteilung in jedem Zeitpunkt. Hall *et al.* (2008) nutzen funktionale Hauptkomponentenanalyse (siehe Anhang A.2), um longitudinale, nichtnormalverteilte Zielgrößen mittels latenter Gaußprozesse vorherzusagen:

$$\begin{aligned}\mathsf{E}(\mathcal{Y}(t_1)\dots\mathcal{Y}(t_T)|\mathcal{X}) &= \prod_{j=1}^{T} h(\mathcal{X}(t_j)),\\ \mathsf{E}\left(\mathcal{Y}(t)^2|\mathcal{X}\right) &\leq h_1(\mathcal{X}(t)),\end{aligned}$$

wobei $t_1,\dots,t_m \in \mathcal{T},\ \mathcal{X}$ ein Gaußprozess auf $\mathcal{T},\ h$ eine monoton steigende, hinreichend glatte Responsefunktion und $h_1$ eine beschränkte Funktion ist. Dieses Verfahren sollte jedoch nicht für dicht beobachtbare funktionale Zielgrößen verwendet werden, da es Unabhängigkeit zwischen $\mathcal{Y}(t_i)$ und $\mathcal{Y}(t_j), i \neq j$ voraussetzt.

Ein bayesscher Ansatz von Goldsmith *et al.* (2015) erweitert generalisierte hierarchische lineare Modelle um funktionale Kovariablen und Zielgrößen:

$$\mathsf{E}(\mathcal{Y}_{i\ell}(t)|b_i(t), v_{i\ell}(t)) = h\left(\beta_0(t) + \sum_{j=1}^{p} x_{i\ell,j}(t)\beta_j(t) + b_i(t) + v_{i\ell}(t)\right),$$

wobei $b_i(t)$ *subject-specific random deviation*, $v_{i\ell}(t)$ *subject- and visit-specific random deviation* sowie $h$ eine Responsefunktion bezeichnen. Mit Hilfe der Basisfunktionen der hierarchischen funktionalen Hauptkomponentenanalyse werden pönalisierte Splines bestimmt, deren Koeffizienten gemeinsam mit den PC Scores und PC Spline-Koeffizienten in einer bayesschen Analyse geschätzt werden.

Eine Variante generalisierter additiver Modelle für funktionale Kovariablen und Zielgrößen schlagen Scheipl *et al.* (2016) vor. Die zugehörige Modellgleichung lautet

$$\mathsf{E}(\boldsymbol{\mathcal{F}}(t)|\mathbf{X}, t, \boldsymbol{\nu}) = h\left(\sum_{j=1}^{p} f_j(\mathbf{X}_{\bullet j}, t)\right),$$

wobei $\boldsymbol{\nu}$ ein optionaler Vektor mit weiteren Parametern der Verteilung von $\boldsymbol{\mathcal{F}}(t)$ ist und $f_j, j = 1,\dots,p$ unbekannte, zu schätzende, glatte Funktionen. Die obige Formel kann vereinfacht werden in ein generalisiertes lineares Modell mit funktionalen Zielgrößen, wenn die Funktionen $f_j$ linear sind. Dann ist die Schreibweise des

linearen Prädiktors analog zur nicht-generalisierten Version:

$$\mathsf{E}\left(\boldsymbol{\mathcal{F}}(t)|\mathbf{X}, t, \boldsymbol{\nu}\right) = h\left(\mathbf{X}\boldsymbol{\Psi}\boldsymbol{\phi}(t)\right).$$

Wegen der Generalisiertheit des Modells ist die KQ-Schätzung der Parameter nicht sinnvoll. Es existiert jedoch ein restringierter Maximum-Likelihood-Ansatz von Scheipl *et al.* (2016), der kurz vorgestellt wird. Die Idee dieses Ansatzes ist, den multi-funktionalen Zufallsprozess $\boldsymbol{\mathcal{F}}(t)$ als diskretisierten Zufallsvektor $\boldsymbol{\mathcal{Y}}$ zu schreiben: Seien unter der Vernachlässigung der Achsenabschnitt-Terme $\boldsymbol{\mathcal{Y}}_i = (\mathcal{F}_i(t_1), \ldots, \mathcal{F}_i(t_T))^\top, i = 1, \ldots, n, \boldsymbol{\mathcal{Y}} = (\boldsymbol{\mathcal{Y}}_1^\top, \ldots, \boldsymbol{\mathcal{Y}}_n^\top)^\top \in \mathbb{R}^{nT}, \boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{\Psi}) \in \mathbb{R}^{pK}$ und $K$ die Anzahl der Basisfunktionen je Prädiktor. Um ein Modell der Form

$$\mathsf{E}(\boldsymbol{\mathcal{Y}}|\mathbf{X}, t, \boldsymbol{\nu}) = h(\mathbf{Z}(\mathbf{X})\boldsymbol{\beta}) \tag{3.13}$$

aufzustellen, muss $\mathbf{X} \in \mathbb{R}^{n\times p}$ in die erweiterte Designmatrix $\mathbf{Z}(\mathbf{X}) \in \mathbb{R}^{nT\times pK}$ umgeformt werden, was durch das Zeilentensorprodukt geschieht (siehe Herleitung von (3.14) im Anhang A.3). Sei $\mathbf{X} = (x_{ij})_{i=1,\ldots,n;j=1,\ldots,p}$ gegeben. Dann folgt für die erweiterte Designmatrix der generalisierten *function-on-scalar regression* (GFOSR) mit

$$\mathbf{Z}_j(\mathbf{X}) = \begin{pmatrix} x_{1j}\phi_1(t_1) & \cdots & x_{1j}\phi_K(t_1) \\ \vdots & & \vdots \\ x_{1j}\phi_1(t_T) & \cdots & x_{1j}\phi_K(t_T) \\ \vdots & & \vdots \\ x_{nj}\phi_1(t_1) & \cdots & x_{nj}\phi_K(t_1) \\ \vdots & & \vdots \\ x_{nj}\phi_1(t_T) & \cdots & x_{nj}\phi_K(t_T) \end{pmatrix} \in \mathbb{R}^{nT\times K}, \text{ dass} \tag{3.14}$$

$$\mathbf{Z}(\mathbf{X}) = (\mathbf{Z}_1(\mathbf{X}), \ldots, \mathbf{Z}_p(\mathbf{X})).$$

In (3.13) ist der unbekannte Parametervektor $\boldsymbol{\beta}$ zu schätzen, welcher die Koeffizienten der Linearkombinationen der Elemente des Basisfunktionensystems enthält. Aus der GFOSR wird somit ein generalisiertes lineares Modell, wobei die Zielgrößen – ähnlich wie in gemischten Modellen – gegeben $t$ als unabhängig angesehen

werden. Die Verteilung der Zielgröße $\mathcal{Y}_{ij}$ gehört zur Exponentialverteilungsfamilie. Weiterhin sind die üblichen ML-Schätzer verwendbar. Die Log-Likelihood bezüglich der Verteilung $\mathcal{F}(\mu_{ij}, \boldsymbol{\nu})$ von $\mathcal{Y}_{ij}$ lautet

$$l(\boldsymbol{\mu}, \boldsymbol{\nu}|\mathbf{y}) = \sum_{i=1}^{n} \sum_{j=1}^{T} l(\mu_{ij}, \boldsymbol{\nu}|y_{ij}),$$

wobei $\boldsymbol{\nu}$ ein nicht interessierender Parameter ist. Auch in dieser Darstellung ist der ML-Schätzer anfällig für Überanpassung. Aus diesem Grund wird die Log-Likelihood durch einen Strafterm erweitert, der starke Schwankungen der jeweiligen (diskretisierten) Regressionsfunktionen $\mathbf{Z}_j(\mathbf{X})\boldsymbol{\beta}_j$ pönalisiert:

$$l_p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\nu}|\mathbf{y}) = l(\boldsymbol{\mu}, \boldsymbol{\nu}|\mathbf{y}) - \frac{1}{2} \sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j^\top \mathbf{P}_j \boldsymbol{\beta}_j, \tag{3.15}$$

wobei $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$ der Vektor der positiven Glättungsparameter, $\lambda_j \boldsymbol{\beta}_j^\top \mathbf{P}_j \boldsymbol{\beta}_j$ der Strafterm und $\mathbf{P}_j$ mit $p_{j,ik} = \int \phi''_{j,i}(t)\phi''_{j,k}(t)dt$ definiert sind. Häufig wird für alle skalaren Kovariablen das gleiche Basissystem angenommen, dann gilt $\mathbf{P} = \mathbf{P}_j \; \forall \; j = 1, \ldots, p$. Für die gleichzeitige Maximierung von (3.15) bzgl. $\boldsymbol{\beta}$ und $\boldsymbol{\lambda}$ schlagen Wood *et al.* (2016) eine Laplace-Approximation der marginalen Likelihood der Glättungsparameter vor. Hierbei wird iterativ bzgl. $\boldsymbol{\lambda}$ und dann bzgl. $\boldsymbol{\beta}$ gegeben $\boldsymbol{\lambda}$ optimiert, siehe Scheipl *et al.* (2016). Die Laplace-Approximation der Log-Likelihood lautet

$$\begin{aligned} l_{LA}(\boldsymbol{\lambda}, \boldsymbol{\nu}|\mathbf{y}) = l(\tilde{\boldsymbol{\mu}}, \boldsymbol{\nu}|\mathbf{y}) - \frac{1}{2} \sum_{j=1}^{p} \lambda_j \widetilde{\boldsymbol{\beta}}_j^\top \mathbf{P}_j \widetilde{\boldsymbol{\beta}}_j + \frac{1}{2} \ln \left| \sum_{j=1}^{p} \lambda_j \mathbf{P}_j \right|^+ \\ - \frac{\ln |\mathbf{H}|}{2} + \frac{d_\emptyset \ln(2\pi)}{2}, \end{aligned}$$

wobei $\tilde{\boldsymbol{\mu}} = h(\mathbf{Z}(\mathbf{X})\widetilde{\boldsymbol{\beta}}), \widetilde{\boldsymbol{\beta}} = \operatorname{argmax}_{\boldsymbol{\beta}} l_p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\nu}|\mathbf{y})$, $\mathbf{H} = -\frac{\partial^2 l_p(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\nu}|\mathbf{y})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}$ die negative Hessematrix, $|\cdot|^+$ das Produkt der positiven Eigenwerte von $\cdot$ und $d_\emptyset$ die algebraische Vielfachheit des Eigenwerts 0 der Matrix $\sum_{j=1}^{p} \lambda_j \mathbf{P}_j, j = 1, \ldots, p$ bezeichnen. In R ist GFOSR in der Funktion `pffr` im Paket `refund` (Goldsmith *et al.*, 2016) implementiert. Für die konkrete Maximierung der Laplace-Approximation der Log-Likelihood nutzt `refund` Funktionen aus dem Paket `mgcv` (Wood, 2011).

Auch im GLM mit funktionalen Zielgrößen ist die (typischerweise grafische) Überprüfung, ob die Annahmen an das Modell als gerechtfertigt angesehen werden können, ein wichtiger Aspekt. So kann mittels Streudiagrammen überprüft werden, ob die geschätzten Residuen abgetragen gegen ihren Index ein Muster aufweisen, was gegen die Annahme einer konstanten Varianz spricht. Im Bezug auf funktionale Zielgrößen muss jedoch beachtet werden, dass zum Beispiel das Streudiagramm der geschätzten Zielgrößen gegen die geschätzten Residuen für jedes $t_i, i = 1, \ldots, T$ separat betrachtet werden muss (Faraway, 1997). Für großes $T$ ist dieses Vorgehen nicht mehr handhabbar, weshalb eine Alternative in Abschnitt 5.3 vorgestellt wird.

### *3.2.3 Weitere Modellierungsansätze funktionaler Daten*

Beiträge zur Modellierung funktionaler Daten, die in dieser Arbeit nicht thematisiert werden, obwohl sie eine gewisse inhaltliche Nähe zu dieser Arbeit aufweisen, umfassen generalisierte lineare Modelle mit funktionalen Kovariablen (James, 2002; Müller und Stadtmüller, 2005). Solche Modelle existieren sowohl für skalare als auch für funktionale Zielgrößen. Bei skalaren Zielgrößen muss das Produkt aus Kovariable und Regressionsfunktion über der gemeinsamen Indexmenge integriert werden. Bei funktionalen Zielgrößen gibt es zwei Methoden: Einerseits einen simultanen Ansatz *(concurrent model)*, bei dem Zielgröße, Kovariablen und Regressionsfunktion dieselbe Indexmenge haben (Ramsay und Silverman, 2005, Kap. 14); andererseits ein bi-funktionales Modell, bei dem Zielgröße und Kovariablen potenziell unterschiedliche Indexmengen besitzen und die Regressionsfunktionen auf dem kartesischen Produkt der beiden Indexmengen definiert sind (Ramsay und Silverman, 2005, Kap. 16). Durch die Integration über die Indexmenge der Kovariablen erhält man den linearen Prädiktor. Die Besonderheit des simultanen Modells ist, dass $x(t)$ nur auf $y(t)$ einen Einfluss ausübt, nicht jedoch auf $y(\tilde{t}), \tilde{t} \neq t$. Selbst bei identischen Indexmengen im bi-funktionalen Modell ist das nicht der Fall. Hier hat $x(t)$ einen Einfluss auf $y(\tilde{t}) \; \forall \; \tilde{t} \in \mathcal{T}$. Insbesondere durch den aus chronologischer Perspektive kontraintuitiven Einfluss von $x(t)$ auf $y(\tilde{t})$ mit $t > \tilde{t}$ ist das historische funktionale lineare Modell (Malfait und Ramsay, 2003) motiviert, das nur Einflüsse von $x(t)$ auf

$y(\tilde{t})$ zulässt, wenn $t \leq \tilde{t}$.

Zur statistischen Inferenz in (generalisierten) linearen Modellen mit funktionalen Zielgrößen und/oder Kovariablen lassen sich nur selten exakte oder asymptotische Resultate gewinnen, da das Konzept der Wahrscheinlichkeitsdichte im funktionalen Kontext nicht ohne Weiteres definiert ist (Delaigle und Hall, 2010). Daraus resultiert vor allem für Tests wie den $F$-Test im GLM mit funktionalen Kovariablen oder Anpassungstests von Verteilungen (Cuesta-Albertos *et al.*, 2007) die Schwierigkeit, dass die Verteilungen der jeweiligen Teststatistiken nicht (oder nicht explizit) bestimmt werden können. Mögliche Auswege sind wie in der endlichdimensionalen Statistik die Verwendung von Permutationstests oder die Bestimmung von approximativen p-Werten über Bootstrap-Verfahren. Eine weitere Anwendung von Bootstrap-Verfahren im funktionalen Kontext wird im Bootstrap-Algorithmus, Abschnitt 5.1 beschrieben.

Die folgenden beiden Kapitel beschäftigen sich mit der Identifikation von Ausreißern in den bisher eingeführten Modellen. Bevor in Kapitel 5 die in den Abschnitten 3.2.1 und 3.2.2 vorgestellten Modelle als Grundlage zur Identifikation funktionaler Ausreißer verwendet werden, wird im folgenden Kapitel 4 basierend auf den Modellen aus Abschnitt 3.1 die Ausreißeridentifikation in Kontingenztafeln thematisiert.

# 4. AUSREISSERIDENTIFIKATION IN KONTINGENZTAFELN

In Abschnitt 2.1 wurden $\alpha$-Ausreißerregionen für bestimmte unabhängig identisch verteilte Zufallsvariablen behandelt. Eine andere Situation liegt vor, wenn Daten in Form einer Kontingenztafel erhoben wurden. Für die Zellhäufigkeiten kann die Annahme der identischen Verteilung nicht aufrechterhalten werden. Im Zentrum dieses Kapitels werden Ausreißeridentifizierer analysiert, welche speziell für die Erkennung einzelner Zellen sowie ganzer Tafeln als Ausreißer konstruiert wurden.

In loglinearen Poissonmodellen (siehe Kap. 3.1) wird jede Zelle einer Tafel als Realisation einer poissonverteilten Zufallsvariable aufgefasst. Daraus lässt sich für jede Zelle eine $\alpha$-Ausreißerregion ableiten. Im loglinearen Poissonmodell, welches in Abschnitt 4.1 betrachtet wird, können damit keine, eine oder mehrere Zellen als $\alpha$-Ausreißer identifiziert werden.

In loglinearen Multinomialmodellen wird die vollständige Tafel in Vektorschreibweise als Realisation einer multinomialverteilten Zufallsvariable aufgefasst, kurz: $\boldsymbol{\mathcal{Y}} \sim Mult(n, \boldsymbol{\theta})$. Dann bezieht sich eine $\alpha$-Ausreißerregion auf die ganze Tafel, wodurch die ganze Tafel als $\alpha$-Ausreißer klassifiziert wird oder nicht. In Abschnitt 4.2 wird näher auf diesen Fall eingegangen.

## 4.1 IDENTIFIKATION VON ZELLEN ALS AUSREISSER

Gegeben seien die Zellhäufigkeiten $y_i, i = 1, \ldots, k$ einer Kontingenztafel. Wenn die Gesamtanzahl an Beobachtungen $n = \sum_{i=1}^{k} y_i$ *a priori* nicht bekannt war, können die Einträge von $\mathbf{y} = (y_1, \ldots, y_k)^\top$ als Realisationen unabhängiger poissonverteilter Zufallsvariablen angesehen werden, kurz: $\mathcal{Y}_i \sim Poi(\theta_i), i = 1, \ldots, k$. Die Schätzung des unbekannten Parametervektors $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^\top$ erfolgt mit den aus Ab-

schnitt 3.1.2 bekannten Methoden. Die Wahrscheinlichkeit $P(\mathcal{Y}_i = y_i \,|\, \theta_i = \hat{\theta}_i)$ kann bestimmt werden und damit über das Konzept der $\alpha$-Ausreißer Zellhäufigkeiten als $\alpha$-Ausreißer klassifiziert werden. Für eine formale Definition von $\alpha$-Ausreißern in Kontingenztafeln wird auf Kuhnt (2004) zurückgegriffen.

DEFINITION 4.1. *Eine beobachtete Zellhäufigkeit $y_i$ wird* $\alpha$-Ausreißer *in Bezug auf ein loglineares Poissonmodell genannt, falls sie sich in der $\alpha$-Ausreißerregion*

$$\mathrm{out}(\alpha, Poi(\theta_i)) = \{y \in \mathbb{N}_0 : P(\mathcal{Y}_i = y) < B(\alpha)\},$$

*befindet, wobei $\mathcal{Y}_i \sim Poi(\theta_i)$, $\alpha \in ]0,1[$ und*

$$B(\alpha) = \sup\{K > 0 : \sum_{y \in \mathbb{N}_0} P(\mathcal{Y}_i = y)\mathbf{1}_{[0,K]}(P(\mathcal{Y}_i = y)) \leq \alpha\}$$

*mit der Indikatorfunktion $\mathbf{1}_A(x)$.*

Mit Hilfe dieser Begriffe und geeigneter (vorzugsweise robuster) Schätzer können für poissonverteilte Zufallsvariablen einschrittige $\alpha$-Ausreißeridentifizierer wie in Definition 2.3 hergeleitet werden. Zum Vergleich von auf verschiedenen Schätzern basierenden Ausreißeridentifizierern werden zwei vom Bruchpunkt eines Schätzers (Def. 3.8) abgeleitete Gütekriterien vorgestellt. Dabei unterscheidet man zwischen der Intensität der $\delta$-Ausreißer und der Sensitivität des $\alpha$-Ausreißeridentifizierers.

DEFINITION 4.2. Masking-Bruchpunkt eines Ausreißeridentifizierers *(Kuhnt, 2000, Def. 5.5)*
*Seien $\delta, \alpha \in ]0,1[$ und $\mathrm{OI}(\cdot; \alpha, P_{\widehat{\boldsymbol{\theta}}})$ ein gegebener $\alpha$-Ausreißeridentifizierer. Sei weiter in der Situation von Definition 3.6 $\mathbf{y}$ eine Realisation des Zufallsvektors $\boldsymbol{\mathcal{Y}}$. Bezeichne $\mathbf{y}^{\mathfrak{z}}$ einen beliebigen Vektor, der durch Ersetzen von $\mathfrak{z}$ Beobachtungen $y_{\mathbf{i}}$ durch Werte $y_{\mathbf{i}}^0 \in \mathrm{out}(\delta, Poi(\theta_{\mathbf{i}}))$ erzeugt wird. Dann ist*

$$\begin{aligned}
&\xi^{\mathfrak{M}}(\mathrm{OI}, \alpha, \mathbf{y}, \mathfrak{z}, \delta) \\
&:= \inf\{\xi > 0 : \exists\, \mathbf{y}^{\mathfrak{z}},\ \textit{der eine Beobachtung } y_{\mathbf{i}}^{\mathfrak{z}} \textit{ enthält mit } y_{\mathbf{i}}^{\mathfrak{z}} \in \mathrm{out}(\xi, Poi(\theta_{\mathbf{i}})), \\
&\qquad \textit{die von } \mathrm{OI} \textit{ auf der Basis von } \mathbf{y}^{\mathfrak{z}} \textit{ nicht als } \alpha\textit{-Ausreißer identifiziert wird}\}
\end{aligned}$$

*der* Masking-Punkt *und*

$$\varepsilon^{\mathfrak{M}}(\mathrm{OI}, \alpha, \mathbf{y}, \delta) := \frac{\min\{\mathfrak{z} | \xi^{\mathfrak{M}}(\mathrm{OI}, \alpha, \mathbf{y}, \mathfrak{z}, \delta) = 0\}}{k}$$

*der* Masking-Bruchpunkt des $\alpha$-Ausreißeridentifizierers OI.

Der *Swamping-Bruchpunkt eines Ausreißeridentifizierers* $\varepsilon^{\mathfrak{S}}$ ist definiert durch den kleinsten Anteil von $\delta$-Ausreißern, der dazu führen kann, dass Beobachtungen, die keine $\alpha$-Ausreißer sind, als $\alpha$-Ausreißer klassifiziert werden. Die Verwendung robuster Schätzer sollte gegenüber nicht-robusten Schätzern die oben genannten Bruchpunkte des Ausreißeridentifizierers verbessern. Ein Ausreißeridentifizierer, der auf einem Schätzer mit einer gewissen Robustheit im loglinearen Unabhängigkeitsmodell beruht, wird in Definition 4.3 vorgestellt.

DEFINITION 4.3. *Sei* $\alpha \in ]0, 1[$ *gegeben. Ein* einschrittiger $\alpha$-Ausreißeridentifizierer basierend auf dem $L_1$-Schätzer (OL1-Identifizierer) *im loglinearen Unabhängigkeitsmodell sei durch folgende Prozedur definiert:*

*(i) Schätze* $\theta_i, i = 1, \ldots, k$*, für das loglineare Modell basierend auf der kompletten Kontingenztafel mittels des* $L_1$*-Schätzers.*

*(ii) Identifiziere Zellhäufigkeiten* $y_i \in \mathrm{out}(\alpha, Poi(\hat{\theta}_i))$ *als Ausreißer.*

Wie in Bemerkung 2.4 angedeutet, kann $\alpha$ bezüglich $k$ adjustiert werden. Im Folgenden wird diese Adjustierung zugunsten einer höheren Richtig-Positiv-Rate jedoch nicht verwendet. In Abschnitt 4.1.1 werden auf Minimalmustern basierende Ausreißeridentifizierer hergeleitet, die in den Abschnitten 4.1.2 und 4.1.3 mit dem OL1-Identifizierer verglichen werden.

### *4.1.1 Minimalmusterverfahren*

In diesem Abschnitt wird ein Verfahren vorgestellt, um eine oder mehrere Zellen einer Kontingenztafel als $\alpha$-Ausreißer zu identifizieren. Dieses Verfahren basiert

auf dem loglinearen Poissonmodell mit $\mathsf{E}(\boldsymbol{\mathcal{Y}}|\mathbf{Z}) = \exp(\mathbf{X}\boldsymbol{\beta})$ (Def. 3.6), in dem die Einträge einer Kontingenztafel als Zielgrößen fungieren und die Designmatrix $\mathbf{X}$ anhand der Dimensionen der Kontingenztafel und der angenommenen Abhängigkeitsstruktur der Merkmale konstruiert wird. Wie in (3.5) bzw. (A.2) am Beispiel einer $3 \times 3$-Tafel veranschaulicht wurde, korrespondiert jede Zeile der Designmatrix zu einem Eintrag in der Kontingenztafel. Falls ein Ausreißer in der $i^*$-ten Stelle von $\mathbf{y}$ vermutet wird, kann $\boldsymbol{\beta}$ auch ohne $y_{i^*}$ geschätzt werden, indem die $i^*$-te Zeile von $\mathbf{X}$ entfernt wird. Die reduzierte Schätzfunktion lautet dann $\widehat{\boldsymbol{\beta}} = T(\{(\mathbf{X}_{i\bullet}, \mathcal{Y}_i), i = 1, \ldots, i^* - 1, i^* + 1, \ldots, k\}, g) =: T(\{(\underline{\mathbf{X}}_{i\bullet}, \underline{\mathcal{Y}}_i), i = 1, \ldots, k-1\}, g)$. Falls mehrere Ausreißer vermutet werden, können auch mehrere Zeilen von $\mathbf{X}$ entfernt werden. Das sogenannte Minimalmusterverfahren basiert auf solchen reduzierten Schätzfunktionen. Es legt die Annahme zugrunde, dass die Mehrheit der Zellhäufigkeiten einem gemeinsamen loglinearen Poissonmodell folgt. Stark davon abweichende Werte sollen als Ausreißer identifiziert werden. Ein Minimalmuster ist nach Kuhnt (2000) sowie Kuhnt *et al.* (2014) wie folgt definiert:

DEFINITION 4.4. Minimalmuster
*Sei $\mathbf{X} \in \mathbb{R}^{k \times p}$ die Designmatrix eines loglinearen Poissonmodells. Eine Teilmenge von Zellen der zugehörigen Kontingenztafel wird* Minimalmuster *genannt, falls*

*(i) sie mindestens $\left\lfloor \frac{k}{2} \right\rfloor + 1$ Elemente enthält;*

*(ii) für die korrespondierende Matrixpartition $\underline{\mathbf{X}}$ gilt, dass $Rg(\underline{\mathbf{X}}) = p$ und*

*(iii) die Teilmenge die minimale Anzahl von Elementen besitzt, die für die Erfüllung von (i) und (ii) nötig sind.*

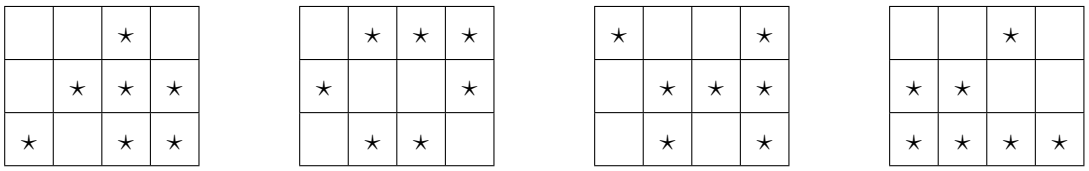
*Falls eine $p$-elementige Teilmenge von Zellen nur Bedingung (ii) erfüllt, ist durch sie ein* strenges Minimalmuster *definiert. Des Weiteren wird die Menge aller Minimalmuster mit $\mathbb{M} = \{\mathbb{M}_1, \ldots, \mathbb{M}_V\}$ bezeichnet und die Menge aller strengen Minimalmuster mit $\mathbb{S} = \{\mathbb{S}_1, \ldots, \mathbb{S}_{V^{str}}\}$ bezeichnet.*

Bei einem strengen Minimalmuster handelt es sich um die kleinstmögliche Teilmenge einer Kontingenztafel, mit der die Parameter des loglinearen Poissonmodells

noch eindeutig schätzbar sind. Minimalmuster beinhalten hingegen mindestens die Hälfte der Einträge der Kontingenztafel. Die Idee hinter diesem Verfahren ist, jedes einzelne Minimalmuster als potenziell ausreißerfreie Teilmenge der Kontingenztafel anzusehen.

BEISPIEL 4.5. Minimalmuster einer $3 \times 4$ Tafel

*Es sei eine Kontingenztafel mit den Merkmalen $\mathcal{D}_1$ und $\mathcal{D}_2$ gegeben. $\mathcal{D}_1$ habe drei und $\mathcal{D}_2$ habe vier mögliche Ausprägungen. Die beiden Merkmale werden als unabhängig angenommen. Ausgehend davon bestehen die Minimalmuster dieser $3 \times 4$ Tafel aus sieben Zellen. Unten finden sich vier Beispiele der 612 möglichen Minimalmuster, die entsprechenden Zellen sind durch Sterne gekennzeichnet.*



In Tabelle 4.1 ist abzulesen, wie viele Modellparameter, Minimalmusterelemente, Teilmengen mit $\lfloor k/2 \rfloor + 1$ Elementen, Minimalmuster und strenge Minimalmuster die Unabhängigkeitsmodelle ausgewählter zweidimensionaler Kontingenztafeln haben. Die Anzahl von Minimalmustern steigt demnach sehr schnell, weshalb die explizite Betrachtung aller Minimalmuster für größere Tafeln nicht mehr handhabbar ist. Stattdessen wird auf hinreichend große Stichproben von Minimalmustern zurückgegriffen.

Tabelle 4.1: Anzahl von Minimalmustern im Unabhängigkeitsmodell

| Dimension | $3 \times 3$ | $2 \times 5$ | $3 \times 4$ | $3 \times 5$ | $4 \times 4$ | $3 \times 6$ | $4 \times 5$ |
|---|---|---|---|---|---|---|---|
| $p = k_1 + k_2 - 1$ | 5 | 6 | 6 | 7 | 7 | 8 | 8 |
| $\lfloor \frac{k}{2} \rfloor + 1$ | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| $\binom{k}{\lfloor k/2 \rfloor + 1}$ | 126 | 210 | 792 | 6435 | 11440 | 43758 | 167960 |
| $\lvert\mathbb{M}\rvert$ | 81 | 80 | 612 | 3780 | 9552 | 26325 | 139660 |
| $\lvert\mathbb{S}\rvert$ | 81 | 80 | 432 | 2025 | 4096 | 41066 | 105408 |

Minimalmuster enthalten immer mehr als die Hälfte der Elemente einer Kontingenztafel. Falls $p = \lfloor \frac{k}{2} \rfloor + 1$ gilt, stimmen strenge Minimalmuster und Minimalmuster überein. Falls $p < \lfloor \frac{k}{2} \rfloor + 1$, genügen $\lfloor \frac{k}{2} \rfloor + 1 - p$ beliebig gewählte Zellen,

um ein strenges Minimalmuster zu einem Minimalmuster zu ergänzen. Nicht alle Teilmengen mit $p$ Zellen ergeben Designmatrizen mit vollem Rang.

Strenge Minimalmuster unterscheiden sich von sog. *strictly reconstructable replacement patterns* (Kuhnt, 2010). Letztere definieren Ausreißermuster, die eindeutig identifizierbar sind und zur Beschreibung des Zusammenbruch-Verhaltens von Schätzern und Identifizierern verwendet werden. Sie basieren auf unbedingten identifizierbaren Interaktionsmustern bei zweifaktoriellen Klassifikationsmodellen, siehe Terbeck und Davies (1998).

BEMERKUNG 4.6. *Definition 4.4 gilt auch für die Minimalmuster höherdimensionaler Kontingenztafeln. Bei gleicher Dimension, aber einem vom Unabhängigkeitsmodell abweichenden Modell hat ein Minimalmuster mehr Elemente, da $p$ größer wird. Ein zu einem saturierten Modell einer zweidimensionalen Tafel gehörendes Minimalmuster hat eine $p \times p$ Designmatrix. Diese kann (ii) aus Definition 4.4 nur erfüllen, wenn sie nicht reduziert wird, also alle Elemente der zugehörigen Tafel im Minimalmuster enthalten sind. Bei höherdimensionalen Tafeln können jedoch auch bedingte Unabhängigkeitsstrukturen vorgegeben werden, so dass das Modell nicht saturiert ist und eine Betrachtung von Minimalmustern sinnvoll ist.*

Im Folgenden werden Minimalmuster-basierte Ausreißeridentifizierer in Pseudo-Code vorgestellt. Zunächst wird etwas Notation benötigt. Für $v = 1, \ldots, V$ werden die Minimalmuster $\mathbb{M}_v$ durch Indikatormatrizen $\mathbf{W}_v$ dargestellt, wobei $V$ die Anzahl möglicher Minimalmuster bezeichnet. Die Matrix $\mathbf{W}_v$ hat die Dimension $\phi \times k$, wobei $\phi$ die Anzahl der Elemente eines Minimalmusters bezeichnet. Die $i$-te Zeile enthält Nullen außer an der Stelle, die im Originalbeobachtungsvektor $\mathbf{y}$ zum $i$-ten Element des Minimalmusters gehört, dort wird der Wert $1$ eingetragen. Mit Hilfe der Indikatormatrizen können Schätzer $\widehat{\boldsymbol{\beta}} = T(\{((\mathbf{W}_v\mathbf{X})_{i\bullet}, (\mathbf{W}_v\mathbf{y})_i), i = 1, \ldots, \phi\}, g)$ des unbekannten Parametervektors basierend auf der durch das Minimalmuster gegebenen Teilstichprobe definiert werden. Als Kurzschreibweise wird $\boldsymbol{\mathcal{Y}}_v := \mathbf{W}_v\boldsymbol{\mathcal{Y}}$ bzw. $\mathbf{X}_v := \mathbf{W}_v\mathbf{X}$ verwendet.

BEISPIEL 4.7. *Gegeben sei eine Kontingenztafel $\mathbf{N} \in \mathbb{R}^{3\times3}$ und das loglineare Poisson-Unabhängigkeitsmodell. Die nachfolgend in $\mathbf{N}$ durch Sterne gekennzeichneten Elemente*

| | | |
|---|---|---|
| $\star$ | $\star$ | |
| | $\star$ | |
| | $\star$ | $\star$ |

*d. h.* $n_{11}, n_{12}, n_{22}, n_{32}, n_{33}$ *bilden einen Kandidaten für ein Minimalmuster. Im Vektor* $\mathrm{vec}(\mathbf{N}^\top) = (n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{31}, n_{32}, n_{33})^\top$ *korrespondiert das mit den Elementen* $1, 2, 5, 8, 9$*. Die zugehörige Minimalmustermatrix ist durch*

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

*gegeben, welche in den Spalten* $1, 2, 5, 8, 9$ *jeweils den Wert* $1$ *besitzt, ansonsten Nullen.*

Der OMP-Algorithmus (*outlier detection with minimal patterns,* Kuhnt (2000)) wird für jedes $\mathbb{M}_v, v = 1, \ldots, V$ wiederholt: Es wird zunächst für jede Zelle $i$ in Abhängigkeit von $\mathbb{M}_v$ der geschätzte Parameter der Poissonverteilung $\hat{\theta}_i^v$ und die dazu korrespondierende $\alpha$-Ausreißerregion bestimmt. Die beobachteten Zellhäufigkeiten $y_1, \ldots, y_k$ werden jeweils mit ihrer korrespondierenden $\alpha$-Ausreißerregion verglichen. Die Menge der so identifizierten $\alpha$-Ausreißer wird Ausreißermenge genannt. Für jedes Minimalmuster existiert dann eine (eventuell leere) Ausreißermenge. Im letzten Schritt wird das Minimalmuster mit der kleinsten Ausreißermenge ausgewählt. Folgende Fälle sind zu unterscheiden:

1. Genau ein Minimalmuster identifiziert $n_{\mathrm{out}} \in \mathbb{N}_0$ Ausreißer, alle anderen Minimalmuster identifizieren $n_{\mathrm{out}} + 1$ oder mehr Ausreißer. Dann werden die von diesem Minimalmuster gefundenen Ausreißer als OMP-Ausreißer klassifiziert.

2. Es gibt mehrere Minimalmuster, die keinen Ausreißer erkennen. Dann wird keine Zellhäufigkeit als OMP-Ausreißer identifiziert.

3. Es gibt mehr als ein Minimalmuster, das die minimale Anzahl von Ausreißern

findet, mindestens jedoch einen. Sind die von den Minimalmustern gefundenen Ausreißer an denselben Stellen der Tafel, werden sie als OMP-Ausreißer klassifiziert. Sind sie an unterschiedlichen Stellen der Tafel, existieren konkurrierende potenzielle OMP-Ausreißermuster.

Der OMP-Algorithmus basiert auf der Idee, dass in der Regel nicht alle Minimalmuster mindestens einen Ausreißer identifizieren, falls keine Ausreißer vorliegen. Falls andererseits ein Ausreißer in der Tafel ist, sollte er von allen Minimalmustern, die ihn nicht enthalten, als solcher erkannt werden. In den Minimalmustern, in denen der Ausreißer enthalten ist, wird der zugehörige Schätzer womöglich so verzerrt, dass mehr als ein Ausreißer erkannt wird. Falls die minimale Anzahl von Ausreißern bei unterschiedlichen Ausreißermengen erreicht wird, existiert keine eindeutige Lösung (Kuhnt, 2000). Es ist jedoch basierend auf Hintergrundwissen möglich, eine der Ausreißermengen als „plausibelste Ausreißermenge“ zu definieren. Eine andere, konservativere Möglichkeit bei mehreren gleichgroßen Ausreißermengen besteht darin, den Schnitt über die minimalen Ausreißermengen als OMP-Ausreißermenge zu definieren. Diese Ausreißermenge könnte auch als Ausgangspunkt für einen mehrschrittigen Identifizierer dienen. In dieser Arbeit werden jedoch nur einschrittige Identifizierer betrachtet. Eine weitere Alternative zur „plausibelsten Ausreißermenge“ besteht darin, diejenige Ausreißermenge auszuwählen, welche von den Ausreißermengen minimaler Größe am häufigsten angegeben wurde.

**OMP-Algorithmus:**

**for** $v = 1$ **to** $V$ **do**
  $\widehat{\boldsymbol{\beta}}_v^{\mathrm{ML}} \leftarrow \mathrm{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left( \sum_{i=1}^{\phi} \left( \mathbf{e}_i^\top \mathbf{y}_v \, (\mathbf{X}_v)_{i\bullet}\boldsymbol{\beta} - \exp((\mathbf{X}_v)_{i\bullet}\boldsymbol{\beta}) \right) \right)$;
  **for** $i = 1$ **to** $k$ **do**
    Bestimme $\mathrm{out}(\alpha, Poi(\hat{\theta}_i^v))$ für $\hat{\theta}_i^v = \exp\left( \mathbf{X}_{i\bullet} \widehat{\boldsymbol{\beta}}_v^{\mathrm{ML}} \right)$
  **end**
  $NUMB.OUT(v) \leftarrow$ Anzahl der Ausreißer bei Minimalmustermatrix $\mathbf{W}_v$
**end**
$v^* \leftarrow \mathrm{argmin}_v \, NUMB.OUT(v)$;
Ausreißermuster $\leftarrow$ Zellen mit Ausreißern basierend auf $\mathrm{out}(\alpha, Poi(\widehat{\boldsymbol{\theta}}_{v^*}))$

Ein im Rahmen dieser Dissertation entwickelter Algorithmus, der ebenfalls auf dem

Minimalmusterprinzip basiert, aber eindeutige Ausreißermengen erzwingt, ist der OMPC-Algorithmus. Er zählt die Anzahl $m_i$, die definiert, wie oft Zelle $i$ *über alle Minimalmuster hinweg* als Ausreißer erkannt wird. Wie viele Ausreißer *pro Minimalmuster* identifiziert werden, spielt somit im Gegensatz zum OMP-Algorithmus keine Rolle. Zelle $i$ wird als OMPC-Ausreißer klassifiziert, falls $m_i$ einen bestimmten Grenzwert überschreitet. Der konkrete Grenzwert ist abhängig von der Größe der Tafel $k$, der Anzahl der Elemente pro Minimalmuster $\phi$ und der Anzahl an Minimalmustern $V$. Jede Zelle ist in $\frac{\phi}{k}V$ Minimalmustern enthalten und in $(1-\frac{\phi}{k})V$ Minimalmustern nicht enthalten. Zur Identifizierung möglicher Ausreißer werden für jede Zelle nur die Minimalmuster betrachtet, die die Zelle nicht enthalten. Eine Zelle wird als OMPC-Ausreißer bezeichnet, wenn sie in mehr als $\zeta(1-\frac{\phi}{k})V$ der $\alpha$-Ausreißerregionen enthalten ist, wobei $\zeta \in [0.5, 1]$. Die Wahl von $\zeta$ als Schwellenwert wird in Abschnitt 4.1.2 thematisiert.

**OMPC-Algorithmus:**

**for** $v = 1$ **to** $V$ **do**
  $\widehat{\boldsymbol{\beta}}_v^{\mathrm{ML}} \leftarrow \mathrm{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left( \sum_{i=1}^{\phi} \left( \mathbf{e}_i^\top \mathbf{y}_v \left(\mathbf{X}_v\right)_{i\bullet}\boldsymbol{\beta} - \exp((\mathbf{X}_v)_{i\bullet}\boldsymbol{\beta})\right)\right)$;
  $\widehat{\boldsymbol{\theta}}^v \leftarrow \exp\left(\mathbf{X}\widehat{\boldsymbol{\beta}}_v^{\mathrm{ML}}\right)$
**end**
**for** $i = 1$ **to** $k$ **do**
  $\widetilde{\mathbb{M}} \leftarrow$ Indexmenge der Minimalmuster, die Zelle $i$ nicht enthalten;
  **if** *$y_i$ ist in mindestens $\zeta|\widetilde{\mathbb{M}}|$ der auf $\hat{\theta}_i^v, v \in \widetilde{\mathbb{M}}$ basierenden $\alpha$-Ausreißerregionen enthalten* **then**
    Klassifiziere $y_i$ als Ausreißer
  **end**
**end**

Die Minimalmuster werden als potenziell ausreißerfreie Teilmengen der Kontingenztafeln angesehen, weswegen nichts gegen die Verwendung des ML-Schätzers spricht. In dieser Arbeit wird dennoch auch der $L_1$-Schätzer zum Vergleich herangezogen. Der auf dem $L_1$-Schätzer basierende Algorithmus wird OMPCL1-Algorithmus genannt und unterscheidet sich vom OMPC-Algorithmus bezüglich des Pseudo-Codes nur im Optimierungsproblem der ersten For-Schleife. Die im Rahmen dieser Dissertation erhaltenen ersten Ergebnisse in Form einer Simulati-

onsstudie wurden vorab in Kuhnt, Rapallo und Rehage (2014) publiziert und können dem Anhang (Tab. A.1) entnommen werden. Dabei wurde im OMPC- und OMPCL1-Algorithmus $\zeta = 0.5$ gewählt. Die in Abschnitt 4.1.2 dargestellten Simulationsergebnisse basieren auf einer situativen Wahl von $\zeta$, wodurch die Vergleichbarkeit der untersuchten Ausreißeridentifizierer innerhalb jedes Szenarios erhöht wird.

Eine Alternative zum OMPC-Algorithmus basiert auf dem in Definition 3.9 vorgestellten LTCS-Schätzer, der ebenfalls zur Bestimmung von Ausreißermustern in Kontingenztafeln verwendet werden kann.

BEMERKUNG 4.8. *Aufgrund der vielen lokalen Minima der zu minimierenden Funktion* $\widehat{\boldsymbol{\beta}}_{\mathfrak{h}}^{LTCS} = \operatorname{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{k} \mathbf{1}_{[0,\chi^2_{(\mathfrak{h})}]}(\chi^2_{(i)})\chi^2_{(i)}$ (3.9) *verwenden Shane und Simonoff (2001) keine Gradientenverfahren, sondern werten das Minimierungsproblem nur an endlich vielen Stellen* $\boldsymbol{\beta}$ *aus. Hier bieten strenge Minimalmuster die Möglichkeit eines generellen Vorgehens, indem aus jedem strengen Minimalmuster ein potenzieller Kandidat für* $\widehat{\boldsymbol{\beta}}^{LTCS}$ *resultiert, unter denen der beste ausgewählt wird.*

Mit Hilfe des so bestimmten LTCS-Schätzers kann gemäß OLTCS-Algorithmus ein Ausreißeridentifizierer definiert werden. Für große Tafeln ist die Minimalmustermethode nicht mehr handhabbar, was auch die Anwendbarkeit von OLTCS einschränkt. Simulationsergebnisse von Shane und Simonoff (2001) für $5 \times 5$ und $8 \times 8$ Kontingenztafeln zeigen, dass sich die Schätzer nur marginal ändern, wenn eine zufällige Auswahl von $500, 1000$ oder $1500$ verschiedenen statt $V^{str}$ möglichen (von ihnen *elemental subsets* genannten) strengen Minimalmustern erfolgt. Dabei ziehen die Autoren per Brute-Force-Suche zufällig $p$-elementige Teilmengen der Zellen, bis sie die gewünschte Anzahl von Mustern mit zugehöriger nicht-singulärer Designmatrix erhalten. Als Kurzschreibweisen für die zu strengen Minimalmustern korrespondierenden Partitionen von $\mathbf{y}$ und $\mathbf{X}$ werden $\check{\mathbf{y}}_v$ und $\check{\mathbf{X}}_v$ verwendet.

Alternativ kann im OLTCS-Algorithmus der Pearson *least median of chi-squared residuals* (LMCS) Schätzer verwendet werden. Wie Shane und Simonoff (2001) per Simulation zeigen, unterscheiden sich die LTCS- und LMCS-Parameterschätzwerte allerdings kaum. In Abschnitt 4.1.2 werden die vorgestellten Ausreißeridentifizierer OL1, OMPC, OMPCL1 und OLTCS in einer Simulationsstudie verglichen.

**OLTCS-Algorithmus:**

Tuningparameter $\mathfrak{h} \leftarrow \lfloor (k+p+2)/2 \rfloor$ (Bruchpunkt-optimierender Wert, (3.10)) ;
**for** $v = 1$ **to** $V^{str}$ **do**
  $\widehat{\boldsymbol{\beta}}_v^{\mathrm{ML}} \leftarrow \mathrm{argmax}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left( \sum_{i=1}^{p} \left( \mathbf{e}_i^\top \check{\mathbf{y}}_v \, (\check{\mathbf{X}}_v)_{i\bullet}\boldsymbol{\beta} - \exp((\check{\mathbf{X}}_v)_{i\bullet}\boldsymbol{\beta}) \right) \right)$;
  $\widehat{\mathbf{y}}_v \leftarrow \exp\left( \mathbf{X}\widehat{\boldsymbol{\beta}}_v^{\mathrm{ML}} \right)$;
  **for** $i = 1$ **to** $k$ **do**
    $\chi^2_{v,i} \leftarrow (y_{v,i} - \hat{y}_{v,i})^2 / \hat{y}_{v,i}$
  **end**
**end**
$v^* \leftarrow \mathrm{argmin}_v \sum_{i=1}^{k} \mathbf{1}_{[0,\chi^2_{v,(\mathfrak{h})}]}(\chi^2_{v,(i)}) \chi^2_{v,(i)}$;
**for** $i = 1$ **to** $k$ **do**
  Bestimme $\mathrm{out}(\alpha, Poi(\hat{\theta}_i^{v^*}))$ für $\hat{\theta}_i^{v^*} = \exp(\mathbf{X}_{i\bullet}\widehat{\boldsymbol{\beta}}_{v^*}^{\mathrm{ML}})$;
  **if** $y_i \in \mathrm{out}(\alpha, Poi(\hat{\theta}_i^{v^*}))$ **then**
    Klassifiziere $y_i$ als Ausreißer
  **end**
**end**

**Eigenschaften des Minimalmusterverfahrens.** In diesem Abschnitt werden Eigenschaften von Minimalmustern untersucht. Der Fokus liegt auf dem loglinearen Poisson-Unabhängigkeitsmodell zweidimensionaler $k_1 \times k_2$ Kontingenztafeln, wobei o. B. d. A. $k_1 \leq k_2$ gelte. Es wird unter anderem ein Konzept vorgestellt, um die Betrachtung von Mustern, die Definition 4.4 nicht erfüllen, direkt zu vermeiden. Dieses Konzept basiert auf sog. Schlingen (*cycles*) und wurde gemeinsam mit Dr. Fabio Rapallo (Universität Ostpiemont) entwickelt.

Tabelle 4.1 zeigt, dass die Anzahl als Minimalmuster infrage kommender Teilmengen schnell sehr groß wird. Um nicht alle Teilmengen untersuchen zu müssen, wird im Folgenden die Struktur strenger Minimalmuster intensiver untersucht.

BEISPIEL 4.9. *Gegeben sei eine $3 \times 3$ Kontingenztafel. Dann haben die hier dargestellten Teilmengen unterschiedliche Auswirkungen auf den Rang der jeweiligen Designmatrizen:*

*(a)*

| | | |
|---|---|---|
| ⋆ | ⋆ | |
| ⋆ | ⋆ | |
| | | ⋆ |

*(b)*

| | | |
|---|---|---|
| ⋆ | ⋆ | |
| | ⋆ | ⋆ |
| | | ⋆ |

*Die Sterne markieren die jeweiligen Elemente der Teilmengen. Muster (a) ist kein Mini-*

*malmuster, da die korrespondierende Designmatrix nicht vollen Rang hat. Muster (b) ist hingegen ein strenges Minimalmuster. Es ist zu erkennen, dass Muster (a) im Gegensatz zu Muster (b) eine komplette* $2 \times 2$ *„Teiltafel" enthält. In größeren Kontingenztafeln reicht dies jedoch nicht mehr, um strenge Minimalmuster hinreichend zu charakterisieren:*

*(c)*

| | | | |
|---|---|---|---|
| ⋆ | ⋆ | | |
| ⋆ | | ⋆ | |
| | ⋆ | ⋆ | |
| | | | ⋆ |

*(d)*

| | | | |
|---|---|---|---|
| ⋆ | ⋆ | | |
| ⋆ | ⋆ | | |
| | | ⋆ | ⋆ |

*Muster (c) enthält keine* $2 \times 2$ *Teiltafel; die zugehörige Designmatrix hat allerdings keinen vollen Rang. Daraus wird klar, dass die Abwesenheit einer* $2 \times 2$ *Teiltafel kein hinreichendes Kriterium für ein strenges Minimalmuster ist. Eine Gemeinsamkeit mit Muster (a) besteht darin, dass in der letzten Zeile und Spalte nur ein Element, die sogenannte isolierte Zelle, zur Teilmenge gehört. Muster (d), dessen zugehörige Designmatrix ebenfalls keinen vollen Rang hat, enthält eine komplette* $2 \times 2$ *Teiltafel, aber keine isolierte Zelle wie die Muster (a) und (c). Somit folgt, dass die Abwesenheit von isolierten Zellen ebenfalls kein hinreichendes Kriterium für ein strenges Minimalmuster ist.*

Um ein Kriterium zur Identifikation strenger Minimalmuster zu finden, das die Fälle (a), (c) und (d) von Beispiel 4.9 umfasst, wird der Begriff der $\mathfrak{k}$-Schlinge eingeführt. Diese Definition und entsprechende Ergebnisse wurden bereits in Kuhnt, Rapallo und Rehage (2014) veröffentlicht.

DEFINITION 4.10. *Sei* $\mathfrak{k} \geq 2$*. Eine* $\mathfrak{k}$*-Schlinge ist gegeben durch eine Menge von* $2\mathfrak{k}$ *Zellen, die in einer* $\mathfrak{k} \times \mathfrak{k}$ *Teiltafel enthalten sind und für die gilt, dass sich genau je zwei Zellen in jeder Zeile und Spalte der Teiltafel befinden.*

BEISPIEL 4.11. *Laut Definition 4.10 ist eine* 2*-Schlinge eine* $2 \times 2$ *Teiltafel. Eine* 3*-Schlinge ist definiert durch sechs Zellen der Form*

| | | |
|---|---|---|
| ⋆ | ⋆ | |
| ⋆ | | ⋆ |
| | ⋆ | ⋆ |

*genauso wie durch alle anderen Permutationen von Zeilen oder Spalten dieses Musters.*

Im Fall des Unabhängigkeitsmodells verknüpft das folgende Theorem Schlingen und strenge Minimalmuster.

THEOREM 4.12. *Eine Menge von $p = k_1 + k_2 - 1$ Zellen formt ein strenges Minimalmuster im Unabhängigkeitsmodell genau dann, wenn sie keine $\mathfrak{k}$-Schlingen beinhaltet, $\mathfrak{k} = 2, \ldots, \min\{k_1, k_2\}$.*

*Beweis:* Die beiden Implikationen werden jeweils über ihre Kontraposition hergeleitet.

Eine Schlinge kann in zwei $\mathfrak{k}$-elementige Teilmengen $A_1$ und $A_2$ zerlegt werden, die jeweils eine Zelle pro Zeile und Spalte enthalten. Werden nun die zu $A_1$ korrespondierenden Zeilen der Designmatrix $\mathbf{X}$ addiert, erhält man dieselbe Summe wie für die Zeilen, die zu $A_2$ gehören. Damit hat die durch die Schlinge und somit auch die durch die $p$-elementige Menge bestimmte Teilmatrix von $\mathbf{X}$ keinen vollen Rang und das Muster ist nach Definition 4.4 kein strenges Minimalmuster.

Andererseits, falls die Matrixpartition $\underline{\mathbf{X}} \in \mathbb{R}^{p \times p}$ singulär ist und somit kein strenges Minimalmuster, existiert eine Linearkombination der Zeilen von $\underline{\mathbf{X}}$:

$$\gamma_1 c_{(i_1,j_1)} + \ldots + \gamma_p c_{(i_p,j_p)} = 0, \tag{4.1}$$

wobei $c_{(i,j)}$ die Zeile von $\underline{\mathbf{X}}$ ist, die zu Zelle $(i,j)$ gehört und die Koeffizienten $\gamma_1, \ldots, \gamma_p$ sind nicht alle gleich null. Sei o. B. d. A. $\gamma_1 > 0$. Da der Indikatorvektor von Zeile $i_1$ im Spaltenraum von $\underline{\mathbf{X}}$ enthalten ist und das auch für den Indikatorvektor der Spalte $j_1$ gilt, muss eine Zelle $(i_2, j_2)$ in derselben Zeile sein, also $(i_1, j_2)$, wobei der Koeffizient in (4.1) negativ sein muss. Für eine Zelle in derselben Spalte $(i_3, j_1)$ gilt dasselbe. Daher muss die Zelle $(i_3, j_2)$ einen positiven Koeffizienten haben. Dann gilt folgende Fallunterscheidung:

1. Zelle $(i_3, j_2)$ ist im Muster enthalten. Dann liegt eine 2-Schlinge vor.
2. Zelle $(i_3, j_2)$ ist nicht im Muster enthalten. Dann muss das obige Vorgehen sukzessive wiederholt werden, um eine 3-Schlinge (4-Schlinge, usw.) aufzufinden. Dabei werden die Zellen $(i_1, j_2)$ und $(i_3, j_1)$ als Ausgangspunkte gewählt.

Dies zeigt, dass eine bestimmte Anzahl $\ell$ an Zeilen und Spalten existiert, die je zwei Zellen mit einem Koeffizienten ungleich null haben. Dies ist per Definition eine $\ell$-Schlinge. □

Damit produziert der nachfolgende Algorithmus strenge Minimalmuster:

1. Sei $\mathcal{S}_{all}$ die Menge aller Zellen einer Kontingenztafel und $\mathcal{S}_{sel}$ die Menge der gewählten Zellen, mit initial $\mathcal{S}_{sel} = \emptyset$.

2. Für $q \in \{1, \ldots, k_1 + k_2 - 1\}$:

   (a) Ziehe eine zufällige Zelle aus $\mathcal{S}_{all}$, füge sie zu $\mathcal{S}_{sel}$ hinzu und lösche sie aus $\mathcal{S}_{all}$.

   (b) Finde alle Tupel mit drei, fünf, usw. Elementen in $\mathcal{S}_{sel}$, die die gewählte Zelle beinhalten, und lösche aus $\mathcal{S}_{all}$ alle Zellen (falls vorhanden), die Schlingen produzieren.

Eine 2-Schlinge besteht aus $4$ Zellen. Daher können die ersten drei Zellen immer rein zufällig ausgewählt werden. Außerdem ist der Algorithmus symmetrisch in Bezug auf Zeilen- und Spaltenpermutationen. Daraus folgt, dass die Auftretenswahrscheinlichkeit aller strengen Minimalmuster gleich ist.

Für $3 \times 3$ Kontingenztafeln ist Theorem 4.12 äquivalent zum folgenden Kriterium von Kuhnt (2000, S. 87). Aufgrund des hier eingeführten Konzepts der Schlinge kann das Kriterium erstmals analytisch bewiesen werden.

KOROLLAR 4.13. *Im Unabhängigkeitsmodell von* $3 \times 3$ *Kontingenztafeln ist die Existenz von 2-Schlingen ausgeschlossen, falls*

*(i) keine leeren Zeilen;*

*(ii) keine leeren Spalten;*

*(iii) und für jede gewählte Zelle mindestens eine weitere Zelle in der gleichen Zeile oder Spalte gewählt wurde.*

*Beweis:* Angenommen, es gebe eine leere Zeile. In den übrigen zwei Zeilen müssen fünf Zellen des Minimalmusters gewählt werden. Daraus folgt, dass eine 2-Schlinge existiert. Aus dem gleichen Grund gilt dies auch für leere Spalten. Falls eine beliebige Zelle $n_{ij}$ gewählt wurde und in Zeile $i$ und Spalte $j$ keine weiteren Zellen ausgewählt werden, verbleiben nur vier Zellen in je zwei Zeilen und Spalten, die daher eine 2-Schlinge bilden.

Andererseits: Es existiere eine 2-Schlinge. O. B. d. A. sei diese durch die Zellen $n_{11}, n_{12}, n_{21}$ und $n_{22}$ definiert. Die letzte zu wählende Zelle des Minimalmusters kann entweder in den Zellen $n_{13}$ oder $n_{23}$ liegen, was eine leere Zeile verursacht; in den Zellen $n_{31}$ oder $n_{32}$, was eine leere Spalte verursacht, oder in der Zelle $n_{33}$, womit sie die einzige Zelle in Zeile $3$ und Spalte $3$ wäre. □

Für allgemeine $k \times k$ Kontingenztafeln kann jedoch nicht gefolgert werden, dass aus den Bedingungen in Korollar 4.13 auch die Nicht-Existenz von $2$-Schlingen folgt. Dazu sei im Fall $k = 4$ das Muster

| | | | |
|---|---|---|---|
| ⋆ | ⋆ | | |
| ⋆ | ⋆ | | |
| | | | ⋆ |
| | | ⋆ | ⋆ |

herangezogen, das die Bedingungen (i), (ii) und (iii) von Korollar 4.13 erfüllt, aber eine $2$-Schlinge enthält.

Für allgemeine loglineare Modelle kann ein effizienter Algorithmus zur Ziehung von Minimalmustern wie folgt definiert werden:

(a) Ziehe ein strenges Minimalmuster.

(b) Falls nötig, füge so viele Zellen zufällig hinzu, bis die für ein Minimalmuster benötigte Anzahl Zellen erreicht ist.

Diese Prozedur kann wiederholt werden, bis jedes mögliche Minimalmuster gefunden wurde. Alternativ kann die Prozedur nach einer bestimmten Zeit oder Anzahl

gefundener Minimalmuster beendet werden. In diesem Fall produziert das Vorgehen zufällig ausgewählte Minimalmuster, falls in Schritt (a) das strenge Minimalmuster zufällig gezogen wird.

In $3 \times 3$ und $2 \times 5$ Kontingenztafeln ist jedes strenge Minimalmuster gleichzeitig ein Minimalmuster, siehe Tabelle 4.1. Da in diesen Fällen $p = k_1 + k_2 - 1 = \lfloor k/2 \rfloor + 1$ gilt, müssen $p$ Unbekannte durch $p$ Gleichungen bestimmt werden. Dies ist exakt möglich, weshalb ML-Schätzer basierend auf Minimalmustern in $3 \times 3$ und $2 \times 5$ Kontingenztafeln die Exact-Fit-Eigenschaft (Def. 3.7) besitzen.

Auch bei gewissen Unabhängigkeitsstrukturen des loglinearen Poissonmodells kann es zum Phänomen der exakten Anpassung kommen. Das saturierte Modell, in dem alle möglichen Mehrfachwechselwirkungen zwischen den einzelnen Parametern enthalten sind, liefert hierfür ein extremes Beispiel, da nicht nur mehr als die Hälfte der Beobachtungen, sondern alle Beobachtungen exakt geschätzt werden.

Rousseeuw und Leroy (1987, S. 123) merken an, dass der auf der Exact-Fit-Eigenschaft basierende Exact-Fit-Punkt eine obere Schranke für den Finite-Sample-Bruchpunkt ist. Daher ist es naheliegend, die Exact-Fit-Eigenschaft als Robustheitsmaß aufzufassen. Es kann aber auch der Fall auftreten, in dem ein Schätzer mit Exact-Fit-Eigenschaft Schätzwerte liefert, die alles andere als robust sind: Für eine $3 \times 3$ Kontingenztafel existieren 81 Minimalmuster. Jede Zelle ist in 45 Minimalmustern, also in mehr als der Hälfte, enthalten. Wird ein beliebig extremer Ausreißer an einer Stelle der Kontingenztafel platziert, werden 45 der 81 ML-Schätzer für $\boldsymbol{\beta}$ ihn exakt schätzen. Diese exakte Anpassung führt dazu, dass Beobachtungen außerhalb des Minimalmusters beliebig weit von ihrer ML-Schätzung entfernt sind: Sei $y_i \to \infty$ die kontaminierte Beobachtung und $i \in \widetilde{\mathbb{M}}_v$, wobei $\widetilde{\mathbb{M}}_v$ die Indexmenge des $v$-ten Minimalmusters bezeichne. Dann gibt es mindestens ein $y_j$ in derselben Zeile oder Spalte wie $y_i$, für das $|y_j - \hat{y}_j^v| \to \infty$ gilt. Wegen $||\mathbf{y} - \widehat{\mathbf{y}}^v|| \to \infty$ ist der Finite-Sample-Bruchpunkt von $\widehat{\mathbf{y}}^v$ gleich $1/9$.

Dieses Beispiel zeigt, dass ein einzelner ML-Schätzwert basierend auf einem Minimalmuster keine guten Robustheitseigenschaften haben muss. Im Gegensatz dazu kann der vektorwertige Ausreißeridentifizierer $\mathbf{OI}(\mathbf{y}; \alpha, \mathrm{OMPC})$ gute Robustheitseigenschaften haben, was anhand von Masking- und Swamping-Bruchpunkt ana-

lysiert wird. Betrachtet man nur ein Minimalmuster im $3 \times 3$-Fall, liegen Masking- und Swamping-Bruchpunkt bei $1/9$, wenn der Ausreißer im Minimalmuster enthalten ist. Der optimale Swamping-Bruchpunkt von $1$ kann erreicht werden, wenn die Nicht-Ausreißer alle innerhalb des Minimalmusters liegen und die Ausreißer alle außerhalb. Werden sukzessive Nicht-Ausreißer durch Ausreißer ersetzt, ändert sich die Klassifikation der verbleibenden Nicht-Ausreißer aufgrund der Exact-Fit-Eigenschaft nicht. Ist der beliebig extreme Wert nicht im Minimalmuster enthalten, wird er korrekt als Ausreißer klassifiziert. Dies ist solange der Fall, bis der erste Nicht-Ausreißer im Minimalmuster durch einen Ausreißer ersetzt wird, also liegt der Masking-Bruchpunkt bei $5/9$.

Anhand dieser Analyse wird deutlich, dass die allgemeine Herleitung von Masking- und Swamping-Bruchpunkten bei Minimalmustern schwierig oder unmöglich ist. Das folgende Lemma dient der effizienteren Herleitung der OMPC-Ausreißer.

LEMMA 4.14. *Seien $\mathbf{y} = (y_1, \ldots, y_k)^\top$ die Einträge einer Kontingenztafel und das loglineare Poisson-Unabhängigkeitsmodell $\mathsf{E}(\mathcal{Y}_i|\mathbf{X}) = \exp(\mathbf{X}_{i\bullet}\boldsymbol{\beta}), i = 1, \ldots, k$ mit $\boldsymbol{\beta} \in \mathbb{R}^p$ und der zugehörigen, vollständigen Designmatrix $\mathbf{X} \in \mathbb{R}^{k \times p}$ gegeben. Seien $\mathbf{y}_v \in \mathbb{R}^{\lfloor \frac{k}{2} \rfloor + 1}$ die Einträge eines Minimalmusters und $\mathbf{X}_v \in \mathbb{R}^{\lfloor \frac{k}{2} \rfloor + 1 \times p}$ die reduzierte Designmatrix. Außerdem sei $\widehat{\mathbf{y}}^{ML(v)} \in \mathbb{R}^k$ die ML-Anpassung, die auf dem $v$-ten Minimalmuster basiert. Dann gilt: Falls $\lfloor \frac{k}{2} \rfloor + 1 = p$, folgt $\widehat{\mathbf{y}}^{ML(v)} = \exp(\mathbf{X}\mathbf{X}_v^{-1} \ln(\mathbf{y}_v))$.*

*Beweis:* Nach (3.7) gilt mit $\mathbf{y}_v$ und $\mathbf{X}_v$ für $\mathbf{y}$ und $\mathbf{X}$:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\mathrm{ML}} &= \operatorname*{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{j=1}^{p} \left( (\mathbf{y}_v)_j \, (\mathbf{X}_v)_{j\bullet}\boldsymbol{\beta} - \exp((\mathbf{X}_v)_{j\bullet}\boldsymbol{\beta}) \right) \\ &= \operatorname*{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p} \left( \mathbf{y}_v^\top \mathbf{X}_v \boldsymbol{\beta} - \mathbb{1}_p^\top \exp(\mathbf{X}_v\boldsymbol{\beta}) \right) =: \operatorname*{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^p} f(\boldsymbol{\beta}).\end{aligned}$$

Dann folgt mit Hilfe des Matrixkalküls

$$\begin{aligned}\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}_v^\top \mathbf{y}_v - \mathbf{X}_v^\top \exp(\mathbf{X}_v\boldsymbol{\beta}) &\stackrel{!}{=} 0 \\ \Leftrightarrow \mathbf{X}_v^\top \mathbf{y}_v &= \mathbf{X}_v^\top \exp(\mathbf{X}_v\boldsymbol{\beta}).\end{aligned}$$

Da $\mathbf{X}_v \in \mathbb{R}^{p \times p}$ vollen Rang hat, existiert ihre Inverse $\mathbf{X}_v^{-1}$. Somit ergibt sich:

$$\mathbf{y}_v = \exp(\mathbf{X}_v\boldsymbol{\beta}) \Leftrightarrow \ln(\mathbf{y}_v) = \mathbf{X}_v\boldsymbol{\beta} \Leftrightarrow \boldsymbol{\beta} = \mathbf{X}_v^{-1}\ln(\mathbf{y}_v).$$

Folglich ist der ML-Schätzer in geschlossener Form durch $\widehat{\boldsymbol{\beta}}_v^{\mathrm{ML}} = \mathbf{X}_v^{-1}\ln(\mathbf{y}_v)$ darstellbar, was für die geschätzten Zellhäufigkeiten der vollständigen Tafel $\widehat{\mathbf{y}}^{\mathrm{ML}(v)} = \exp(\mathbf{X}\mathbf{X}_v^{-1}\ln(\mathbf{y}_v))$ ergibt. □

Im folgenden Abschnitt wird die Performanz der vorgestellten Ausreißeridentifizierer in verschiedenen Ausreißerszenarien untersucht.

### *4.1.2 Simulationsstudie*

Um das Verhalten der hier vorgestellten Ausreißeridentifizierer in Kontingenztafeln mit einer moderaten bis hohen Anzahl an Beobachtungen zu vergleichen, wird die folgende Simulationsstudie durchgeführt. Teilergebnisse dieser Studie wurden bereits in Kuhnt, Rapallo und Rehage (2014) veröffentlicht, vgl. Tabelle A.1. Zwei Maße stehen im Fokus des Vergleichs: Der Anteil korrekt identifizierter $\alpha$-Ausreißer (Richtig-Positiv-Rate, RPR) und der Anteil korrekt identifizierter $\alpha$-Nicht-Ausreißer (Richtig-Negativ-Rate, RNR). Dabei ist ein wichtiger Aspekt, welche Auswirkungen mehrere Ausreißer in gleichen Zeilen (oder Spalten) der Kontingenztafel haben, da sie sich im Unabhängigkeitsmodell auf die Schätzung desselben Regressionsparameters auswirken. Andererseits soll auch die Performanz auf den unkontaminierten Daten untersucht werden. Dafür werden drei loglineare Poisson-Unabhängigkeitsmodelle aufgestellt, für die unterschiedliche Ausreißersituationen angenommen werden. Überprüft werden die einschrittigen Methoden mit $L_1$-Schätzer, LTCS-Schätzer, Minimalmusterverfahren (OMPC) und Minimalmusterverfahren mit $L_1$-Schätzer (OMPCL1). Das Minimalmusterverfahren aus dem OMP-Algorithmus wird dabei nicht untersucht, da die Ergebnisse nicht eindeutig sein können und daher schwierig zu vergleichen sind.

Es werden Daten anhand von Poissonmodellen zu verschieden großen Kontingenztafeln ($3 \times 3$, $4 \times 4$ und $10 \times 10$) generiert. Simulierte Daten, die laut des vorgege-

benen Modells $\alpha$-Ausreißer sind, werden durch neu simulierte Daten ersetzt. Ein oder zwei Ausreißer werden künstlich in vorher festgelegte Zellen eingefügt. Die Ausreißer werden analog zu von Eye (2002) als „Typen“ und „Antitypen“ bezeichnet. Ein Ausreißer wird als Typ bezeichnet, wenn er größer als die obere Grenze der $(1-\alpha)$-Inlierregion ist und als Antityp, wenn er kleiner als die untere Grenze der $(1-\alpha)$-Inlierregion ist.

Die sieben simulierten Szenarien mit $N = 500$ Wiederholungen werden unten beschrieben. Es werden $\delta$-Ausreißer in unkontaminierte Kontingenztafeln eingefügt, wobei wegen $\delta \in \{10^{-8}, 10^{-4}\}$ die $10^{-4}$-Ausreißer „moderat“ genannt werden und die $10^{-8}$-Ausreißer „extrem“. Inwieweit $10^{-4}$-Ausreißer tatsächlich (noch) als moderat angesehen werden können, hängt von der jeweiligen Anwendung ab. Allgemein gilt in den Szenarien für den Eintrag $\mathcal{Y}_i, i = 1, \ldots, k$ einer Kontingenztafel: $\mathcal{Y}_i \sim Poi(\mathbf{X}_{i\bullet}\boldsymbol{\beta})$. Die Simulationen wurden mit R (R Core Team, 2016) durchgeführt, wobei die Zufallszahlen mit Hilfe der Funktion `rpois` aus der Poissonverteilung generiert werden.

1. $3 \times 3$ Kontingenztafeln werden mit der Effektkodierungsdesignmatrix $\widetilde{\mathbf{X}} = \widetilde{\mathbf{X}}_{3\times 3}$ (siehe Anhang A.5, Formel (A.2)) und $\boldsymbol{\beta}_1 = (4, 0.2, -0.2, 0.4, 0.3)^\top$ erzeugt. Je ein moderater $\delta$-Ausreißer ($\delta = 10^{-4}$) wird in $n_{11}$ eingesetzt. Die spezifische Position des Ausreißers ist dabei nicht relevant. Die $\delta$-Ausreißerregion ist für $n_{11}$ durch

$$\mathrm{out}(\delta, Poi(\exp(\widetilde{\mathbf{X}}_{1\bullet}\boldsymbol{\beta}_1))) = \{0, 1, \ldots, 62\} \cup \{141, 142, \ldots\}$$

   gegeben, so dass als Antityp $n_{11} = 62$ und als Typ $n_{11} = 141$ gewählt wird. Im Gegensatz zu Kuhnt (2000) werden im Folgenden auch größere Tafeln untersucht.

2. $4 \times 4$ Kontingenztafeln werden mit $\widetilde{\mathbf{X}}_{4\times 4}$ (siehe Anhang A.5, Formel (A.3)) und $\boldsymbol{\beta}_2 = (3.8, 0.2, -0.2, 0.1, 0.25, 0.3, -0.1)^\top$ generiert. Es wird ein moderater Ausreißer ($\delta = 10^{-4}$) in Zelle $n_{11}$ eingefügt: $n_{11} = 39$ als Antityp und $n_{11} = 105$ als Typ.

3. $4 \times 4$ Kontingenztafeln werden mit $\boldsymbol{\beta}_2$ und $\widetilde{\mathbf{X}}_{4\times 4}$ generiert. Dabei wird ein wei-

terer Ausreißer ($\delta = 10^{-4}$) hinzugefügt, woraus drei verschiedene Situationen entstehen: Zwei Typen, zwei Antitypen und je ein Typ und Antityp. Die Ausreißer werden in $n_{11}$ und $n_{12}$ eingefügt. Damit sind in einer Zeile die Hälfte der Einträge Ausreißer. Dieser Fall stellt Ausreißeridentifizierer vor große Probleme, da hier bei zwei ersetzten Typen bzw. Antitypen zwei konkurrierende Modelle existieren können, für die entweder die Nicht-Ausreißer *oder* die Ausreißer unkontaminierte Beobachtungen sind.

4. Es gelte das Modell aus Szenario 3 mit $\boldsymbol{\beta}_2$ und $\widetilde{\mathbf{X}}_{4\times 4}$. Die Ausreißerzellen liegen jedoch auf der Hauptdiagonale in $n_{11}$ und $n_{22}$. Es werden also unterschiedliche Parameterschätzungen beeinflusst.

5. Es gelte das Modell aus Szenario 3 mit $\boldsymbol{\beta}_2$ und $\widetilde{\mathbf{X}}_{4\times 4}$, wobei die ersetzten Werte in $n_{11}$ und $n_{12}$ extremer sind, da nun $\delta = 10^{-8}$ verwendet wird.

6. $10 \times 10$ Kontingenztafeln mit

$$\begin{aligned}\boldsymbol{\beta}_3 = (&3.3, 0.2, -0.2, 0.1, 0.25, 0.3, -0.1, 0.4, 0.2, 0.1,\\ &0.2, -0.4, 0.2, -0.2, 0.1, 0.0, 0.1, -0.3, 0.1)^\top\end{aligned}$$

   und $\widetilde{\mathbf{X}}_{10\times 10}$, der Effektkodierungsdesignmatrix im Unabhängigkeitsmodell wie in Abschnitt 3.1 beschrieben, werden generiert. Die Ausreißer werden in $n_{11}$ und $n_{23}$ eingesetzt, wobei $\delta = 10^{-4}$.

7. Es gelte das Modell aus Szenario 6 mit $\delta$-Ausreißern in $n_{11}$ und $n_{12}$, wobei $\delta = 10^{-4}$.

Für die hier thematisierten Ausreißeridentifizierer wird $\alpha = 0.01$ verwendet. Zur Wahl des Tuningparameters im OMPC- und OMPCL1-Algorithmus werden gemäß der obigen Szenarien generierte Datensätze ohne Ausreißer hinzugezogen. Abbildung 4.1 zeigt das Verhältnis der Richtig-Negativ-Rate zum Parameter $\zeta$. Ziel ist es, einen möglichst universalen Tuningparameter zu finden, für den die RNR über $95\%$ liegt. Je kleiner die Tafel ist, desto ähnlicher sind sich die Kurven von OMPC und OMPCL1 im Verlauf. Der OMPC-Algorithmus ist dabei jeweils konservativer als der

OMPCL1-Algorithmus. Für eher kleine Tafeln liefert $\zeta = 0.75$ eine RNR über $95\%$. Im $10 \times 10$-Fall wird die Anzahl der verwendeten Minimalmuster generell auf $500$ beschränkt. Für den OMPC-Algorithmus ist hier bereits $\zeta = 0.6$ hinreichend, um eine RNR über $95\%$ zu erhalten. Für den OMPCL1-Algorithmus muss $\zeta$ deutlich größer gewählt werden, damit nicht zu viele Nicht-Ausreißer als Ausreißer klassifiziert werden. Im Folgenden wird dann $\zeta = 0.9$ festgelegt.

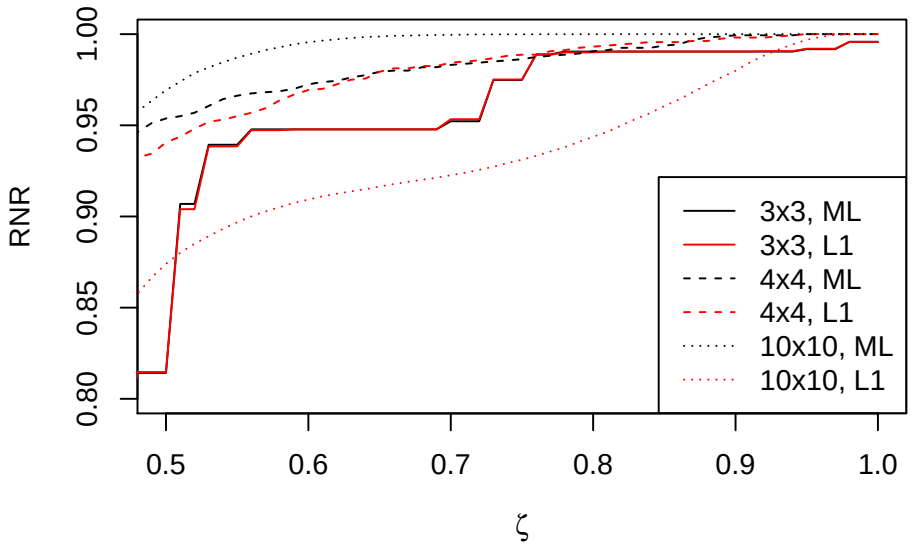



Abbildung 4.1: Richtig-Negativ-Rate des OMPC- und OMPCL1-Algorithmus für $\zeta \in [0.5, 1]$

Tabelle 4.2 zeigt an, wie häufig über die je 500 generierten Kontingenztafeln hinweg Ausreißer bzw. Nicht-Ausreißer korrekt identifiziert wurden. Jedes „A“ steht für einen Antitypen, jedes „T“ für einen Typen. In Tabelle 4.3 ist abzulesen, wie hoch der Anteil der als Ausreißer identifizierten Nicht-Ausreißer in jedem Verfahren ist, wenn die Kontingenztafeln frei von $0.01$-Ausreißern sind. Dabei scheint die Größe der Tafel einen gewissen monotonen Einfluss auf die RNR der Verfahren zu haben. Durch die Wahl des Tuningparameters ist dieses Phänomen bei OMPC- und OMPCL1-Algorithmus weniger ausgeprägt.

Die Ergebnisse können wie folgt zusammengefasst werden:

1. In der $3 \times 3$ Tafel (Szenario 1) sind die Ergebnisse von OL1, OMPC und OMPCL1 relativ ähnlich. Wegen der Exact-Fit-Eigenschaft müssten die Erkennungsraten bei OMPC und OMPCL1 exakt gleich sein, da auch die ML- und

Tabelle 4.2: Richtig-Positiv-Rate (RPR) und Richtig-Negativ-Rate (RNR) der vier eindeutigen Ausreißeridentifizierer in sieben Szenarien mit Antitypen (A) und Typen (T)

| | | *OL1* | | *OMPC* | | *OMPCL1* | | *OLTCS* | |
|---|---|---|---|---|---|---|---|---|---|
| Szenario | | RPR | RNR | RPR | RNR | RPR | RNR | RPR | RNR |
| 1 | A | 0.428 | 0.961 | 0.354 | 0.960 | 0.356 | 0.958 | 0.536 | 0.951 |
| | T | 0.392 | 0.965 | 0.384 | 0.972 | 0.384 | 0.970 | 0.500 | 0.950 |
| 2 | A | 0.636 | 0.982 | 0.668 | 0.980 | 0.660 | 0.985 | 0.724 | 0.942 |
| | T | 0.552 | 0.982 | 0.634 | 0.985 | 0.498 | 0.986 | 0.642 | 0.937 |
| 3 | AA | 0.202 | 0.964 | 0.023 | 0.957 | 0.114 | 0.970 | 0.274 | 0.897 |
| | AT | 0.715 | 0.983 | 0.878 | 0.987 | 0.750 | 0.985 | 0.657 | 0.933 |
| | TT | 0.109 | 0.971 | 0.043 | 0.973 | 0.036 | 0.977 | 0.242 | 0.907 |
| 4 | AA | 0.737 | 0.976 | 0.805 | 0.964 | 0.798 | 0.973 | 0.805 | 0.941 |
| | AT | 0.452 | 0.970 | 0.397 | 0.969 | 0.412 | 0.973 | 0.642 | 0.925 |
| | TT | 0.589 | 0.970 | 0.728 | 0.976 | 0.577 | 0.974 | 0.659 | 0.923 |
| 5 | AA | 0.567 | 0.943 | 0.208 | 0.918 | 0.586 | 0.954 | 0.341 | 0.871 |
| | AT | 0.953 | 0.982 | 0.995 | 0.984 | 0.975 | 0.978 | 0.841 | 0.936 |
| | TT | 0.259 | 0.944 | 0.147 | 0.949 | 0.144 | 0.960 | 0.383 | 0.896 |
| 6 | AA | 0.956 | 0.991 | 0.923 | 0.991 | 0.004 | 0.982 | 0.873 | 0.935 |
| | AT | 0.930 | 0.991 | 0.929 | 0.994 | 0.001 | 0.982 | 0.841 | 0.935 |
| | TT | 0.936 | 0.992 | 0.970 | 0.995 | 0.000 | 0.983 | 0.821 | 0.935 |
| 7 | AA | 0.965 | 0.990 | 0.955 | 0.990 | 0.002 | 0.981 | 0.897 | 0.936 |
| | AT | 0.967 | 0.991 | 0.988 | 0.993 | 0.007 | 0.983 | 0.866 | 0.937 |
| | TT | 0.888 | 0.991 | 0.924 | 0.994 | 0.000 | 0.982 | 0.799 | 0.931 |

Tabelle 4.3: Richtig-Negativ-Rate (RNR) der vier eindeutigen Ausreißeridentifizierer in den drei Szenarien ohne $0.01$-Ausreißer

| | *OL1* | *OMPC* | | *OMPCL1* | | *OLTCS* |
|---|---|---|---|---|---|---|
| Szenario | RNR | RNR | $\zeta$ | RNR | $\zeta$ | RNR |
| 1 | 0.987 | 0.989 | 0.75 | 0.989 | 0.75 | 0.968 |
| 2 | 0.988 | 0.990 | 0.75 | 0.978 | 0.75 | 0.944 |
| 6 | 0.991 | 0.996 | 0.60 | 0.980 | 0.90 | 0.936 |

$L_1$-Schätzer bei gleichem Minimalmuster identisch sind. Der Algorithmus zur Bestimmung des $L_1$-Schätzers spiegelt das in sehr wenigen Fällen nicht wider. Die beste Richtig-Positiv-Rate hat der OLTCS-Identifizierer mit etwa $50\%$.

2. Auch für $4 \times 4$ Tafeln mit einem Ausreißer (Szenario 2) hat OLTCS die be-

sten RPR, wobei die RNR der anderen drei Verfahren deutlich höher sind. Im Antityp-Fall ist der Unterschied zwischen den Verfahren bei Berücksichtigung der RNR eher klein, im Typ-Fall stellt der OMPC-Identifizierer wegen hoher RPR und RNR einen guten Kompromiss dar. Des Weiteren fällt auf, dass die Antitypen für alle Verfahren häufiger erkannt werden als die Typen.

3. Bei zwei moderaten Ausreißern (Szenario 3) in der $4 \times 4$ Tafel fällt generell auf, dass die RNR bei gleichgerichteten Ausreißern etwas (OMPC, OMPCL1) oder deutlich (OL1, OLTCS) niedriger sind als bei entgegengesetzten Ausreißern. Noch deutlicher ist der Unterschied bei der RPR. Bei entgegengesetzten Ausreißern hat der OMPC-Identifizierer das beste Ergebnis ($87.8\%$). Bei gleichgerichteten Ausreißern ist der OL1-Identifizierer dem OLTCS trotz niedrigerer RPR vorzuziehen, da seine RNR deutlich höher ist. Bei extremeren Ausreißern (Szenario 5) verbessern sich wie erwartet alle RPR. Die RNR bei gleichgerichteten Ausreißern werden allerdings auch kleiner. Bei entgegengerichteten Ausreißern ist der Unterschied sehr gering. Für zwei Antitypen ist OMPCL1 die beste Methode, für zwei Typen OL1 und für einen Antitypen und einen Typen ist OMPC die beste Wahl.

4. Wenn die Ausreißer verschiedene Parameter verzerren wie in Szenario 4, sind die RPR bei gleichgerichteten Ausreißern in allen Fällen deutlich besser als in Szenario 3 bzw. 5. Bei entgegengerichteten Ausreißern sind die RPR hingegen viel schlechter. Im Fall von zwei Antitypen sind alle Verfahren ähnlich gut, in dieser Simulationsstudie ist OMPCL1 jedoch der beste Kompromiss. Für zwei Typen ist OMPC deutlich besser als die anderen Verfahren. Für zwei entgegengesetzte Ausreißer ist das Resultat weniger eindeutig: OLTCS hat zwar die mit Abstand größte RPR, aber auch die deutlich kleinste RNR. Eine etwas konservativere Wahl wäre der OL1-Identifizierer.

5. Für die $10 \times 10$ Tafel (Szenario 6) sind OL1- und OMPC-Identifizierer klar die beste Wahl. Sowohl RPR als auch RNR liegen über $92\%$. Die im Verhältnis wenigen Ausreißer, die sich zudem in unterschiedlichen Zeilen und Spalten befinden, haben keinen spürbaren Einfluss auf den $L_1$-Schätzer. Der OMPC-Algorithmus ist besser bei Typen, der OL1-Algorithmus bei Antitypen. Der

OMPCL1-Identifizierer scheint für große Tafeln weniger geeignet zu sein, da kaum Ausreißer erkannt werden. Nur 500 Minimalmuster zur Bestimmung des OMPCL1-Identifizierers scheinen nicht auszureichen.

6. In Szenario 7 bestätigen sich die Erkenntnisse aus Szenario 6, obwohl die Ausreißer nun in derselben Zeile sind. Nur bei der Erkennung zweier Typen sind die Resultate von OL1 und OMPC etwas schlechter als in Szenario 6, ansonsten sind sie besser. Insgesamt ist OMPC etwas besser geeignet als OL1. Für die übrigen Identifizierer lassen sich keine auffälligen Veränderungen in Bezug auf Szenario 6 erkennen.

Diese Simulationsstudie liefert keinen klaren Favoriten unter den Ausreißeridentifizierern. Generell ist OLTCS liberaler als die anderen Verfahren, erkennt dadurch oft aber auch mehr Ausreißer korrekt. Ein liberaleres Vorgehen kann jedoch auch durch einen kleineren Tuningparameter im OMPC-Algorithmus oder durch ein größeres $\alpha$ erreicht werden. In der Regel ist der OMPC- dem OMPCL1-Identifizierer vorzuziehen, vor allem bei großen Kontingenztafeln. In Bezug auf den OL1-Identifizierer fällt auf, dass häufig die Faustregel „*Wenn der OL1-Identifizierer gut ist, ist der OMPC-Identifizierer besser, und wenn der OL1-Identifizierer schlecht ist, ist der OMPC-Identifizierer schlechter*“ gilt. „Schlecht“ ist der OL1-Identifizierer meist dann, wenn gleichgerichtete Ausreißer in derselben Zeile oder Spalte oder entgegengerichtete Ausreißer in unterschiedlichen Zeilen und Spalten vorliegen. Die Wahl des Identifizierers sollte idealerweise davon abhängen, welches Ausreißerszenario in der Datensituation naheliegender ist. In der Praxis beinhaltet ein gängiges Szenario gleichgerichtete Ausreißer auf der Hauptdiagonale einer quadratischen Tafel, siehe Beispiel 4.15. Daher ist die Verwendung des OMPC-Identifizierers in quadratischen Tafeln zu empfehlen, wenn keine weiteren Informationen über naheliegende Ausreißerszenarien vorliegen und keine Beschränkungen bzgl. der Rechenzeit bestehen.

### *4.1.3 Fallstudien*

In diesem Abschnitt werden die vorgestellten Ausreißeridentifikationsverfahren OL1, OMPC, OMPCL1 und OLTCS auf echte Daten angewendet.

BEISPIEL 4.15. Sozialer Status von Vätern und Söhnen

*In diesem Beispiel wird untersucht, ob in einer Kontingenztafel bzgl. des sozialen Status von Vätern und Söhnen Ausreißer identifiziert werden können. Hierzu stellt Brown (1974) anhand einer Kontingenztafel mit 775 Beobachtungen von Pearson (1904) fest, dass Söhne dazu tendieren, einen Beruf in derselben beruflichen Kategorie zu erlernen wie ihre Väter. Falls sie dies jedoch nicht tun, sei der Beruf des Sohnes in der Regel unabhängig vom Beruf des Vaters, so Brown. Die Ergebnisse einer solchen Befragung sollten demnach – wenn sie durch ein Unabhängigkeitsmodell repräsentiert werden – Hinweise auf Ausreißer (Typen) auf der Hauptdiagonale liefern. Ein neuerer und größerer Datensatz zu diesem Thema wird von Glass und Berent (1954) mit einer $7 \times 7$ Kontingenztafel beschrieben. Goodman (1971) fasst die Klassen zu den Kategorien hoch, mittel und niedrig zusammen, siehe Tabelle 4.4.*

Tabelle 4.4: Zellhäufigkeiten des sozialen Status von Vätern und ihren Söhnen (links) und Ergebnisse der Ausreißeridentifizierer (rechts) mit $\alpha = 0.01$ und $\zeta = 0.75$

| *Kontingenztafel* | | | | *Ausreißeridentifizierer* | | | | |
|---|---|---|---|---|---|---|---|---|
| Zelle | Vater | Sohn | Anzahl | OL1 | OMP | OMPC | OMPCL1 | OLTCS |
| $n_{11}$ | hoch | hoch | 588 | * | * | * | * | * |
| $n_{21}$ | mittel | hoch | 395 | | | * | * | |
| $n_{31}$ | niedrig | hoch | 159 | * | | * | * | * |
| $n_{12}$ | hoch | mittel | 349 | | | * | * | |
| $n_{22}$ | mittel | mittel | 714 | | * | * | * | |
| $n_{32}$ | niedrig | mittel | 447 | | | * | * | |
| $n_{31}$ | hoch | niedrig | 114 | * | | * | * | * |
| $n_{32}$ | mittel | niedrig | 320 | | | * | * | |
| $n_{33}$ | niedrig | niedrig | 411 | * | * | * | * | * |

*Die OMP-Methode liefert nur die Beobachtungen auf der Hauptdiagonale als* $0.01$*-Ausreißer und damit das erwartete Ergebnis. Die OL1- sowie die OLTCS-Methode identifizieren* $n_{11}, n_{13}, n_{31}$ *und* $n_{33}$ *als Ausreißer. Dies erscheint ebenfalls sinnvoll, da es eher selten passiert, dass sich der soziale Status innerhalb einer Generation vom einen ins andere Extrem verändert. Allerdings fehlt in diesem Ausreißermuster die Beobachtung* $n_{22}$*.*

*Die OMPC- und OMPCL1-Methode identifizieren jede der neun Zellen als Ausreißer. Dieses überraschende Ergebnis ist sinnvoll, da es darauf hinweist, dass das zugrunde liegende Unabhängigkeitsmodell ungeeignet zur Beschreibung der Struktur der Tafel ist.*

BEISPIEL 4.16. Hochgeschwindigkeitsflammspritzen *(Fortsetzung von Beispiel 3.3) In Tabelle 3.3 ist eine zweidimensionale Kontingenztafel bezüglich der Form und Größe einzelner Splats dargestellt. In der Fortsetzung dieses Beispiels liegt mit der Position der Splats eine weitere Dimension vor. Dabei interessiert bei der Analyse, ob der unter der Unabhängigkeitsannahme auffällig hohe Wert für kleine talförmige Splats nach Hinzunahme der Position als weitere erklärende Dummyvariable als Ausreißer identifiziert wird. Die Kontingenztafel sowie die Ergebnisse der Ausreißeranalyse sind Tabelle 4.5 zu entnehmen.*

Tabelle 4.5: Zellhäufigkeiten der Splatmorphologie (links) und Ergebnisse der Ausreißeridentifizierer (rechts) mit $\alpha = 0.01$ und $\zeta = 0.75$

| *Kontingenztafel* | | | | *Ausreißeridentifizierer* | | | | |
|---|---|---|---|---|---|---|---|---|
| Zelle | Position | Größe | Form | OL1 | OMP | OMPC | OMPCL1 | OLTCS |
| $n_{111} = 27$ | Mitte | groß | Berg | | | | | |
| $n_{211} = 18$ | Rand | groß | Berg | | | | | |
| $n_{121} = 7$ | Mitte | klein | Berg | | | | | |
| $n_{221} = 7$ | Rand | klein | Berg | | | | | |
| $n_{112} = 25$ | Mitte | groß | flach | | | | | |
| $n_{212} = 23$ | Rand | groß | flach | | | | | |
| $n_{122} = 10$ | Mitte | klein | flach | | | | | |
| $n_{222} = 13$ | Rand | klein | flach | | | | | |
| $n_{113} = 24$ | Mitte | groß | Tal | | | | * | * |
| $n_{213} = 6$ | Rand | groß | Tal | * | | * | * | * |
| $n_{123} = 46$ | Mitte | klein | Tal | * | | * | * | |
| $n_{223} = 33$ | Rand | klein | Tal | * | | | * | |

*In Bemerkung 4.6 werden Minimalmuster in höherdimensionalen Tafeln behandelt. Die OMP-Methode findet basierend auf dem Muster* $n_{111}, n_{121}, n_{212}, n_{122}, n_{213}, n_{123}, n_{223}$ *als einzige keine Ausreißer. Die anderen Verfahren identifizieren zwei bis vier Ausreißer, die alle die Anzahl der talförmigen Splats betreffen. Besonders markant ist, dass der OMPCL1-Identifizierer alle vier Zellhäufigkeiten talförmiger Splats identifiziert und damit die Unabhängigkeitsannahme des loglinearen Poissonmodells am deutlichsten in Frage stellt. Am häufigsten wird die Anzahl großer, talförmiger, am Rand befindlicher Splats als Ausreißer, genauer gesagt als Antityp, identifiziert.*

BEISPIEL 4.17. Soziale Netzwerke junger Mütter

*McKinlay (1973) untersucht in einer Studie anhand von Frauen aus der Arbeiterklasse Aberdeens, wie intensiv Schwangere soziale Dienste in Anspruch nehmen. Die Merkmale dieser dreidimensionalen Kontingenztafel (siehe Tab. 4.6) sind:*

- $\mathcal{D}_1$: *Häufigkeit der Interaktion mit Freunden (täglich, mindestens einmal pro Woche, weniger als einmal pro Woche),*
- $\mathcal{D}_2$: *Entfernung zwischen eigener Wohnung und Wohnung der Freunde (zu Fuß erreichbar, per Bus erreichbar),*
- $\mathcal{D}_3$: *ob die Schwangere ihr erstes oder ein weiteres Kind austrägt.*

*Für das Modell wird die bedingte Abhängigkeit von $\mathcal{D}_1$ und $\mathcal{D}_3$ gegeben $\mathcal{D}_2$ angenommen, was mit der Designmatrix $\mathbf{X}$ in* (A.4)*, Anhang A.5 einhergeht. Für die OL1-Methode werden die beiden absoluten Extrema der Tafel als Ausreißer identifiziert: $n_{111} = 30$ und $n_{121} = 2$. Die OLTCS-Methode identifiziert nur Zelle $n_{121}$ als Ausreißer.*

*Um Ergebnisse der Minimalmustermethode für diese Situation zu erhalten, muss zunächst der veränderten Modellannahme Rechnung getragen werden. Von den $\binom{12}{8} = 495$ achtelementigen Teilmengen erfüllen $144$ Definition 4.4. Die Minimalmuster ergeben $40$ dreielementige, $88$ zweielementige und $16$ einelementige Ausreißermuster. Das OMP-Verfahren findet je acht Mal in Zelle $n_{121}$ bzw. $n_{122}$ den einzigen Ausreißer. Dieses Verfahren verursacht zwei unterschiedliche Lösungen. Das OMPC-Verfahren liefert ähnliche Resultate. Jede Zelle ist in $48$ Minimalmustern nicht enthalten. Die Zellen $n_{111}, n_{121}$ und $n_{122}$ werden $48$ Mal als Ausreißer identifiziert, wohingegen die Zellen $n_{311}$ und $n_{312}$ überhaupt nicht identifiziert werden. Die übrigen Zellen werden in der Hälfte der Fälle als Ausreißer identifiziert und somit als nicht als OMPC-Ausreißer angesehen. Das OMPCL1-Verfahren identifiziert dieselben Zellen wie das OMPC-Verfahren als Ausreißer. Eine Zusammenfassung der Ergebnisse ist in Tabelle 4.6 zu finden.*

*Diese Kontingenztafel wurde bereits von weiteren Autoren in Bezug auf Ausreißer analysiert. Upton (1980) und Upton und Guillen (1995) stellen fest, dass $n_{122}$ als Ausreißer angesehen werden sollte: Viele Frauen, die zum ersten Mal schwanger werden, sind unmittelbar davor berufstätig gewesen. Ihre sozialen Kontakte beinhalten auch Arbeitskollegen,*

Tabelle 4.6: Zellhäufigkeiten des sozialen Netzwerke junger Mütter (links) und Ergebnisse der Ausreißeridentifizierer (rechts) mit $\alpha = 0.01$ und $\zeta = 0.75$

| *Kontingenztafel* | | | | *Ausreißeridentifizierer* | | | | |
|---|---|---|---|---|---|---|---|---|
| Zelle | Frequenz | Weg | Kind | OL1 | OMP | OMPC | OMPCL1 | OLTCS |
| $n_{111} = 30$ | täglich | zu Fuß | 1 | * | | * | * | |
| $n_{112} = 6$ | täglich | zu Fuß | 2+ | | | | | |
| $n_{121} = 2$ | täglich | Bus | 1 | * | *[1] | * | * | * |
| $n_{122} = 13$ | täglich | Bus | 2+ | | *[2] | * | * | |
| $n_{211} = 19$ | wöchentlich | zu Fuß | 1 | | | | | |
| $n_{212} = 12$ | wöchentlich | zu Fuß | 2+ | | | | | |
| $n_{221} = 16$ | wöchentlich | Bus | 1 | | | | | |
| $n_{222} = 8$ | wöchentlich | Bus | 2+ | | | | | |
| $n_{311} = 5$ | seltener | zu Fuß | 1 | | | | | |
| $n_{312} = 2$ | seltener | zu Fuß | 2+ | | | | | |
| $n_{321} = 10$ | seltener | Bus | 1 | | | | | |
| $n_{322} = 4$ | seltener | Bus | 2+ | | | | | |

*die nicht in der Nähe wohnen. Diese Zelle wird nur von den Minimalmuster-basierten Verfahren entdeckt. Allerdings identifizieren diese auch weitere Zellen als Ausreißer. Wenn von einem sinnvollen Modell ausgegangen werden kann, aber aus Sicht des Anwenders zu viele Ausreißer entdeckt werden, kann ein kleineres $\alpha$ oder eine andere Wahl des Tuningparameters im OMPC- oder OMPCL1-Algorithmus für ein zufriedenstellenderes Ergebnis sorgen. Wenn davon ausgegangen wird, dass hier ein Unabhängigkeitsmodell vorliegt, ändern sich die identifizierten Ausreißermengen kaum. OMP identifiziert keine Ausreißer mehr, OMPC und OMPCL1 nur noch $n_{111}$ und $n_{121}$. Die restlichen Verfahren identifizieren dieselben Beobachtungen wie zuvor im bedingten Modell als Ausreißer.*

## 4.2 IDENTIFIKATION GANZER TAFELN ALS AUSREISSER

Liegen $N$ Kontingenztafeln $\mathbf{N}_1, \ldots, \mathbf{N}_N \in \mathbb{R}^{k_1 \times k_2}$ mit fester Stichprobengröße $n$ vor, ist es möglich, eine oder mehrere von ihnen als $\alpha$-Ausreißer zu identifizieren. Unter der Bedingung, dass alle Kontingenztafeln als Realisationen von $\boldsymbol{\mathcal{N}}$ mit $\mathrm{vec}(\boldsymbol{\mathcal{N}}) \sim Mult(n, \boldsymbol{\theta})$ aufgefasst werden, lässt sich – wie in Abschnitt 2.1, Algorithmus 1 beschrieben – eine gemeinsame $\alpha$-Ausreißerregion finden. Im allgemeine-

ren Fall, dass unterschiedliche Stichprobengrößen $n_1, \ldots, n_N$ vorliegen, lässt sich für jede Stichprobengröße eine separate $\alpha$-Ausreißerregion herleiten. In der Praxis wird bei gleicher oder unterschiedlicher Stichprobengröße zunächst die Unabhängigkeitsstruktur festgelegt und darauf basierend die Parameter des Multinomialmodells geschätzt. Das folgende Beispiel veranschaulicht die Anwendung der vorgestellten Modelle und $\alpha$-Ausreißer im Fall gleicher Stichprobengrößen. Dabei werden auch naive Schätzer hinzugezogen, die die Struktur der Kontingenztafeln nicht berücksichtigen: Das arithmetische Mittel $\widehat{\boldsymbol{\theta}}_{\text{aM}}$, der Median $\widehat{\boldsymbol{\theta}}_{\text{Med}}$ und das $0.25$-getrimmte Mittel $\widehat{\boldsymbol{\theta}}_{\text{0.25-tM}}$. Die zwei letztgenannten Schätzer werden pro Zelle betrachtet und normiert, da sich – nach der Trunkierung – die geschätzten Zellwahrscheinlichkeiten nicht notwendigerweise zu $1$ summieren.

BEISPIEL 4.18. *Gegeben seien acht Realisationen einer $2\times 2$ Kontingenztafel, wobei $\boldsymbol{\theta}$ mit $\theta_{11} = \frac{2}{5}, \theta_{12} = \frac{4}{15}, \theta_{21} = \frac{1}{5}, \theta_{22} = \frac{2}{15}$ und $n = 75$. Damit gilt: $\theta_{ij} = \theta_{i+}\theta_{+j}, i = 1, 2, j = 1, 2$. In der Tafel $\mathbf{N}_1$ seien $n_{12}$ und $n_{21}$ vertauscht eingetragen. Mit Hilfe diverser Schätzer wurde die exakte $\alpha$-Ausreißerregion bestimmt ($\alpha = 0.1$) und daraufhin jede Tafel als Ausreißer oder als Nicht-Ausreißer klassifiziert, siehe Tabelle 4.7.*

Tabelle 4.7: Acht beispielhafte Kontingenztafeln

| | | | | | *Ausreißer erkannt mittels* | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $n_{11}$ | $n_{12}$ | $n_{21}$ | $n_{22}$ | $\boldsymbol{\theta}$ | $\widehat{\boldsymbol{\theta}}_{\text{aM}}$ | $\widehat{\boldsymbol{\theta}}_{\text{Med}}$ | $\widehat{\boldsymbol{\theta}}_{\text{0.25-tM}}$ | $\widehat{\boldsymbol{\beta}}^{\text{ML}}$ | $\widehat{\boldsymbol{\beta}}^{L_1}$ |
| $\mathbf{N}_1$ | 32 | 8 | 22 | 13 | ja | ja | ja | ja | ja | ja |
| $\mathbf{N}_2$ | 37 | 17 | 16 | 5 | nein | nein | nein | nein | nein | nein |
| $\mathbf{N}_3$ | 30 | 22 | 15 | 8 | nein | nein | nein | nein | nein | nein |
| $\mathbf{N}_4$ | 28 | 27 | 10 | 10 | nein | nein | nein | nein | nein | nein |
| $\mathbf{N}_5$ | 31 | 22 | 14 | 8 | nein | nein | nein | nein | nein | nein |
| $\mathbf{N}_6$ | 30 | 18 | 17 | 10 | nein | nein | nein | nein | nein | nein |
| $\mathbf{N}_7$ | 28 | 27 | 10 | 10 | nein | nein | nein | nein | nein | nein |
| $\mathbf{N}_8$ | 26 | 21 | 21 | 7 | nein | nein | nein | nein | nein | nein |

*Korrekterweise wurde Tafel $\mathbf{N}_1$ von allen Verfahren als einzige als $0.1$-Ausreißer identifiziert.*

Im Folgenden werden die Ergebnisse einer kurzen Simulationsstudie vorgestellt. Dabei wurden die Parameter aus Beispiel 4.18 als Grundlage verwendet. Von den 100 Wiederholungen zufällig generierter $P = Mult(75, (\frac{2}{5}, \frac{4}{15}, \frac{1}{5}, \frac{2}{15})^\top)$-verteilter

Stichproben mit je acht Tafeln waren 49 Läufe der je sieben unkontaminierten Tafeln keine $0.1$-Ausreißer bezüglich ihrer wahren Verteilung. Die folgende Auswertung beschränkt sich auf diese Läufe. Des Weiteren ist festzuhalten, dass durch das Vertauschen der Nebendiagonalelemente der Kontingenztafel $\mathbf{N}_1$ nur $40.8\%$ der Tafeln zu $0.1$-Ausreißern bezüglich der wahren Verteilung wurden. Ob in der Praxis auch die übrigen kontaminierten Tafeln identifiziert werden sollen, hängt von der konkreten Anwendung ab. In Tabelle 4.8 wird nur der Anteil der Tafeln angegeben, die wie gewünscht klassifiziert wurden.

Tabelle 4.8: Simulationsergebnisse: $0.1$-Ausreißer im Multinomial-Unabhängigkeitsmodell

| | *Parametervektor für $P$* | | | | | |
|---|---|---|---|---|---|---|
| | $\boldsymbol{\theta}$ | $\widehat{\boldsymbol{\theta}}_{\text{aM}}$ | $\widehat{\boldsymbol{\theta}}_{\text{Med}}$ | $\widehat{\boldsymbol{\theta}}_{\text{0.25-tM}}$ | $\widehat{\boldsymbol{\beta}}^{\text{ML}}$ | $\widehat{\boldsymbol{\beta}}^{L_1}$ |
| Anteil der $\mathbf{N}_1 \in \text{out}(0.1, P)$ | 0.408 | 0.286 | 0.347 | 0.347 | 0.327 | 0.408 |
| Anteil der $\mathbf{N}_2, \ldots, \mathbf{N}_8 \notin \text{out}(0.1, P)$ | 1.000 | 0.980 | 0.962 | 0.971 | 0.988 | 0.965 |

Die Anteile der korrekt erkannten Nicht-Ausreißer sind bei allen Verfahren ähnlich und sehr gut. Bezüglich der geschätzten Parameter der Multinomialverteilung und den daraus resultierenden $\alpha$-Ausreißerregionen gibt es deutliche Unterschiede. Die kontaminierte Tafel $\mathbf{N}_1$ wird am schlechtesten vom Identifizierer basierend auf dem arithmetischen Mittel erfasst, da alle Tafeln mit demselben Gewicht eingehen und zusätzlich die Information über die Unabhängigkeit der Tafel nicht genutzt wird. Ähnlich gut erkennen die Identifizierer basierend auf Median, getrimmten Mittel und ML-Schätzer die kontaminierte Tafel, da sie entweder robust gegenüber Ausreißern sind oder die Unabhängigkeit der Merkmale der Kontingenztafel mit berücksichtigen. Der Identifizierer basierend auf dem robusten $L_1$-Schätzer hat eine gleich hohe Erkennungsrate der kontaminierten Tafel wie der Identifizierer basierend auf dem wahren Parametervektor. Es ist allerdings nicht so, dass alle wahren $0.1$-Ausreißer auch von $\widehat{\boldsymbol{\beta}}^{L_1}$ erkannt würden: Jeweils vier wahre $0.1$-Ausreißer werden als Nicht-Ausreißer klassifiziert und umgekehrt.

Nach dieser umfassenden Untersuchung der wichtigsten Ausreißer-Szenarien in Kontingenztafeln werden im folgenden Kapitel 5 Methoden entwickelt und eruiert, um Ausreißer in funktionalen Daten identifizieren zu können.

# 5. Ausreisseridentifikation in funktionalen Daten

Ähnlich wie im Fall kategorialer Daten kann sich im Fall funktionaler Daten der unbekannte datengenerierende Prozess verändern. Dies kann unerwünschte Folgen haben, z. B. Parameterschätzer in Modellen verzerren oder Klassifikationsverfahren unbrauchbar machen. Betreffen die Veränderungen des datengenerierenden Prozesses einen kleinen Anteil der funktionalen Beobachtungen, so werden diese als „Ausreißer in funktionalen Daten“ oder „funktionale Ausreißer“ bezeichnet. Des Weiteren werden so auch solche Beobachtungen genannt, die extrem von dem Muster abweichen, das durch den datengenerierenden Mechanismus erzeugt wird.

Im vorigen Kapitel wurden Einträge der Kontingenztafel als abhängige, vektorwertige Zielgrößen aufgefasst. Im funktionalen Fall muss zwischen unabhängigen und abhängigen funktionalen Beobachtungen unterschieden werden. Mit unabhängigen Beobachtungen sind in diesem Kapitel solche Beobachtungen gemeint, die als Realisationen eines zeitstetigen stochastischen Prozesses $\mathcal{Y} = \{\mathcal{Y}(t)|t \in \mathcal{T}\}$ aufgefasst werden. Folglich muss lediglich ein geeignetes Ausreißeridentifikationsverfahren (siehe Bootstrap-Algorithmus) auf den kompletten funktionalen Datensatz angewendet werden. Bei abhängigen Beobachtungen wird davon ausgegangen, dass bestimmte Kovariablen einen Einfluss auf die Parameter des datengenerierenden stochastischen Prozesses haben. Durch die Annahme eines Modells kann der datengenerierende Prozess – bis auf die unbekannten Parameter – als bekannt angesehen werden. Diese Kenntnis ist in das Ausreißeridentifikationsverfahren zu integrieren. Eine flexible Klasse zur Beschreibung abhängiger funktionaler Beobachtungen ist die in Abschnitt 3.2.2 vorgestellte *generalisierte function-on-scalar regression* (GFOSR). Über die Verwendung geeigneter funktionaler Residuen (z. B. Devianz, Pearson,

Anscombe) können statt der abhängigen Zielgrößen die – unter den Modellannahmen – unabhängig identisch verteilten Residuen auf das Vorhandensein potenzieller Ausreißer analysiert werden. Daher wird der Fokus dieses Kapitels auf der Ausreißeridentifikation unabhängiger funktionaler Daten liegen. In Abschnitt 5.3 wird zudem anhand einer Simulationsstudie gezeigt, wie die im Folgenden vorgestellten Verfahren im GFOSR-Kontext zu beurteilen sind.

Ein weiterer Unterschied zur Ausreißeridentifikation in Kontingenztafeln besteht darin, dass nicht nur unerwartet große („Typ“) und kleine („Antityp“) Beobachtungen als Ausreißer klassifiziert werden können, sondern dass auch im Hinblick auf Ausmaß und Form der funktionalen Daten differenziert werden muss. Dementsprechend unterscheiden Lopez-Pintado und Romo (2009) zwischen *magnitude* (Ausmaß, Lage) und *shape* (Form) Ausreißern. Die Relevanz der Form wird auch von Horváth und Kokoszka (2012, S. 5) hervorgehoben: „For functional data the information contained in the *shape* of the curves matters a great deal.“ Abbildung 5.1 zeigt links zwei gestrichelte Ausmaß-Ausreißer, die weit entfernt von einer Schar grauer Kurven liegen, aber prinzipiell der gleichen Form folgen. Auf der rechten Seite von Abbildung 5.1 liegt der gestrichelte Form-Ausreißer innerhalb der Schar grauer Kurven, hat dabei aber eine deutlich abweichende Form.

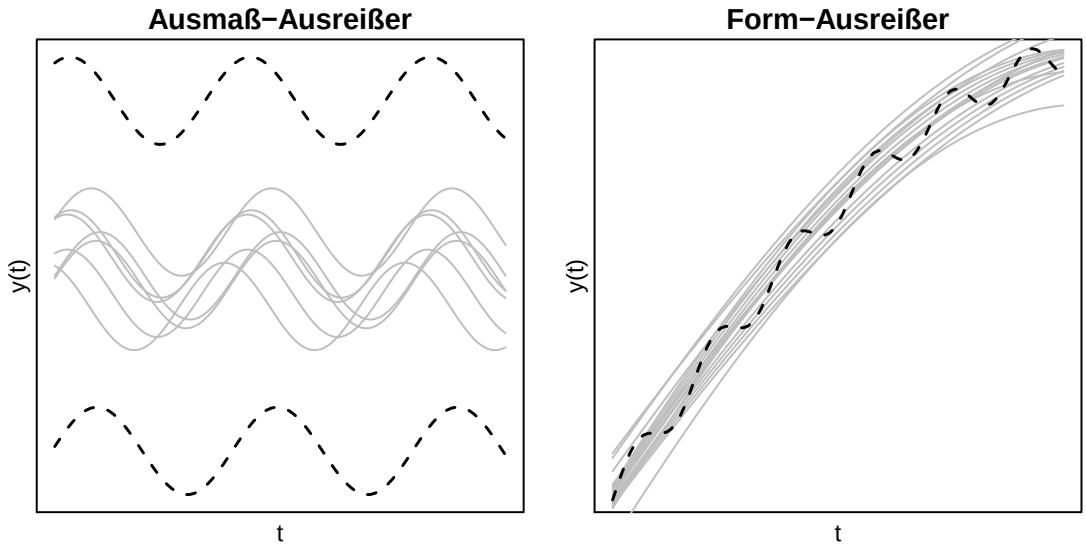



Abbildung 5.1: Ausmaß- und Form-Ausreißer (schwarz, gestrichelt) in einer Schar regulärer künstlich erzeugter Beobachtungen (grau, durchgezogen)

Ausmaß-Ausreißer können bereits häufig mittels klassischer univariater Analysemethoden wie dem arithmetischen Mittel einer Funktion entdeckt werden. Bei Form-Ausreißern wie in Abbildung 5.1 rechts wäre solch ein Vorgehen mit dem arithmetischen Mittel nicht erfolgreich. Stattdessen könnte die Varianz der Funktion oder alternativ das Integral über die zweite Ableitung herangezogen werden.

In Situationen, bei denen Mittelwert und Varianz gleich oder sehr ähnlich sind, müssen andere Maßzahlen in Betracht gezogen werden.

BEISPIEL 5.1. *In Abbildung 5.2 sind drei Realisationen eines Gaußprozesses eingezeichnet (durchgezogene, gestrichelte und punktierte Kurve). Die Werte wurden an 100 äquidistanten Zeitpunkten realisiert. Die abgebildeten Kurven sind so skaliert, dass ihre empirischen Mittelwerte und Varianzen übereinstimmen.*

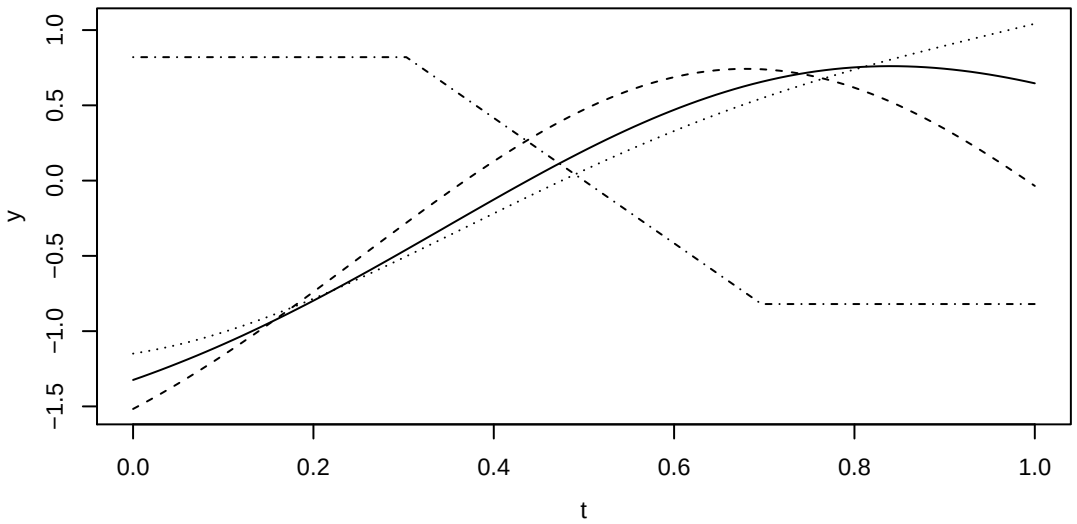



Abbildung 5.2: Vier Funktionen mit gleichem Mittelwert und gleicher Streuung

*Zusätzlich ist die strichpunktierte Funktion*

$$y(t) = \begin{cases} 0.82, & t \in [0, 0.31[ \\ 2.12 - 0.042t, & t \in [0.31, 0.7[ \\ -0.82, & t \in [0.7, 1] \end{cases}$$

*eingetragen, die über dieselben empirische Maße verfügt. Der sich offensichtlich unterschei-*

*dende datengenerierende Mechanismus kann folglich nicht mit Hilfe der genannten Lage- und Streuungsmaßen erkannt werden.*

Für die Identifikation von Ausreißern wird eine Ordnung der Funktionen bezüglich ihrer *outlyingness* benötigt. Das Konzept der Datentiefe (siehe (2.13)), mit dem Tukey (1975) multivariate Daten von innen nach außen ordnet, wurde von einigen Autoren bereits auf funktionale Daten übertragen und steht im Zentrum von Abschnitt 5.1. In den Abschnitten 5.2 und 5.3 wird anhand von realen und künstlichen Datensätzen überprüft, wie gut die Ausreißeridentifikation der funktionalen Datentiefen ist, wenn der datengenerierende Prozess ganz oder teilweise unbekannt ist. In Abschnitt 5.4 werden neue Verfahren zur Bestimmung von Ausreißern für vollständig bekannte Gaußprozesse entwickelt und evaluiert.

## 5.1 DATENTIEFEN

Datentiefen sind ein beliebtes Werkzeug, um Beobachtungen einer Stichprobe in Bezug auf ihre Zentralität „von innen nach außen“ zu ordnen. Die am wenigsten zentralen Beobachtungen können dann als potenzielle Ausreißer angesehen werden.

Das Konzept der Datentiefe wurde von Tukey (1975) eingeführt: Seine Halbraumtiefe (2.13) gilt als wegweisender systematischer Vorschlag, das Rangkonzept für multivariate Datensituationen zu generalisieren und somit die Bestimmung des multivariaten Medians zu ermöglichen. Seither wurden viele alternative Datentiefen (unter anderem Barnett, 1976; Oja, 1983; Liu, 1990; Koshevoy und Mosler, 1997) vorgestellt. Eine Datentiefe soll vor allem anhand der Berechenbarkeit und – abhängig von der Datensituation – der Robustheit gewählt werden (Mosler, 2013).

Die oben genannten Datentiefen können in der Regel nicht ohne Weiteres auf funktionale Datensituationen übertragen werden. Für funktionale Datentiefen existieren jedoch bereits einige spezielle Vorschläge (z. B. die integrierte Datentiefe (Fraiman und Muniz, 2001), $h$-modal Datentiefe (Cuevas *et al.*, 2006, 2007) oder modifizierte Bandtiefe (Lopez-Pintado und Romo, 2009)). Im Allgemeinen ist die Stichprobenversion einer funktionalen Datentiefe eine Abbildung vom Funktionenraum in die

reellen Zahlen, $d : \mathbb{B} \to \mathbb{R}$. Nieto-Reyes und Battey (2016) stellen grundlegende theoretische Anforderungen an funktionale Datentiefen auf: Distanzinvarianz, Maximalität im Zentrum, streng fallend bzgl. des tiefsten Punktes, obere Semistetigkeit in $x$, Aufnahmefähigkeit der Breite der konvexen Hülle, Stetigkeit in $P$. Die Autorinnen halten jedoch auch fest, dass keine der sechs von ihnen betrachteten funktionalen Datentiefen all diese Anforderungen erfüllt.

Um auf Basis einer Datentiefe einen einschrittigen $\alpha$-Ausreißeridentifizierer wie in Definition 2.3 zu konstruieren, müsste ihre Verteilung bekannt sein. Febrero *et al.* (2008) weisen darauf hin, dass die Verteilung funktionaler Datentiefen unbekannt ist. Stattdessen schlagen die Autoren vor, das $p$-Quantil der Verteilung, wobei $p \in ]0, 1[$ vorab zu wählen ist, mit Hilfe des Bootstrap (Efron, 1979) zu approximieren. Der approximierte Wert wird als $C(p)$ bezeichnet. Eine Beobachtung $y \in \mathbb{B}$ wird als $p$-Ausreißer klassifiziert, falls $d(y) < C(p)$ ist. Die explizite Prozedur zur Bestimmung von $C(p)$ ist im folgenden Bootstrap-Algorithmus skizziert, wobei $\widehat{\mathbf{\Sigma}}_{\mathcal{Y}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top, \mathbf{y}_i := (y_i(t_1), \ldots, y_i(t_T))^\top$ und $\bar{\mathbf{y}} := (\frac{1}{n}\sum_{i=1}^{n} y_i(t_1), \ldots, \frac{1}{n}\sum_{i=1}^{n} y_i(t_T))^\top$ mit Hilfe der diskretisierten Beobachtungen bestimmt werden.

**Bootstrap-Algorithmus:** Geglätteter Bootstrap vom Umfang $B$

$\varsigma \leftarrow$ Glättungsparameter für die geschätzte Kovarianzmatrix;
Bestimme funktionalen Tiefewert $d(\mathbf{y}_i), i = 1, \ldots, n$ jeder Kurve;
**for** $b = 1$ **to** $B$ **do**
    Ziehe Standard-Bootstrap-Stichprobe $\mathbf{Y}^b = (\mathbf{y}_1^b, \ldots, \mathbf{y}_n^b)^\top$, in die jede Kurve mit einer Wahrscheinlichkeit proportional zu ihrer Tiefe aufgenommen wird;
    $\widetilde{\mathbf{Y}}^b \leftarrow (\widetilde{\mathbf{y}}_1^b, \ldots, \widetilde{\mathbf{y}}_n^b)^\top$, wobei $\widetilde{\mathbf{y}}_i^b = \mathbf{y}_i^b + \mathbf{e}_i^b, \mathbf{e}_i^b = (e_i^b(t_1), \ldots, e_i^b(t_T))^\top \sim \aleph(\mathbf{0}, \varsigma\widehat{\mathbf{\Sigma}}_{\mathcal{Y}})$;
    $C^b(p) \leftarrow$ empirisches $p$-Quantil von $(d(\widetilde{\mathbf{y}}_1^b), \ldots, d(\widetilde{\mathbf{y}}_n^b))$
**end**
$C(p) \leftarrow \text{median}(\{C^1(p), \ldots, C^B(p)\})$

Ähnlich wie das Minimalmusterverfahren aus Abschnitt 4.1.1 basiert der hier verwendete Bootstrap auf wiederholt gezogenen Teilstichproben. Die asymptotische Validität des geglätteten Bootstrap konnte noch nicht nachgewiesen werden (Horváth und Kokoszka, 2012, S. 8); in der Praxis liefert er jedoch sinnvolle Resultate.

Im folgenden Abschnitt 5.1.1 werden Datentiefen vorgestellt, die vor allem zur

Ordnung von Funktionen bezüglich ihres Ausmaßes genutzt werden. Zwei eigene Pseudo-Datentiefen zur expliziten Erkennung von Form-Ausreißern werden in den Abschnitten 5.1.2 und 5.1.3 entwickelt und in Bezug auf ihre theoretischen Eigenschaften untersucht.

### *5.1.1 Populäre funktionale Datentiefen*

In diesem Abschnitt werden mit der integrierten Datentiefe (Fraiman und Muniz, 2001, FMD), der $h$-modal Datentiefe (Cuevas *et al.*, 2006, 2007, HMD) und der modifizierten Bandtiefe (Lopez-Pintado und Romo, 2009, MBD) drei populäre Datentiefen vorgestellt. Da die *random projection depth* (Cuevas *et al.*, 2007) gegenüber der HMD sowohl einen höheren Rechenaufwand hat als auch sich in empirischen Untersuchungen als weniger geeignet zur Ausreißeridentifikation herausgestellt hat, wird sie hier nicht berücksichtigt (vergleiche auch Febrero *et al.*, 2008).

Die Datentiefen werden nachfolgend in ihrer Stichprobenversion angegeben. In den Definitionen der folgenden Datentiefen sei $\tilde{y} \in \mathbb{B}$ eine beliebige funktionale Beobachtung. Die Datentiefe von $\tilde{y}$ wird bezüglich der Stichprobe $\mathbf{y} = (y_1, \ldots, y_n)^\top$ angegeben.

DEFINITION 5.2. Integrierte Datentiefe *(Fraiman und Muniz, 2001)*
*Sei $\tilde{y} = \{\tilde{y}(t) \mid t \in \mathcal{T} = [a, b]\} \in \mathbb{B}$ eine beliebige funktionale Beobachtung und $\mathbf{y}$ eine gegebene Stichprobe. $F_{n,t}$ bezeichne die empirische Verteilungsfunktion der reellwertigen Beobachtungen $y_1(t), \ldots, y_n(t)$. Dann ist die* integrierte Datentiefe (integrated depth, FMD) *von $\tilde{y}$ gegeben durch*

$$d^{\mathrm{FM}}(\tilde{y}; \mathbf{y}) = \frac{1}{b-a} \int_a^b \left[1 - \left|\frac{1}{2} - F_{n,t}(\tilde{y}(t))\right|\right] dt. \tag{5.1}$$

Die FMD ist die erste Datentiefe, die speziell für funktionale Daten definiert wurde und hängt wie die Stichprobenversion der Halbraumtiefe von der empirischen Verteilungsfunktion ab. Das Maximum $d^{\mathrm{FM}}(\tilde{y}; \mathbf{y}) = 1$ wird erreicht, wenn $F_{n,t}(\tilde{y}(t)) \equiv 0.5$ ist. Somit kann das die FMD maximierende $\tilde{y}$ als funktionales Analogon des

univariaten Medians aufgefasst werden. Mit der HMD hat auch der Modus seine Entsprechung im funktionalen Fall.

DEFINITION 5.3. $h$-modal Datentiefe *(Cuevas* et al.*, 2006, 2007)*
*Sei $\tilde{y} \in \mathbb{B}$ eine beliebige funktionale Beobachtung und $\mathbf{y}$ eine gegebene Stichprobe. Weiter sei $h \in \mathbb{R}^+$ ein gegebener Tuningparameter, $K_h(t) = \frac{1}{h} K_G\left(\frac{t}{h}\right)$ ein reskalierter Gaußkern (siehe* (2.10)*) und $\|\cdot\|$ die $L_2$-Norm. Dann ist die $h$-*modal Datentiefe (HMD) *von $\tilde{y}$ definiert als*

$$d^{\mathrm{HM}}(\tilde{y}; \mathbf{y}, h) = \frac{1}{n} \sum_{i=1}^{n} K_h(\|\tilde{y} - y_i\|) = \frac{1}{n} \sum_{i=1}^{n} K_h\left(\sqrt{\int (\tilde{y}(t) - y_i(t))^2 \, dt}\right). \quad (5.2)$$

Folglich misst $d^{\mathrm{HM}}$, wie dicht $\tilde{y}$ von der Stichprobe $\mathbf{y}$ umgeben ist. Je „dichter" $\tilde{y}$ von der Stichprobe umgeben ist, desto mehr spricht dafür, dass $\tilde{y}$ als funktionaler Modus der Stichprobe angesehen werden kann. Wegen des Gaußkerns erinnert sie an Kerndichteschätzer, deren Maximalstelle in multivariaten Datensituation als geschätzter Modus der Verteilung aufgefasst wird. Im R-Paket `fda.usc` (Febrero-Bande und Oviedo de la Fuente, 2012) existieren Implementierungen für (5.1) (`depth.FM`) und (5.2) (`depth.mode`). In `depth.mode` ist neben $h$ und der Norm auch ein Trimmungsparameter einstellbar: Standardmäßig wird ein getrimmtes Mittel bestimmt, bei dem die kleinsten $25\%$ der Summanden in (5.2) nicht verwendet werden.

Die unter den funktionalen Datentiefen anschaulichste Interpretation besitzt die MBD. Sie ist ähnlich motiviert wie die multivariate *convex hull peeling depth* von Barnett (1976): Alle Beobachtungen auf der konvexen Hülle der Stichprobe gehören zu Level 1. Dann werden diese Beobachtungen entfernt und eine konvexe Hülle der verbleibenden Stichprobe bestimmt, deren Elemente zu Level 2 gehören. Dieses Vorgehen wird wiederholt bis die verbleibenden Elemente keine konvexe Hülle mehr bilden können oder keine Elemente mehr vorhanden sind. In je mehr konvexen Hüllen eine Beobachtung enthalten ist, desto höher ist ihre Datentiefe.

DEFINITION 5.4. Modifizierte Bandtiefe *(Lopez-Pintado und Romo, 2009)*
*Sei $\tilde{y} \in \mathbb{B}$ eine beliebige funktionale Beobachtung und $\mathbf{y}$ eine gegebene Stichprobe. Des Weiteren sei $A(\tilde{y}; y_{i_1}, y_{i_2}) = \{t \in \mathcal{T} : \min_{r=i_1,i_2} y_r(t) \leq \tilde{y}(t) \leq \max_{r=i_1,i_2} y_r(t)\}$ das Intervall, in dem sich $\tilde{y}$ innerhalb des Gebiets befindet, das durch die Funktionen $y_{i_1}$ und $y_{i_2}$ beschränkt*

*wird. Sei außerdem $\lambda$ das Lebesguemaß auf $\mathcal{T}$ und $\lambda_r(A(y;\cdot)) = \lambda(A(y;\cdot))/\lambda(\mathcal{T})$. Dann ist die* modifizierte Bandtiefe (modified band depth, MBD) *von $\tilde{y}$ definiert durch*

$$d^{\mathrm{MB}}(\tilde{y};\mathbf{y}) = \binom{n}{2}^{-1} \sum_{1\leq i_1<i_2\leq n} \lambda_r(A(\tilde{y}; y_{i_1}, y_{i_2})).$$

Mit anderen Worten misst $\lambda_r$ den „Zeitanteil", in dem sich $\tilde{y}$ innerhalb des Bands $A(\tilde{y};\cdot)$ befindet. Das arithmetische Mittel der anteiligen Verweildauer von $\tilde{y}$ in allen von zwei Funktionen aufgespannten Bändern wird mit $d^{\mathrm{MB}}$ bezeichnet. Die MBD kann auch für Bänder definiert werden, die von mehr als zwei Funktionen aufgespannt werden. Wie Lopez-Pintado und Romo anmerken, steigt der dafür benötigte Rechenaufwand sehr stark, wohingegen die Ergebnisse nur leicht variieren, weshalb hier darauf verzichtet wird. Diese Datentiefe ist implementiert im R-Paket `fda` (Ramsay *et al.*, 2014).

BEMERKUNG 5.5. *In der Regel soll in der Praxis für jedes Stichprobenelement der Wert einer zuvor festgelegten Datentiefe bestimmt werden. Dafür muss jedes Stichprobenelement temporär aus der Stichprobe entfernt werden und wird dann als $\tilde{y}$ bezeichnet.*

Nagy *et al.* (2016) stellen fest, dass MBD und FMD zur Klasse der integrierten funktionalen Datentiefen gehören, für die sich eine Reihe asymptotischer Eigenschaften wie Konsistenz und Messbarkeit gemeinsam herleiten lässt. In Bezug auf *finite-sample* Eigenschaften trifft dies jedoch nicht zu.

Tabelle 5.1 liefert eine Übersicht über die Verwandtschaftsbeziehungen zwischen den vorgestellten Datentiefen und uni- sowie multivariaten Maßen.

Tabelle 5.1: Multivariate Datentiefen und die davon inspirierten funktionalen Ansätze zur Identifikation zentraler Parameter einer zugrunde liegenden Verteilung

| | Verfahren zur Bestimmung des Parameters | | |
|---|---|---|---|
| Parameter | *Univariat* | *Multivariat* | *Funktional* |
| Median | emp. Median | Halbraumtiefe | FMD |
| | | Convex hull peeling depth | MBD |
| Modus | Kerndichte | Simplextiefe | HMD |

Weil diese Datentiefen nicht explizit konstruiert wurden, um Form-Ausreißer zu er-

kennen, wird speziell dafür in Abschnitt 5.1.2 ein neues Maß namens *functional tangential angle pseudo-depth* (FUNTA) eingeführt. FUNTA basiert auf den Winkeln, die die Tangenten in den Schnittpunkten einer Beobachtung mit den anderen Tangenten der jeweils geschnittenen Funktionen bildet. Einen ähnlichen Ansatz verfolgen Mosler und Polyakova (2012) mit der *location-slope graph depth*, die auf den Funktionen und ihren Ableitungen basiert. Teile der Abschnitte 5.1.2 und 5.1.3 sind bereits in Kuhnt und Rehage (2016) veröffentlicht.

### *5.1.2 FUNTA Pseudo-Datentiefe*

Die nun folgende Datentiefe wird, anders als die bisher vorgestellten, nicht für die Bestimmung eines (funktionalen) Medians oder Modus konstruiert, sondern speziell zur Identifikation von Form-Ausreißern. Eine wesentliche Annahme ist, dass die Lage einer Kurve nichts darüber aussagt, ob sie ein Form-Ausreißer ist. Daher werden in Definition 5.7 zentrierte Funktionen $y \in \mathbb{B}$ betrachtet, das heißt, es gilt $\int y(t)dt = 0$. Im Hilbertraum-Modell funktionaler Daten ist dies eine gängige Annahme (Horváth und Kokoszka, 2012, S. 21). Ziel dieses Abschnitts ist es, eine Graphen-basierte Datentiefe zu entwickeln, die leicht zu bestimmen ist. Dafür wird als Indikator der Form-Unähnlichkeit einer Funktion mit einer oder mehreren anderen Funktionen ihr durchschnittlicher Schnittwinkel herangezogen.

DEFINITION 5.6. Schnittpunkte und Schnittwinkel eines Funktionenpaares
*Seien zwei zentrierte, differenzierbare Funktionen $y_1, y_2 \in \mathbb{B}$ gegeben. Dann existiert eine Familie $\mathcal{S}^P = (s_k)_{k\in I}$ mit $s_k \in \mathcal{T}$ für alle $k \in I$, für die $y_1(s_k) = y_2(s_k)\ \forall\ k \in I$ gilt. Die Familie $\mathcal{S}^P$ wird* Familie der Schnittpunkte *genannt. Weiterhin bezeichne $y_i'(t)$ die erste Ableitung von $y_i$ an der Stelle $t, i = 1, 2$. Dann ist der* tangentiale Schnittwinkel *der Funktion $y_1$ mit $y_2$ im Punkt $s_k$ durch*

$$w_k := w_k(y_1(s_k), y_2(s_k)) = \arccos\left(\frac{1 + y_1'(s_k)y_2'(s_k)}{\sqrt{(1 + y_1'(s_k)^2)(1 + y_2'(s_k)^2)}}\right)$$

*gegeben, wobei $s_k \in \mathcal{S}^P$ gilt. Der Winkel $w_k \in [0, \pi[$ wird im Bogenmaß angegeben. Dann bezeichne $\mathcal{S}^W = (w_k)_{k\in I}$ die* Familie der tangentialen Schnittwinkel. *Bei endlichen In-*

*dexmengen $|I| < \infty$ werden $\mathcal{S}^P$ und $\mathcal{S}^W$* Schnittpunktmenge *bzw.* Schnittwinkelmenge *genannt.*

Das Konzept des tangentialen Schnittwinkels wird in Abbildung 5.3 visualisiert: Links sind zwei Funktionen (schwarz und grau) abgebildet, welche sich in einem Punkt schneiden. In der Mitte sind die Tangenten der Funktionen im Schnittpunkt gestrichelt eingezeichnet. Der Schnittwinkel $w$ ist im rechten Plot eingetragen und bezieht sich stets auf den inneren Winkel des Dreiecks, das von den Schnittpunkten der Tangenten und der Ordinate gebildet wird.

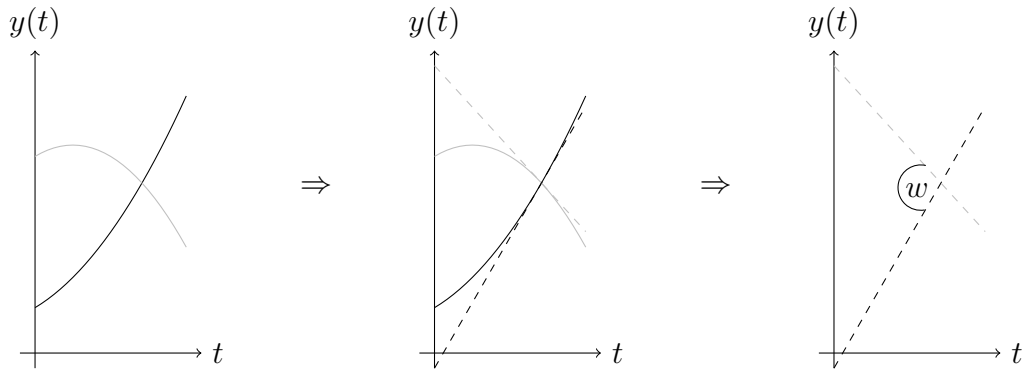



Abbildung 5.3: Konzept des tangentialen Schnittwinkels

In dieser Arbeit wird von endlichen Indexmengen $I$ ausgegangen. Falls Intervalle existieren, so dass $y_1(s) = y_2(s)$ für alle $s \in [t_\ell, t_u]$ mit $t_\ell \neq t_u$, werden nur die beiden Schnittpunkte $t_\ell$ und $t_u$ in $\mathcal{S}^P$ aufgenommen. Dies verhindert, dass unendlich viele Schnittwinkel gleich Null sind. Dies gilt insbesondere für $[t_\ell, t_u] = \mathcal{T}$. Der Fall von unendlich vielen Schnittwinkeln ungleich Null ist eher theoretischer Natur (z. B. $y_1(t) = \cos(1/t)$ mit $y_2(t) = 0$ für $t \in ]0, 1]$) und wird daher vernachlässigt.

Je größer der durchschnittliche Schnittwinkel einer Funktion $\tilde{y}$ mit allen anderen Funktionen $\mathbf{y}$ ist, desto unähnlicher soll $\tilde{y}$ gegenüber $\mathbf{y}$ angesehen werden. Andererseits haben zwei Funktionen, die sich nur um eine Konstante unterscheiden, dieselbe Form, was durch die Datentiefe reflektiert werden sollte. Dass jedes Funktionenpaar mindestens einen Schnittwinkel besitzt, wird durch die Zentrierung sichergestellt (siehe Proposition 5.9). Ein kleiner durchschnittlicher Schnittwinkel spricht dafür, dass die Form einer Funktion gut zur Form der übrigen Funktionen der Stich-

probe passt. Im Kontext der Datentiefe wird dies durch einen hohen Wert gekennzeichnet. Somit sei eine Pseudo-Datentiefe wie folgt konstruiert:

DEFINITION 5.7. FUNTA Pseudo-Datentiefe (Stichprobenversion)
*Sei $\tilde{y} \in \mathbb{B}$ eine beliebige zentrierte differenzierbare funktionale Beobachtung und* $\mathbf{y}$ *eine gegebene Stichprobe zentrierter und differenzierbarer Beobachtungen. Seien $\mathcal{S}^P(\tilde{y}, y_i)$ bzw. $\mathcal{S}^W(\tilde{y}, y_i)$ die Schnittpunktmenge bzw. Schnittwinkelmenge von $\tilde{y}$ mit $y_i, i = 1, \ldots, n$. Bezeichne $s_{i1}, \ldots, s_{im_i}$ die Schnittpunkte und $w_{i1}, \ldots, w_{im_i}$ die Schnittwinkel von $\tilde{y}$ mit $y_i, i = 1, \ldots, n$. Mit $m = \sum_{i=1}^{n} m_i$ ist die* FUNTA Pseudo-Datentiefe (functional tangential angle pseudo-depth) *von $\tilde{y}$ bzgl. der Stichprobe* $\mathbf{y}$ *definiert durch*

$$d^{\text{FUNTA}}(\tilde{y}; \mathbf{y}) = 1 - \frac{1}{m} \sum_{i=1}^{n} \sum_{k=1}^{m_i} \frac{w_{ik}}{\pi}.$$

BEMERKUNG 5.8. *In der Regel ist es nicht sinnvoll, bei Vorliegen mehrerer Winkel den „typischen" Winkel mit Hilfe des arithmetischen Mittels zu bestimmen. Im Falle des in Definition 5.6 eingeführten Schnittwinkels gilt jedoch, dass $w \in [0, \pi[$ ist, wodurch diese Problematik hier nicht existiert.*

Eine Implementierung in R ist im Paket `FUNTA` (Rehage, 2016) in der gleichnamigen Funktion verfügbar. Es genügt bei der Analyse von FUNTA nur eine endliche Anzahl von Schnittpunkten zu betrachten, da unendlich viele Schnittpunkte in der Praxis nicht auftreten können.

PROPOSITION 5.9. *Es seien $y_1, y_2 \in \mathcal{C}([t_\ell, t_u])$ und zentriert. Dann gilt: Es existiert mindestens ein $\tilde{t} \in \mathcal{T} = [t_\ell, t_u]$, so dass $y_1(\tilde{t}) = y_2(\tilde{t})$.*

*Beweis:* Sei für alle $t \in \mathcal{T} : f(t) := y_1(t) - y_2(t)$. Falls $\exists\, \tilde{t} \in \mathcal{T}$ mit $f(\tilde{t}) = 0$, stimmt die Behauptung. Da $f$ aus stetigen Funktionen besteht, ist $f$ stetig. Sei $a \in \mathcal{T}$. Betrachte die folgende Fallunterscheidung:

1. Sei $a$ so, dass $f(a) = 0$. Dann existiert ein Schnittpunkt in $a$.
2. Sei $a$ so, dass $f(a) < 0 \Rightarrow y_1(a) < y_2(a)$. Wegen des Zwischenwertsatzes muss nur $\exists\, b \in \mathcal{T} : y_1(b) > y_2(b)$ gezeigt werden. Angenommen, $\nexists\, b \in \mathcal{T} : y_1(b) > y_2(b)$. Dann folgt: $0 = \int_{\mathcal{T}} y_1(t)dt < \int_{\mathcal{T}} y_2(t)dt = 0$. Dies ist ein Widerspruch.

3. Sei $a$ so, dass $f(a) > 0$. Der Beweis ist analog zum zweiten Fall.

□

Mit anderen Worten hat also jede zentrierte stetige Funktion $y_i, i = 1, \ldots, n$ mindestens je einen Schnittpunkt mit jeder anderen zentrierten stetigen Funktion. Proposition 5.9 stellt die Existenz der FUNTA in beliebigen zentrierten Stichproben sicher.

Um die Interpretation einer Datentiefe zu erleichtern, ist die Beschränktheit eine wichtige Eigenschaft (Zuo und Serfling, 2000):

BEMERKUNG 5.10. *Es seien zentrierte Beobachtungen $\tilde{y}, y_i \in \mathbb{B}, i = 1, \ldots, n$ gegeben. Dann gilt: $d^{\mathrm{FUNTA}}(\tilde{y}; \mathbf{y}) \in ]0, 1]$, die empirische FUNTA Pseudo-Datentiefe ist also normiert. Dies folgt durch Einsetzen der extremen Winkel $w \in [0, \pi[$ und der Monotonie von $d^{\mathrm{FUNTA}}$ bezüglich $w$.*

BEMERKUNG 5.11. *Im Gegensatz zu anderen Datentiefen beruht FUNTA nicht auf Konturen, also Bereichen, in denen eine Funktion liegt, und deshalb mindestens eine vorgegebene Datentiefe erreicht. Des Weiteren existiert FUNTA in sinnvoller Weise für (zentrierte) Funktionenpaare $y_1, y_2 \in \mathbb{B}$ und es gilt: $d^{\mathrm{FUNTA}}(y_1; y_2) = d^{\mathrm{FUNTA}}(y_2; y_1)$. So kann FUNTA auch als Distanzmaß aufgefasst werden. Von den hier vorgestellten Datentiefen ist das nur für die $h$-modal Datentiefe – bei passender Wahl der Tuningparameter – ebenfalls der Fall. Die modifizierte Bandtiefe ist bei nur zwei Beobachtungen nicht definiert, da bezüglich einer Funktion $y_1$ immer ein Band aus zwei weiteren Funktionen erstellt werden muss. Die integrierte Datentiefe ist zwar definiert, aber kein Distanzmaß, da sie nicht symmetrisch ist: $d^{\mathrm{FM}}(y_1; y_2) \neq d^{\mathrm{FM}}(y_2; y_1)$.*

BEMERKUNG 5.12. *Zur Bestimmung der Schnittwinkel in* R *ist eine vorherige Anpassung des Datensatzes mit Hilfe eines gewählten Basissystems nötig. Die Funktionen werden an einer großen Anzahl von Elementen der Indexmenge $\mathcal{T}$ ausgewertet und alle Schnittpunkte bestimmt. Damit können die Tangenten und Schnittwinkel schnell berechnet werden. Die Ergebnisse dieser numerischen Ableitungen unterscheiden sich kaum von denen, die mit Hilfe analytisch bestimmter Ableitungen erreicht werden, sind aber rechnerisch effizienter.*

Durch die Betrachtung einfacher Beispiele (Abb. 5.4) wird klar, warum zum Beispiel

die modifizierte Bandtiefe bei gewissen Ausreißerszenarien zur Ausreißeridentifikation ungeeignet ist. Die linke Grafik zeigt nur sich in der Lage unterscheidende Funktionen, wobei die obere gestrichelte Kurve eher nicht als Ausreißer zu sehen ist, die untere gestrichelte Kurve jedoch schon. MBD ordnet beiden gestrichelten Kurven die gleiche Tiefe zu, da zwischen ihnen und der „zentralen“ Kurve gleich viele andere Kurven liegen. Die MBD unterscheidet, genau wie die Rangstatistik im univariaten Fall, nicht zwischen großen Abständen und kleinen Abständen. Arribas-Gil und Romo (2014) beschreiben, wie diese Erkennung von Form-Ausreißern durch Einbeziehung des *modified epigraph index* verbessert werden kann. Anhand der FUNTA Pseudo-Datentiefe der einzelnen Kurven – nach der Zentrierung – sind keine Unterschiede zu erkennen, da ihre Formen identisch sind. Das ideale Ergebnis liefert in diesem Fall nur die $h$-modal Datentiefe, die den Lage-Ausreißer identifiziert.

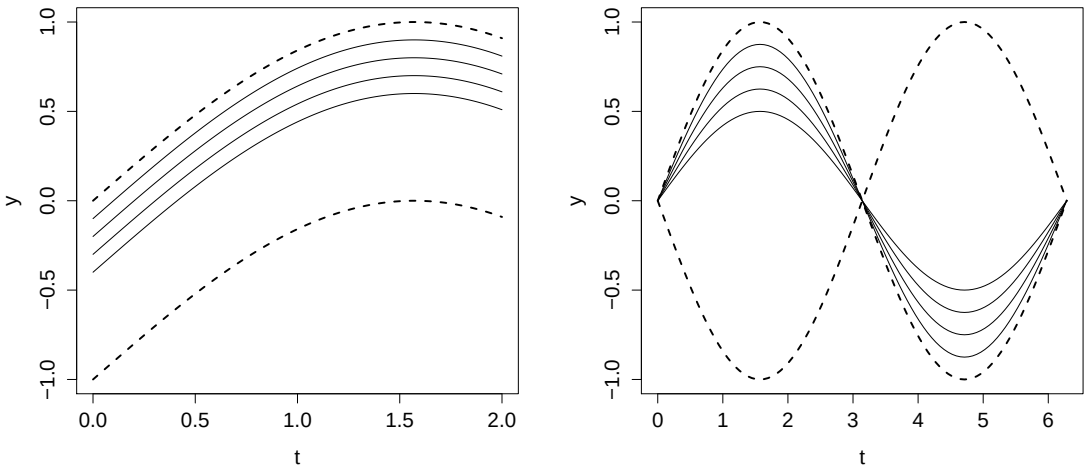



Abbildung 5.4: Zwei Beispiele funktionaler Daten – die gestrichelten Kurven haben jeweils die gleiche MBD

Auf der rechten Seite von Abbildung 5.4 ordnet die MBD den gestrichelten Kurven wieder die gleiche Tiefe zu. HMD und FUNTA finden hingegen den Form-Ausreißer. Ein Vorteil von FUNTA gegenüber HMD ist die Nichtexistenz eines Tu-

ningparameters. Neben der Darstellung in Definition 5.3 kann alternativ durch

$$d^{\mathrm{HM}}_{\mathrm{skal}}(y;\mathbf{y},h) = \frac{d^{\mathrm{HM}}(y;\mathbf{y},h) - \min_{i=1,\dots,n}(d^{\mathrm{HM}}(y_i;\mathbf{y},h))}{\max_{i=1,\dots,n}(d^{\mathrm{HM}}(y_i;\mathbf{y},h)) - \min_{i=1,\dots,n}(d^{\mathrm{HM}}(y_i;\mathbf{y},h))}$$

eine skalierte HMD angegeben werden. Durch diese Darstellung ist es jedoch nicht möglich, anhand der Datentiefe Aussagen über die Heterogenität der Stichprobe zu tätigen. Für die Bestimmung von Grenzwerten für Ausreißer mit dem geglätteten Bootstrap ist die obige Normierung deshalb ungeeignet. Weiterhin ist die Voreinstellung der R-Funktion `depth.mode` (im Paket `fda.usc`) für den Tuningparameter $h$ das $15\%$-Quantil der paarweisen $L_2$-Distanzen der Beobachtungen (Febrero-Bande und Oviedo de la Fuente, 2012). Da die Distanzen auf der Hauptdiagonale der Distanzmatrix mit einbezogen werden, ergibt diese Voreinstellung $h = 0$ für Stichprobenumfänge $n < 6$ und somit den Tiefewert 0 für alle Beobachtungen. Bei einer alternativen Wahl für $h$ muss beachtet werden, dass sowohl zu kleine als auch zu große $h$ zu identischen Tiefewerten führen.

In diesem Kapitel wurden bislang ausschließlich funktionale Datentiefen von Realisationen zeitstetiger stochastischer Prozesse betrachtet. Nachfolgend wird die FUNTA Pseudo-Datentiefe für stochastische Prozesse selbst formuliert. Um das Konzept der Differenzierbarkeit, das in der Stichprobenversion der FUNTA Pseudo-Datentiefe genutzt wird, auch für stochastische Prozesse nutzen zu können, muss der Begriff des *Grenzwerts* im Differenzenquotienten ersetzt werden. Stattdessen kann die *Konvergenz im quadratischen Mittel* verwendet werden.

DEFINITION 5.13. Differenzierbarkeit im quadratischen Mittel (Cramér und Leadbetter, 1967, S. 84)
*Sei $\mathcal{Y} = \{\mathcal{Y}(t) \mid t \in \mathcal{T}\}$ ein schwach stationärer stochastischer Prozess. $\mathcal{Y}$ ist in $t \in \mathcal{T}$* differenzierbar im quadratischen Mittel *mit Ableitung $\mathcal{Y}'(t)$, falls $\frac{\mathcal{Y}(t+h)-\mathcal{Y}(t)}{h}$ im quadratischen Mittel gegen $\mathcal{Y}'(t)$ konvergiert, d. h., falls*

$$\lim_{h\to 0} \mathsf{E}\left(\left[\frac{\mathcal{Y}(t+h)-\mathcal{Y}(t)}{h} - \mathcal{Y}'(t)\right]^2\right) = 0.$$

KOROLLAR 5.14. *Sei $\sigma = \{\sigma(s,t) \mid s,t \in \mathcal{T}\}$ die Kovarianzfunktion eines schwach*

*stationären stochastischen Prozesses $\mathcal{Y} = \{\mathcal{Y}(t) \mid t \in \mathcal{T}\}$. $\mathcal{Y}$ ist in $t \in \mathcal{T}$ differenzierbar im quadratischen Mittel, falls $\frac{\partial^2 \sigma}{\partial s \partial t}$ existiert und endlich im Punkt $(t, t)$ ist.*

*Beweis:* Vergleiche Cramér und Leadbetter (1967, S. 84). □

Falls gilt, dass $\frac{\partial^2 \sigma}{\partial s \partial t}$ existiert und endlich in allen Punkten $(t, t) \in \mathcal{T} \times \mathcal{T}$ ist, wird $\mathcal{Y}$ kurz *differenzierbar im quadratischen Mittel* genannt.

Betrachtet sei mit $\mathcal{Y} = \{\mathcal{Y}(t) : t \in \mathcal{T}\}$ ein zentrierter, zeitstetiger stochastischer Prozess, der differenzierbar im quadratischen Mittel ist. Die unabhängig und identisch verteilten Kopien dieses Prozesses seien mit $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ bezeichnet. Weiterhin sei $\mathcal{W}_{ik}$ die Zufallsvariable, die den Schnittwinkel von $\mathcal{Y}$ mit $\mathcal{Y}_i$ im Schnittpunkt $S_k$ bezeichnet:

$$\mathcal{W}_{ik} = \arccos\left(\frac{1 + \mathcal{Y}'(S_k)\mathcal{Y}_i'(S_k)}{\sqrt{(1 + \mathcal{Y}'(S_k)^2)(1 + \mathcal{Y}_i'(S_k)^2)}}\right)$$

wobei $k = 1, \ldots, M_i, \sum_{i=1}^{n} M_i = M$. Folglich lässt sich FUNTA in einem Populationsszenario wie folgt definieren:

$$D^{\text{FUNTA}}(\mathcal{Y}; \mathcal{Y}_1, \ldots, \mathcal{Y}_n) = 1 - \frac{1}{M} \sum_{i=1}^{n} \sum_{k=1}^{M_i} \frac{\mathcal{W}_{ik}}{\pi}. \tag{5.3}$$

Da $S_k$ eine Zufallsvariable mit unbekannter, aber von $\mathcal{Y}$ abhängiger Verteilung ist, kann der Erwartungswert von $\mathcal{W}_{ik}$ nur unter zusätzlichen Annahmen hergeleitet werden. Da $\mathcal{Y}_1, \ldots, \mathcal{Y}_n$ unabhängig identisch verteilt sind, lässt sich $\mathcal{W}_{ik}$ vereinfachen:

$$\mathcal{W}_{ik} = \mathcal{W}_k = \arccos\left(\frac{1 + \mathcal{Y}'(S_k)\mathcal{Y}_1'(S_k)}{\sqrt{(1 + \mathcal{Y}'(S_k)^2)(1 + \mathcal{Y}_1'(S_k)^2)}}\right).$$

Somit folgt aus (5.3)

$$D^{\text{FUNTA}}(\mathcal{Y}; \mathcal{Y}_1, \ldots, \mathcal{Y}_n) = 1 - \frac{1}{M} \sum_{k=1}^{M} \frac{\mathcal{W}_k}{\pi}. \tag{5.4}$$

Unter der zusätzlichen Voraussetzung, dass die Anzahl der Schnittpunkte $M$ unab-

hängig von der Verteilung der Schnittwinkel ist, lässt sich der Erwartungswert von (5.4) mittels des Gesetzes des iterierten Erwartungswerts umschreiben:

$$\begin{aligned}\mathsf{E}\left(D^{\text{FUNTA}}\left(\mathcal{Y};\mathcal{Y}_1,\ldots,\mathcal{Y}_n\right)\right) &= 1-\frac{1}{\pi}\mathsf{E}\left(M^{-1}\sum_{k=1}^{M}\mathcal{W}_k\right)\\ &= 1-\frac{1}{\pi}\mathsf{E}\left(\mathsf{E}\left(M^{-1}\sum_{k=1}^{M}\mathcal{W}_k\middle|M\right)\right).\end{aligned}$$

Mit Hilfe der Formel von Wald und der Annahme, dass $\mathcal{W}_1,\ldots,\mathcal{W}_M$ unabhängig identisch verteilt sind, folgt

$$\begin{aligned}1-\frac{1}{\pi}\mathsf{E}\left(\mathsf{E}\left(M^{-1}\sum_{k=1}^{M}\mathcal{W}_k\middle|M\right)\right) &= 1-\frac{1}{\pi}\mathsf{E}\left(M^{-1}\mathsf{E}\left(\sum_{k=1}^{M}\mathcal{W}_k\middle|M\right)\right)\\ &= 1-\frac{1}{\pi}\mathsf{E}\left(M\mathsf{E}\left(\mathcal{W}\right)/M\right)\\ &= 1-\frac{1}{\pi}\mathsf{E}\left(\arccos\left(\frac{1+\mathcal{Y}'\mathcal{Y}_1'}{\sqrt{(1+\mathcal{Y}'^2)(1+\mathcal{Y}_1'^2)}}\right)\right).\end{aligned}$$

Um den Erwartungswert analytisch herleiten zu können, wird im Folgenden die erwartete FUNTA Pseudo-Datentiefe einer konkreten Funktion $\tilde{y}$ in Bezug auf einen stationären stochastischen Prozess $\mathcal{Y}$ hergeleitet. Falls die konkrete Funktion $\tilde{y}$ als Realisation eines im quadratisch Mittel differenzierbaren stochastischen Prozesses aufgefasst wird, ist zu beachten, dass sich die Differenzierbarkeit nicht automatisch auf $\tilde{y}$ vererbt (Cramér und Leadbetter, 1967, S. 84). Hierzu ist es jedoch hinreichend zu fordern, dass für den stationären stochastischen Prozess $\mathcal{Y}$ mit Kovarianzfunktion $\sigma$ die vierte Ableitung $\frac{\partial^4\sigma}{\partial s^4}\Big|_{s=0}$ existiert, siehe Cramér und Leadbetter (1967, S. 125). Somit lässt sich analog zu Definition 5.7 die erwartete FUNTA Pseudo-Datentiefe definieren:

DEFINITION 5.15. Erwartete FUNTA Pseudo-Datentiefe

*Sei $\mathcal{Y}=\{\mathcal{Y}(t):t\in\mathcal{T}\}$ ein zentrierter, zeitstetiger stochastischer Prozess, der im quadratischen Mittel differenzierbar ist und dessen Realisationen differenzierbar sind. Sei $\mathcal{W}(\tilde{y},\mathcal{Y};t)$ die Zufallsvariable, die die Verteilung des Verhältnisses zwischen den Steigungen von $\tilde{y}$ und*

*$\mathcal{Y}$ an einem beliebigen Punkt $t$ beschreibt:*

$$\mathcal{W}(\tilde{y}, \mathcal{Y}; t) = \arccos\left(\frac{1 + \tilde{y}'(t)\mathcal{Y}'(t)}{\sqrt{(1 + \tilde{y}'(t)^2)(1 + \mathcal{Y}'(t)^2)}}\right). \tag{5.5}$$

*Unter der Annahme, dass die Verteilungen des Steigungsverhältnisses in Schnittpunkten und Nicht-Schnittpunkten identisch sind, ist die* erwartete FUNTA Pseudo-Datentiefe *der Beobachtung $\tilde{y}$ gegeben durch*

$$D^{\mathrm{FUNTA}}(\tilde{y}; P_{\mathcal{Y}}) = 1 - \frac{\mathsf{E}(\mathcal{W}(\tilde{y}, \mathcal{Y}; t))}{\pi} =: 1 - \frac{\mathsf{E}(\mathcal{W})}{\pi}.$$

Inwieweit die obige Annahme als gerechtfertigt angesehen werden kann, hängt von der Verteilung von $\mathcal{Y}'(t)$ ab.

Die weitere Untersuchung der erwarteten FUNTA Pseudo-Datentiefe soll aus dem Blickwinkel der Gaußprozesse (siehe Definition 3.11) getätigt werden. Es ist von Interesse, inwieweit $D^{\mathrm{FUNTA}}(\tilde{y}; P_{\mathcal{Y}})$ bei gegebenem $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$ mit bekannten $\mu, \sigma$ bestimmt werden kann. Zunächst sei die Beschränkung auf zentrierte Funktionen und stochastische Prozesse nochmals betont, damit $\tilde{y}$ und $\mu$ mindestens einen Schnittpunkt haben. Für die Herleitung der Verteilung von $\mathcal{W}(\tilde{y}, \mathcal{Y}; t)$ muss zunächst die von $\mathcal{Y}'(t)$ bestimmt werden:

PROPOSITION 5.16. *Sei $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$ mit $\mu(t) \equiv \mu$ und $\sigma(s,t) = f(|s-t|)$, wobei $f : \mathcal{T}_0 \to \mathbb{R}_+$ stetig, positiv definit und zweimal stetig differenzierbar für alle $t_0 \in \mathcal{T}_0$ ist. Weiterhin bezeichne $f^{(2)} := \{f^{(2)}(|s-t|) \mid s, t \in \mathcal{T}\}$. Dann ist die Ableitung $\mathcal{Y}' \sim \mathcal{GP}(0, f^{(2)})$, und für alle $t \in \mathcal{T}$ gilt $\mathcal{Y}'(t) \sim \aleph(0, f^{(2)}(0))$.*

*Beweis:* Wegen Adler (1981, Thm. 2.2.2), folgt für die Parameter von $\mathcal{Y}'$:

$$\begin{aligned}
\mu'(t) &= 0, \\
\frac{\partial \sigma}{\partial t} &= \left.\frac{\partial f(x)}{\partial x}\right|_{|s-t|} \frac{s-t}{|s-t|}, \\
\frac{\partial^2 \sigma}{\partial s \partial t} &= \left.\frac{\partial f(x)}{\partial x}\right|_{|s-t|} \frac{-|s-t| + (s-t)\frac{s-t}{|s-t|}}{|s-t|^2} + \frac{(s-t)^2}{|s-t|^2} \left.\frac{\partial^2 f(x)}{\partial x^2}\right|_{|s-t|}
\end{aligned}$$

$$= \left.\frac{\partial f(x)}{\partial x}\right|_{|s-t|} \left[-\frac{|s-t|}{|s-t|^2} + \frac{(s-t)^2}{|s-t|^3}\right] + \frac{(s-t)^2}{|s-t|^2} \left.\frac{\partial^2 f(x)}{\partial x^2}\right|_{|s-t|}$$
$$= f^{(2)}(|s-t|).$$

Für $s = t$ erhält man $\mathcal{Y}'(t) \sim \aleph(0, f^{(2)}(0))$. □

Ein gängiger Spezialfall eines Gaußprozesses hat als Parameter die Erwartungsfunktion $\mu(t) \equiv 0$ und die quadriert-exponentielle Kovarianzfunktion $\sigma(s,t) = \sigma_{QE}(s,t) = \exp\left(-\frac{(s-t)^2}{2\ell^2}\right)$ mit *characteristic length-scale* $\ell$. Dann gilt, dass $\mathcal{Y}'(t) \sim \aleph(0, f^{(2)}(0))$, wobei $f^{(2)}$ die zweite partielle Ableitung der Kovarianzfunktion ist:

$$\left.\frac{\partial \sigma}{\partial s}\right|_{s,t} = -\frac{s-t}{\ell^2} \exp\left(-\frac{(s-t)^2}{2\ell^2}\right)$$
$$\left.\frac{\partial^2 \sigma}{\partial s \partial t}\right|_{s,t} = \left[\frac{1}{\ell^2} - \frac{(s-t)^2}{\ell^4}\right] \exp\left(-\frac{(s-t)^2}{2\ell^2}\right) =: f^{(2)}(s-t).$$

Somit folgt $f^{(2)}(0) = \ell^{-2}$ und $\mathcal{Y}'(t) \sim \aleph(0, \ell^{-2})$. Die Verteilung von $\mathcal{Y}'(t) = \mathcal{Y}'$ hängt nicht von $t$ ab, was die Bestimmung der Verteilung von $\mathcal{W}$ wesentlich erleichtert. Zudem soll angenommen werden, dass $\tilde{y}'(t) \equiv 0$. In diesem Fall wird die FUNTA Pseudo-Datentiefe für $\tilde{y}'(t) = \mu'(t) \; \forall \, t$ hergeleitet. Aus (5.5) folgt

$$\mathcal{W} = \mathcal{W}(\tilde{y}, \mathcal{Y}) = \arccos\left(\frac{1}{\sqrt{1+\mathcal{Y}'^2}}\right).$$

Wegen $\ell\mathcal{Y}' \sim \aleph(0,1)$ folgt $\mathcal{X} := \ell^2\mathcal{Y}'^2 \sim \chi^2_1$. Somit kann die Dichte von $\mathcal{W}$ über die Verteilungsfunktion hergeleitet werden. Für $w \in ]0, \frac{\pi}{2}[$ gilt:

$$P(\mathcal{W} \leq w) = P\left(\arccos\left(\frac{1}{\sqrt{1+\mathcal{Y}'^2}}\right) \leq w\right)$$
$$= P\left(\arccos\left(\frac{1}{\sqrt{1+\frac{\mathcal{X}}{\ell^2}}}\right) \leq w\right) = P\left(\frac{1}{\sqrt{1+\frac{\mathcal{X}}{\ell^2}}} \geq \cos(w)\right),$$

da $\cos$ auf $]0, \frac{\pi}{2}[$ monoton fallend ist. Weiterhin gilt

$$P\left(\frac{1}{\sqrt{1+\frac{\mathcal{X}}{\ell^2}}} \geq \cos(w)\right) = P\left(1 + \frac{\mathcal{X}}{\ell^2} \leq \cos^{-2}(w)\right)$$

$$
\begin{aligned}
&= P\left(\mathcal{X} \leq \ell^2(\cos^{-2}(w) - 1)\right) \\
&= F_{\mathcal{X}}(\ell^2(\cos^{-2}(w) - 1)),
\end{aligned}
$$

wobei $F_{\mathcal{X}}$ die Verteilungsfunktion der $\chi^2_1$-Verteilung ist. Daraus folgt

$$
\begin{aligned}
F_{\mathcal{X}}(\ell^2(\cos^{-2}(w) - 1)) &= \gamma_u(0.5, 0.5\ell^2(\cos^{-2}(w) - 1))/\Gamma(0.5) \\
&= \frac{1}{\sqrt{\pi}} \int_0^{0.5\ell^2(\cos^{-2}(w)-1)} t^{-1/2} \exp(-t) dt
\end{aligned}
$$

mit der unvollständigen unteren Gammafunktion $\gamma_u(\cdot, \cdot)$ und der Gammafunktion $\Gamma(\cdot)$, wobei $\Gamma(0.5) = \sqrt{\pi}$. Durch die Anwendung der Substitutionsregel mit $y := \varphi(t) = \sqrt{t}$ und $f(y) = 2\exp(-y^2)$ erhält man

$$
\begin{aligned}
\frac{1}{\sqrt{\pi}} \int_0^{0.5\ell^2(\cos^{-2}(w)-1)} \frac{\exp(-t)}{\sqrt{t}} dt &= \frac{2}{\sqrt{\pi}} \int_0^{\sqrt{0.5\ell^2(\cos^{-2}(w)-1)}} \exp(-y^2) dy \\
&= \operatorname{erf}\left(\sqrt{0.5\ell^2(\cos^{-2}(w) - 1)}\right) = F_{\mathcal{W}}(w; \ell),
\end{aligned}
$$

wobei die Fehlerfunktion durch $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2) dt$ definiert ist. Mit $f(w; \ell) = \frac{\partial F(x;\ell)}{\partial x}\Big|_{x=w}$ folgt

$$
\begin{aligned}
f(w; \ell) &= \frac{\partial \operatorname{erf}(\sqrt{0.5\ell^2(\cos^{-2}(x) - 1)})}{\partial x}\bigg|_{x=w} \\
&= \frac{\partial \sqrt{0.5\ell^2(\cos^{-2}(w) - 1)}}{\partial w} \cdot \frac{\partial \operatorname{erf}(x)}{\partial x}\bigg|_{x=\sqrt{0.5\ell^2(\cos^{-2}(w)-1)}} \\
&= \frac{\ell\sqrt{2}\sin(w)\exp(-0.5\ell^2(\cos^{-2}(w) - 1))}{\sqrt{\pi(\cos^{-2}(w) - 1)}\cos^3(w)}
\end{aligned}
$$

für alle $w \in ]0, \frac{\pi}{2}[$. In Abbildung 5.5 sind die Dichten (schwarz) beispielhaft für $\ell = 0.1, 0.5, 1, 1.5, 2, 4$ veranschaulicht. Je kleiner $\ell$ ist, desto höher ist die Volatilität des Gaußprozesses. Ein Verhältnis der Steigungen, dass nahe $\pi/2$ ist, wird dadurch wahrscheinlicher als eines nahe $0$. Je größer $\ell$ ist, desto flacher werden die Realisationen des Gaußprozesses. Die Kerndichteschätzungen (rot) basieren auf dem Epanechnikov-Kern und der Bandbreitenschätzung $\hat{h}_{SJ}$ (Sheather und Jones, 1991) für 30 Wiederholungen von Stichproben vom Umfang 50. Es fällt auf, dass die ge-

schätzte Dichte vor allem dann von der analytischen abweicht, wenn $\ell$ klein ist, das heißt falls $\mathcal{Y}'$ eine große Varianz hat. Kleine Schnittwinkel treten dann eher selten auf. Ansonsten zeigt sich für größeres $\ell$, dass die beiden Dichten sehr ähnlich sind, obwohl sie auf unterschiedlichen Grundlagen (Schnittwinkel vs. alle Steigungsverhältnisse) basieren.

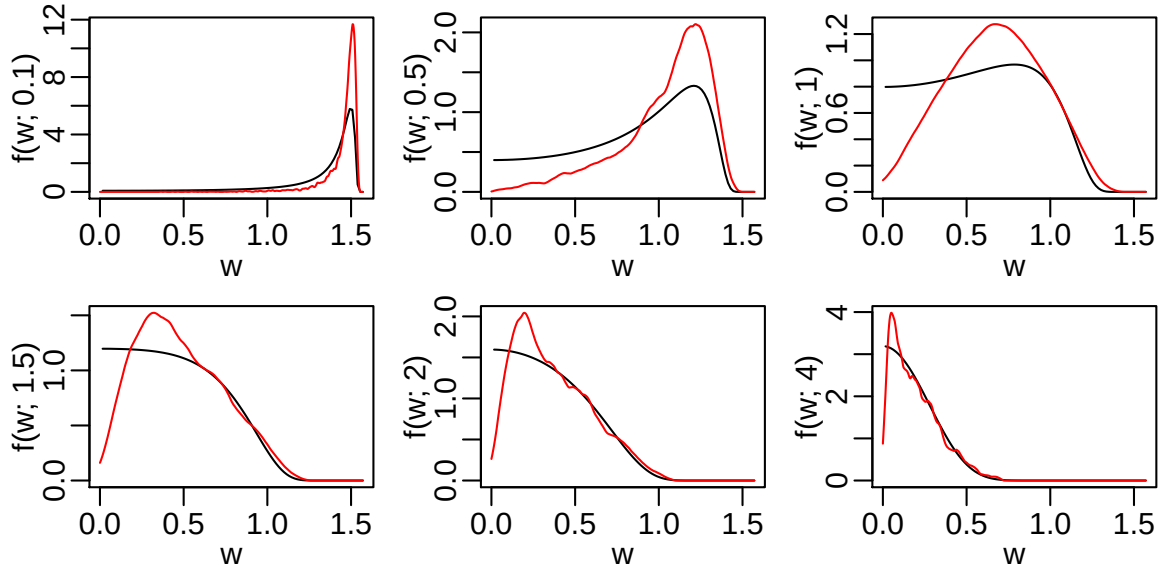



Abbildung 5.5: Dichten der Verteilung von $\mathcal{W}$ für $\ell = 0.1, 0.5, 1, 1.5, 2, 4$ (schwarz) und Kerndichteschätzungen basierend auf allen Steigungsverhältnissen (rot)

Aufgrund der Komplexität dieser Dichtefunktion wird der Erwartungswert nur numerisch für verschiedene $\ell$ bestimmt. Ergebnisse sind in Abbildung 5.6 dargestellt. Außerdem ist eine Approximation der Form $\pi/(2(1 + (\ell/a)^b))$ abgebildet, wobei $a$ und $b$ mittels `nls` in R durch $\hat{a} = 0.5735$ und $\hat{b} = 0.9891$ geschätzt wurden. Da $\mathsf{E}(\mathcal{W})$ fallend in $\ell$ ist und somit $D^{\text{FUNTA}}(y; P_{\mathcal{Y}})$ steigend in $\ell$, ist die Funktion $y(t) \equiv 0$ umso repräsentativer in Bezug auf die Form, je größer $\ell$ ist.

**Eigenschaften von FUNTA.** In den folgenden Sätzen wird untersucht, ob FUNTA typische Eigenschaften (Zuo und Serfling, 2000; Nieto-Reyes und Battey, 2016) einer statistischen Datentiefe erfüllt. Zuo und Serfling (2000) verlangen in *P1* die affine Invarianz, wozu auch die Lokationsinvarianz zählt. Satz 5.17 besagt, dass FUNTA lokationsinvariant ist.

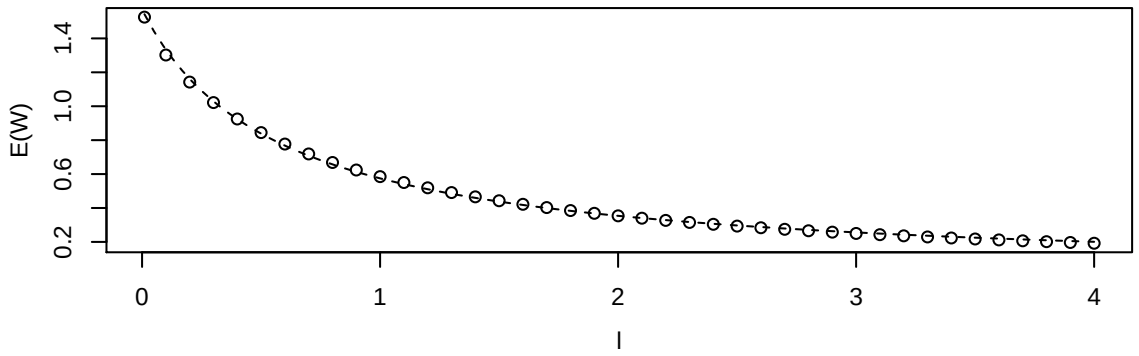



Abbildung 5.6: Numerisch bestimmte Erwartungswerte (Punkte) und Approximation (gestrichelt)

SATZ 5.17. *Sei $\tilde{y}$ eine beliebige Funktion und $\mathbf{y} = (y_1, \ldots, y_n)^\top$ die Realisation einer Stichprobe. Dann gilt: $d^{\mathrm{FUNTA}}(\tilde{y}; \mathbf{y}) = d^{\mathrm{FUNTA}}(\tilde{y} + c; \mathbf{y} + c)$ für $c \in \mathbb{R}$.*

*Beweis:* Es genügt zu zeigen, dass die Schnittwinkel von $\tilde{y}$ mit $\mathbf{y}$ und $\tilde{y} + c$ mit $\mathbf{y} + c$ gleich sind. Dies folgt aus $(y(t) + c)' = y'(t) \; \forall \, y \in \{\tilde{y}, y_1, \ldots, y_n\} \; \forall \, t \in \mathcal{T}$. □

Satz 5.17 gilt nicht für Verschiebungen der Form $l \in \mathbb{B}$. Zwar bleiben die Schnitt*punkte* gleich und FUNTA existiert weiterhin, die Schnittwinkel ändern sich jedoch. Seien hierzu $y_1(t) = t^2 - 1, y_2(t) = 1 - t^2$ mit $\mathcal{T} = [-\sqrt{3}, \sqrt{3}]$ definiert. Damit sind $y_1, y_2$ um 0 zentriert mit zwei Schnittpunkten $s_1 = -1, s_2 = 1$. Für den ersten Schnittwinkel gilt dann: $w_1(y_1(-1), y_2(-1)) = \arccos\left(\frac{1-2\cdot 2}{\sqrt{5\cdot 5}}\right) = \arccos(-3/5) = 2.214$. Sei nun die Funktion $l(t) = t^3$ definiert, welche zu $y_1(t)$ und $y_2(t)$ addiert wird. So erhält man mit $w_1(y_1(-1) + l(-1), y_2(-1) + l(-1)) = \arccos\left(\frac{1+1\cdot 5}{\sqrt{2\cdot 26}}\right) = \arccos(6/\sqrt{52}) = 0.588$ einen anderen Schnittwinkel.

PROPOSITION 5.18. *FUNTA ist nicht skalenäquivariant: Es existieren ein $s \in \mathbb{R}$ und Funktionen $y_1$ und $y_2$, so dass $d^{\mathrm{FUNTA}}(s \cdot y_1; s \cdot y_2) \neq s d^{\mathrm{FUNTA}}(y_1; y_2)$.*

*Beweis:* FUNTA beruht auf der nichtlinearen Arcus-Cosinus-Funktion und ist durch das Intervall $]0, 1]$ beschränkt. Eine Vergrößerung des Schnittwinkels durch Multiplikation einer positiven Konstante bewirkt eine *Verkleinerung* der FUNTA-Werte. □

PROPOSITION 5.19. *FUNTA ist nicht skaleninvariant: Es existieren ein $s \in \mathbb{R}$ und Funktionen $y_1$ und $y_2$, so dass $d^{\mathrm{FUNTA}}(s \cdot y_1; s \cdot y_2) \neq d^{\mathrm{FUNTA}}(y_1; y_2)$.*

*Beweis:* Seien $y_1'(1) = 1, y_2'(1) = 2$ mit einem Schnittpunkt in $t = 1$ gegeben. So ergibt sich der Schnittwinkel $w(y_1(1), y_2(1)) = 0.322$, aber für $s = 3$ ergibt sich $w(3y_1(1), 3y_2(1)) = 0.156$. □

Damit folgt, dass FUNTA Eigenschaft *P1* von Zuo und Serfling (2000) nicht erfüllt. Ihre Werte können jedoch wie die Werte einer statistischen Datentiefe interpretiert werden: Ein kleiner Wert suggeriert, dass die zugehörige Beobachtung (bezüglich ihrer Form) nicht zu den anderen Beobachtungen passt. Daher sei für solche Maße der Terminus „Pseudo-Datentiefe“ verwendet. Es sei darauf hingewiesen, dass die von Zuo und Serfling (2000) bzw. Nieto-Reyes und Battey (2016) postulierten Eigenschaften für Aussagen bezüglich der Lage von Beobachtungen intendiert sind.

### *5.1.3 Robustifizierte FUNTA Pseudo-Datentiefe*

Formausreißer mit einer höheren Anzahl von Schnittpunkten und größeren Winkeln als die Mehrheit der Daten können mit FUNTA (Def. 5.7) und dem Bootstrap-Algorithmus leicht identifiziert werden. Eine größere Anzahl von Schnittwinkeln *mit* Ausreißern beeinflusst jedoch auch die Tiefewerte der regulären Beobachtungen. Sie werden von ihrem „eigentlichen“ Tiefewert in einer Stichprobe ohne den potenziellen Ausreißer hin zu einem niedrigeren Wert verschoben. Ein solches Verhalten von FUNTA kann durch das Ersetzen des nicht-robusten, gepoolten arithmetischen Mittels der Schnittwinkel durch ein robusteres Maß verhindert werden.

DEFINITION 5.20. Robustifizierte FUNTA Pseudo-Datentiefe (Stichprobenversion)
*Mit der Notation von Definition 5.7 sei die* robustifizierte FUNTA Pseudo-Datentiefe *(rFUNTA) gegeben durch*

$$d^{\mathrm{rFUNTA}}(\tilde{y}; \mathbf{y}) = 1 - \pi^{-1} \underset{i=1,\ldots,n}{\mathrm{median}}(\max_k(w_k(\tilde{y}(s_{ik}), y_i(s_{ik})))). \tag{5.6}$$

Über die Verwendung des Medians dieser maximalen Schnittwinkel wird rFUNTA nur dann einen auffällig niedrigen Wert liefern, wenn $\tilde{y}$ zur *Mehrheit* der Funktionen einen besonders großen maximalen Schnittwinkel aufweist. Somit hat eine kleine

Anzahl extremer Beobachtungen keinen Einfluss mehr auf die Tiefe der regulären Beobachtungen.

Ohne die Verwendung des Maximum-Operators wäre rFUNTA zwar immer noch ein robustes Maß, aber nicht mehr sensitiv genug, um Ausreißer zu identifizieren, da der Anteil „unverdächtiger“, kleiner Schnittwinkel auch bei Ausreißern in der Regel zu hoch ist. Im R-Paket `FUNTA` existiert die Funktion `rFUNTA`, in der die Argumente `type.inner` und `type.outer` den inneren bzw. äußeren Operator in Formel (5.6) bezeichnen. Die Voreinstellung `type.inner = "max"` und `type.outer = "median"` führt zur Bestimmung von rFUNTA wie in Definition 5.20 beschrieben. Für beide Argumente kann eine Einstellung aus der Menge {`"max"`, `"median"`, `"mean"`} gewählt werden.

FUNTA und rFUNTA sind als Schätzer für die Repräsentativität der Form einer funktionalen Beobachtung bzgl. der Stichprobe konzipiert. Um zu analysieren, wie diese Schätzer im schlimmsten Fall reagieren, werden Finite-Sample-Bruchpunkte (Def. 3.8) bestimmt. Zuvor muss der supremale Schnittwinkel einer beliebigen Funktion mit anderen Funktionen bestimmt werden, für den folgende Punkte festgehalten werden können:

1. Der supremale Schnittwinkel zweier Funktionen beträgt 180°.

2. Hier steht die Bestimmung des Ersetzungsbruchpunkts im Fokus *(replacement breakdown point)*, bei dem sukzessive jeweils eine Funktion der Stichprobe durch eine „beliebig extreme“ ersetzt wird. Eine alternative Definition ist der Additionsbruchpunkt *(addition breakdown point)*, bei dem beliebig extreme Funktionen zur Stichprobe hinzugefügt werden.

3. Der größte Winkel einer festen Funktion mit einer beliebig extremen Funktion wird in dem Punkt erreicht, in dem die feste Funktion ihre betragsmäßig größte erste Ableitung besitzt.

LEMMA 5.21. *Sei $\tilde{y}$ eine gegebene, differenzierbare Funktion, $c := \operatorname{argmax}_{t\in\mathcal{T}}(|\tilde{y}'(t)|)$ und $g_c := \tilde{y}'(c)$. Dann gilt für den durchschnittlichen Schnittwinkel der Funktion mit einer*

*beliebigen Stichprobe* $\mathbf{y} = (y_1, \ldots, y_n)^\top$*, dass*

$$0 \leq m^{-1} \sum_{i=1}^{n} \sum_{k=1}^{m_i} w_{ik}(\tilde{y}(s_{ik}), y_i(s_{ik})) < w_{\sup}(\tilde{y}(c)),$$

*wobei* $w_{\sup}(\tilde{y}(c)) := \arccos\left(-\sqrt{\frac{g_c^2}{1+g_c^2}}\right)$.

*Beweis:* Mit $c$ und $g_c$ folgt, dass der supremale Schnittwinkel von $\tilde{y}$ mit einer beliebigen Funktion $y_i$ einer Stichprobe nur in $c$ auftreten kann. Sei $g_c < 0$ und $h_c$ die Ableitung einer beliebig extremen zentrierten Funktion an der Stelle $c$, welche o. B. d. A. als $y_1$ bezeichnet werde, wobei $y_1(c) = \tilde{y}(c)$ gilt. Aus $h_c \gg |g_c|$ folgt $1 + g_c h_c < 0$. Für $h_c \to \infty$ erhält man eine obere Schranke für den Schnittwinkel einer Funktion $y_i$ mit $\tilde{y}$:

$$\begin{aligned}
\cos\left(w(\tilde{y}(c), \lim_{h_c \to \infty} y_1(c))\right) &= \lim_{h_c \to \infty} \frac{1 + g_c h_c}{\sqrt{1+g_c^2}\sqrt{1+h_c^2}} \\
&= -\frac{1}{\sqrt{1+g_c^2}} \lim_{h_c \to \infty} \left| \left( \frac{(1+g_c h_c)^2}{1+h_c^2} \right)^{1/2} \right| \\
&= -\frac{1}{\sqrt{1+g_c^2}} \lim_{h_c \to \infty} \left| \left( \underbrace{\frac{1}{1+h_c^2}}_{\to 0} + \underbrace{\frac{2 g_c h_c}{1+h_c^2}}_{\to 0} + \underbrace{\frac{g_c^2 h_c^2}{1+h_c^2}}_{\to g_c^2} \right)^{1/2} \right| \\
&= -\sqrt{\frac{g_c^2}{1+g_c^2}}.
\end{aligned}$$

Daraus folgt $w_{ik}(\tilde{y}(s_{ik}), y_i(s_{ik})) < \arccos\left(-\sqrt{\frac{g_c^2}{1+g_c^2}}\right)$ für alle $i = 1, \ldots, n$ und alle $k = 1, \ldots, m_i$. Für $g_c \geq 0$ und $h_c \to -\infty$ sind die Schritte entsprechend:

$$\begin{aligned}
\cos\left(\tilde{y}(c), \lim_{h_c \to -\infty} y_1(c)\right) &= \lim_{h_c \to -\infty} \frac{1 + g_c h_c}{\sqrt{1+g_c^2}\sqrt{1+h_c^2}} \\
&= -\lim_{h_c \to \infty} \frac{1 + g_c h_c}{\sqrt{1+g_c^2}\sqrt{1+h_c^2}} = -\sqrt{\frac{g_c^2}{1+g_c^2}}.
\end{aligned}$$

Daraus folgt: $w_{\sup}(\tilde{y}(c)) = \arccos\left(-\sqrt{\frac{g_c^2}{1+g_c^2}}\right)$. □

Abbildung 5.7 veranschaulicht den Zusammenhang zwischen maximaler Steigung

einer beliebigen Funktion und dem supremalen Schnittwinkel. Das Minimum $\pi/2$ wird für konstante Funktionen erreicht.

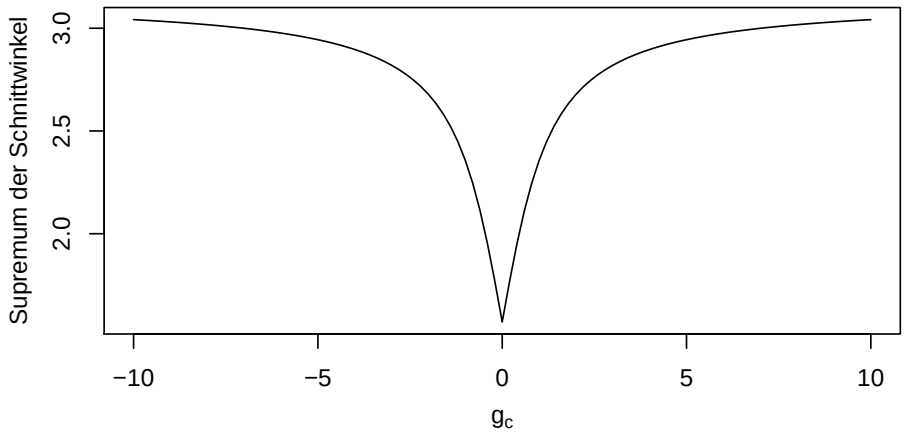



Abbildung 5.7: $w_{\sup}(\tilde{y}(c)) = \arccos\left(-\sqrt{\frac{g_c^2}{1+g_c^2}}\right)$

Somit kann der Wertebereich von FUNTA eingeschränkt werden.

KOROLLAR 5.22. *Aus Lemma 5.21 folgen*

$$d^{\mathrm{FUNTA}}(\tilde{y};\mathbf{y}) \in \,]1 - \frac{w_{\sup}(\tilde{y}(c))}{\pi}, 1]$$

*und*

$$d^{\mathrm{rFUNTA}}(\tilde{y};\mathbf{y}) \in \,]1 - \frac{w_{\sup}(\tilde{y}(c))}{\pi}, 1].$$

Bezüglich der Pseudo-Datentiefen ist der Bruchpunkt definiert als der kleinste Anteil ersetzter Funktionen in der Stichprobe $\mathbf{y} = (y_1, \ldots, y_n)^\top$, so dass $d^{\mathrm{FUNTA}}(\tilde{y};\mathbf{y})$ bzw. $d^{\mathrm{rFUNTA}}(\tilde{y};\mathbf{y})$ einen Wert am Rand des Wertebereichs (Kor. 5.22) erreicht. Da die entwickelten Pseudo-Datentiefen konzeptionell Skalenschätzern näher stehen als Lokationsschätzern, sollen nicht nur Explosions-, sondern auch Implosionsbruchpunkte hergeleitet werden. Anders als bei konventionellen Skalenschätzern ist der Wertebereich nicht durch $[0, \infty[$ begrenzt. Daher wird der Finite-Sample-Implosionsbruchpunkt als maximaler Anteil kontaminierter Beobachtungen definiert, so dass FUNTA nicht den Wert 1 erreicht. Sei $\mathbf{y}^\varepsilon$ die Stichprobe, in der $k$ Funk-

tionen von $\mathbf{y}$ durch beliebige Funktionen ersetzt wurden, wobei $\varepsilon = k/n$. Der Finite-Sample-Implosionsbruchpunkt ist dann durch

$$\varepsilon^*_{imp}(d^{\mathrm{FUNTA}}(\tilde{y};\mathbf{y})) = \min\{\varepsilon : \sup_{\mathbf{y}^\varepsilon} d^{\mathrm{FUNTA}}(\tilde{y};\mathbf{y}^\varepsilon) = 1\} \tag{5.7}$$

gegeben. Der interessantere Fall tritt auf, wenn die ersetzten Funktionen extreme Schnittwinkel hervorrufen. Der Finite-Sample-Explosionsbruchpunkt ist dann gegeben durch

$$\varepsilon^*_{exp}(d^{\mathrm{FUNTA}}(\tilde{y};\mathbf{y})) = \min\{\varepsilon : \inf_{\mathbf{y}^\varepsilon} d^{\mathrm{FUNTA}}(\tilde{y};\mathbf{y}^\varepsilon) = 1 - w_{\sup}(\tilde{y}(c))/\pi\}. \tag{5.8}$$

Die Formeln (5.7) und (5.8) lassen sich analog für rFUNTA definieren. Aus dem supremalen Schnittwinkel eines Funktionenpaares folgt der Explosionsbruchpunkt jedoch nicht unmittelbar. Je mehr Schnittpunkte die kontaminierte Funktion mit einer regulären Funktion hat, desto stärker wird der durchschnittliche Schnittwinkel beeinflusst. Die übrigen Schnittwinkel sind allerdings nur unter bestimmten Bedingungen bekannt:

BEISPIEL 5.23. *Seien $\tilde{y}'(t) \equiv \tilde{c}$ und $y_i'(t) \equiv c_i$ mit $\tilde{c}, c_i > 0$ für $i = 1,\dots,n$. Wird o. B. d. A. $y_1(t)$ durch eine multivariate Funktionenfolge $\check{y}_1^k(t) = \tilde{y}(t) + r^k \sin(r^k t)$ ersetzt, geht die Anzahl der Schnittpunkte von $\tilde{y}$ mit $\check{y}_1$ gegen unendlich, wenn $r^k \overset{k\to\infty}{\longrightarrow} \infty$. Eine Hälfte der Schnittwinkel ist gegeben durch $w_{\sup}(\tilde{y}(c))$, die andere Hälfte durch $\pi - w_{\sup}(\tilde{y}(c))$, so dass der durchschnittliche Schnittwinkel gegen $\pi/2$ tendiert. Daraus folgt auch $d^{\mathrm{FUNTA}}(\tilde{y};\check{y}_1) = 1/2$. Da $\tilde{y}$ nur eine endliche Anzahl von Schnittpunkten mit jeder anderen zentrierten Funktion $y_2,\dots,y_n$ hat, lautet der Finite-Sample-Explosionsbruchpunkt von $d^{\mathrm{FUNTA}}(\tilde{y};\check{y}_1,y_2,\dots,y_n)$ $1/n$. Wenn die Annahme fallen gelassen wird, dass alle Funktionen eine konstante, positive Ableitung haben, ist es möglich, dass der durchschnittliche Schnittwinkel* vor *der Kontaminierung der Stichprobe größer ist als $1/2$. Dies zeigt, dass der Finite-Sample-Explosionsbruchpunkt von FUNTA stark von der Stichprobe abhängt und dass allgemeine Resultate nicht ohne Weiteres herzuleiten sind.*

Anders ist dies beim Implosionsbruchpunkt.

PROPOSITION 5.24. *Sei $\mathbf{y} = (y_1(t),\dots,y_n(t))^\top, t \in [t_l, t_u]$. Der Finite-Sample-Implosionsbruchpunkt von $d^{\mathrm{FUNTA}}(\tilde{y};\mathbf{y})$ ist $1/n$.*

*Beweis:* Ersetze o. B. d. A. $\mathbf{y}$ durch $\check{\mathbf{y}} = (\check{y}_1(t), y_2(t), \ldots, y_n(t))^\top$, wobei $\check{y}_1$ eine Tangente zu $\tilde{y}$ in $m_1$ äquidistanten Punkten $t_l < t_1 < \ldots < t_{m_1} < t_u$ ist. Daher sind $m_1$ Schnittwinkel $w_1 = \ldots = w_{m_1} = 0$. Wenn $m_1$ gegen unendlich strebt, erhält man $\lim_{m_1 \to \infty} m^{-1} \sum_{i=1}^{n} \sum_{k=1}^{m_i} w_k(\tilde{y}(s_{ik}), y_i(s_{ik})) = 0$, woraus $d^{\mathrm{FUNTA}}(\tilde{y}; \check{\mathbf{y}}) = 1$ folgt. □

Für rFUNTA lassen sich die Bruchpunkte unabhängig von der Stichprobe herleiten.

PROPOSITION 5.25. *Seien* $\mathbf{y} = (y_1(t), \ldots, y_n(t))^\top, t \in [t_l, t_u]$. *Der Finite-Sample-Implosions- und Explosionsbruchpunkt von* $d^{\mathrm{rFUNTA}}(\tilde{y}; \mathbf{y})$ *ist jeweils gegeben durch* $\lfloor \frac{n+1}{2} \rfloor / n$.

*Beweis:* Der Median der maximalen Schnittwinkel ist die Grundlage von rFUNTA und vererbt deshalb den Finite-Sample-Bruchpunkt. Wird der maximale Schnittwinkel für den Implosionsbruchpunkt manipuliert, ist dies äquivalent zum Ersetzen eines Stichprobenelements wie im Beweis von Proposition 5.24 beschrieben. Hier wird jedoch für jede der $\lfloor \frac{n+1}{2} \rfloor$ kontaminierten Beobachtungen nur genau ein Schnittpunkt benötigt. Dieser Schnittpunkt hat jeweils den Winkel 0. So strebt $d^{\mathrm{rFUNTA}}(\tilde{y}; \check{\mathbf{y}})$ gegen 1. Eine Explosion der robustifizierten FUNTA Pseudo-Datentiefe kann nur auftreten, wenn mindestens die Hälfte der maximalen Schnittwinkel von $\tilde{y}$ gegen den Wert $w_{\sup}$ aus Lemma 5.21 streben. Beispielhaft ist das der Fall, wenn alle kontaminierten Funktionen identisch sind und $c$ der einzige Schnittpunkt ist. Daher müssen mindestens $\lfloor \frac{n+1}{2} \rfloor$ Funktionen manipuliert werden, deren Steigung in $c$ gegen $-\infty$ (wenn $\tilde{y}(c) > 0$) oder $\infty$ (sonst) geht. Es bezeichne $\check{\check{\mathbf{y}}}$ eine so kontaminierte Stichprobe. Dann strebt $d^{\mathrm{rFUNTA}}(\tilde{y}; \check{\check{\mathbf{y}}})$ gegen die untere Intervallgrenze von Korollar 5.22. □

Nach der Betrachtung theoretischer Eigenschaften von FUNTA und rFUNTA sollen diese Pseudo-Datentiefen in Verbindung mit dem Bootstrap auf ihre Fähigkeit, Ausreißer in realen und künstlich erzeugten Daten zu erkennen, analysiert werden. Diese Ergebnisse werden mit denen der in Abschnitt 5.1.1 vorgestellten Datentiefen verglichen.

## 5.2 AUSREISSERIDENTIFIKATION BEI UNABHÄNGIGEN FUNKTIONALEN DATEN

Die Betrachtung der Eigenschaften der hier entwickelten Pseudo-Datentiefen soll zunächst anhand eines realen Datensatzes erfolgen. Beispiel 5.26 zeigt die Performanz der Ausreißeridentifizierer basierend auf der Kombination aus Datentiefen und geglättetem Bootstrap.

BEISPIEL 5.26. Spektren von Biskuitteig

*Brown* et al. *(2001) untersuchen mit Hilfe der Nahinfrarotspektroskopie (NIR-Spektroskopie) Eigenschaften von $n = 40$ Biskuitteigen. Die Spektren der Teige wurden für $\{1100nm, 1102nm, \ldots, 2498nm\}$ gemessen und können auf $[1100, 2498]$ als funktionale Daten aufgefasst werden. Beobachtung 23 des im* R*-Paket* `fds` *(Shang und Hyndman, 2013) verfügbaren Datensatzes* `nirc` *wird von Brown* et al. *(2001) als Ausreißer bezeichnet. In Abbildung 5.8 ist dieser Ausreißer in einer Schar regulärer, grauer Kurven schwarz hervorgehoben.*

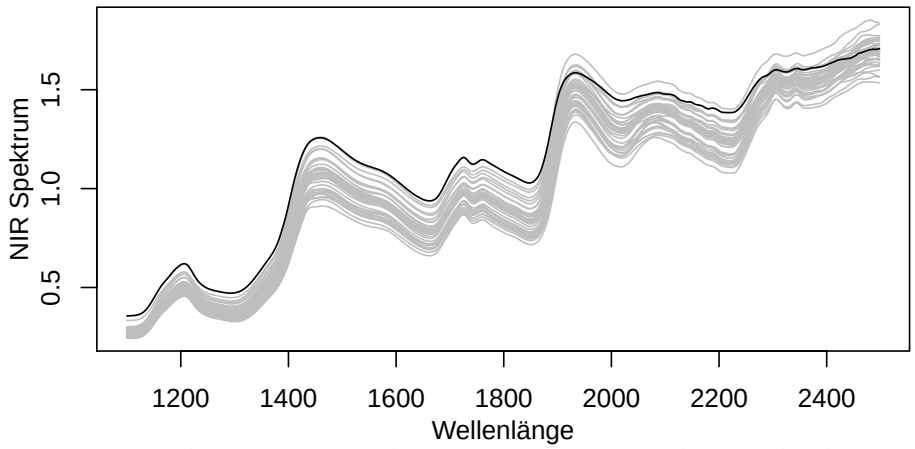



Abbildung 5.8: NIR Spektren von 40 Biskuitteigen. Grau: Reguläre Beobachtungen, schwarz: Ausreißer laut Brown *et al.* (2001).

*In diesem Datensatz wird jeder Beobachtung mit Hilfe der Datentiefen FMD, HMD und MBD sowie der hier entwickelten FUNTA und rFUNTA ein Tiefewert zugewiesen. Mit Hilfe des geglätteten Bootstraps $(B = 200)$ wird das $p$-te Perzentil $(p \in \{1, 5\})$ der jeweiligen Datentiefeverteilung geschätzt und überprüft, welche Beobachtungen Tiefewerte unterhalb der geschätzten Perzentile aufweisen. Da FUNTA und rFUNTA nur für zentrierte Daten bestimmt werden können, wird für die übrigen Datentiefen zur besseren Vergleichbarkeit*

*neben dem unzentrierten auch der zentrierte Datensatz analysiert. Die Ergebnisse sind Tabelle 5.2 zu entnehmen.*

Tabelle 5.2: Anzahl korrekt klassifizierter Ausreißer (RP) und Anzahl korrekt klassifizierter Nicht-Ausreißer (RN) im Datensatz der Biskuitteige

| | | FMD | | HMD | | MBD | | FUNTA | | rFUNTA | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $p$ | RP | RN | RP | RN | RP | RN | RP | RN | RP | RN |
| Unzentriert | 1 | 0 | 38 | 0 | 38 | 0 | 38 | / | / | / | / |
| | 5 | 1 | 37 | 1 | 37 | 0 | 37 | / | / | / | / |
| Zentriert | 1 | 0 | 39 | 0 | 39 | 0 | 39 | 1 | 39 | 1 | 39 |
| | 5 | 1 | 37 | 1 | 39 | 1 | 39 | 1 | 38 | 1 | 39 |

*Das perfekte Resultat mit einer richtig-positiven und 39 richtig-negativen Beobachtungen liefert rFUNTA, da für beide Perzentile ausschließlich der tatsächliche Ausreißer identifiziert wird. Auch FUNTA liefert ein perfektes Resultat für $p = 1$, allerdings wird für $p = 5$ ein Nicht-Ausreißer falsch klassifiziert. Die anderen Datentiefen liefern nur für $p = 5$ sinnvolle Ergebnisse, allerdings werden für die unzentrierten Daten jeweils zwei falsche Ausreißer klassifiziert. Für die zentrierten Daten finden alle Verfahren bei $p = 5$ den korrekten Ausreißer, allerdings findet die integrierte Datentiefe zusätzlich zwei falsche Ausreißer.*

In den folgenden beiden Abschnitten wird die Performanz der Ausreißeridentifizierer bestehend aus einer (Pseudo-)Datentiefe und der Bootstrap-Approximation der Quantile der Verteilung der Datentiefe bei unabhängigen funktionalen Daten untersucht. Dabei wird in Abschnitt 5.2.1 die Planung der Simulationsstudie dargelegt und in Abschnitt 5.2.2 ihre Auswertung vorgenommen.

### *5.2.1 Planung der Simulationsstudie*

In diesem Abschnitt wird geprüft, inwieweit funktionale Form-Ausreißer in einem *IID setting* von den besprochenen Datentiefen und der Bootstrap-Prozedur korrekt klassifiziert werden. Allgemein gilt: Es werden 100 Datensätze mit jeweils 95 identisch verteilten regulären Beobachtungen und fünf identisch verteilten Ausreißern betrachtet. Außerdem werden auch Situationen herangezogen, in denen alle 100 Beobachtungen identisch verteilt sind und somit als regulär angesehen werden. Die

Grenzwerte für den Bootstrap-basierten Ausreißeridentifizierer werden auf Basis von 100 geglätteten Bootstrap-Stichproben gezogen. Als Indexmenge wird das Intervall $[0, 1]$ gewählt, bzw. die Menge $\{0, 0.01, 0.02, \ldots, 1.00\}$ in R. Die Szenarien (beispielhaft abgebildet in Abb. 5.9) werden wie folgt aufgelistet:

1. Gegeben sei ein Gaußprozess $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$ mit $\mu(t) = 8t\sin(40t)+40t^2$, in dem $\sigma(s,t) = \sigma_{QE}(s,t) = \exp(-50(s-t)^2)$ für die regulären Beobachtungen und die Matérn-Kovarianzfunktion $\sigma(s,t) = \sigma_{Mat}(s,t) = \sqrt{2|s-t|/\pi}K_{0.5}(|s-t|)$ für die Ausreißer gilt, wobei $K_{0.5}$ die modifizierte Besselfunktion zweiter Art der Ordnung $0.5$ ist (Abb. 5.9, oben links).

2. Sei $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$ mit $\mu \equiv 0$ gegeben, wobei $\sigma_{QE}(s,t) = \exp(-50(s-t)^2)$ für die regulären Beobachtungen und die Matérn-Kovarianzfunktion $\sigma_{Mat}(s,t) = \sqrt{2|s-t|/\pi}K_{0.5}(|s-t|)$ für die Ausreißer gilt (Abb. 5.9, oben rechts).

3. Sei $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$ gegeben, wobei $\mu(t) = 30(1-t)t^{1.2}$ für die regulären Beobachtungen und $\mu(t) = 30(1-t)^{1.2}t$ für die Ausreißer gilt. Für beide Populationen gilt $\sigma(s,t) = 0.2\exp(-|s-t|/0.3)$. Dieses Szenario wurde auch von Cuevas *et al.* (2007) analysiert (Abb. 5.9, Mitte links).

4. Gegeben sei mit $y(t) = x_0 + \sum_{i=1}^{4}(x_{2i-1}\sin(2\pi it) + x_{2i}\cos(2\pi it))$ ein Prozess in Fourierdarstellung. Die Koeffizienten $x_0, \ldots, x_8$ sind jeweils Realisationen der Zufallsvariable $\mathcal{X} \sim \text{Beta}(p, q)$, wobei für die regulären Beobachtungen $p = q = 2$ und für die Ausreißer $p = q = 0.5$ gilt. Dieses Szenario wurde bereits in Kuhnt und Rehage (2016) in kleinerem Umfang analysiert (Abb. 5.9, Mitte rechts).

5. Gegeben seien Realisationen, die der stochastischen Differentialgleichung $dX_t = \log(X_t)dt + \sigma dW_t$ genügen, wobei $W_t$ ein Wiener Prozess ist. Weiterhin sind der Startwert $X_0 = 20$ und der Driftkoeffizient $\log(x)$ vorgegeben. Für die regulären Beobachtungen ist der Diffusionskoeffizient $\sigma = 1.5$ gegeben, für die Ausreißer gilt $\sigma = 3$. Die Daten wurden in R mittels der Funktion `sde.sim` generiert (Iacus, 2008). Dieses Szenario wurde bereits in Kuhnt und Rehage (2016) in kleinerem Umfang analysiert (Abb. 5.9, unten links).

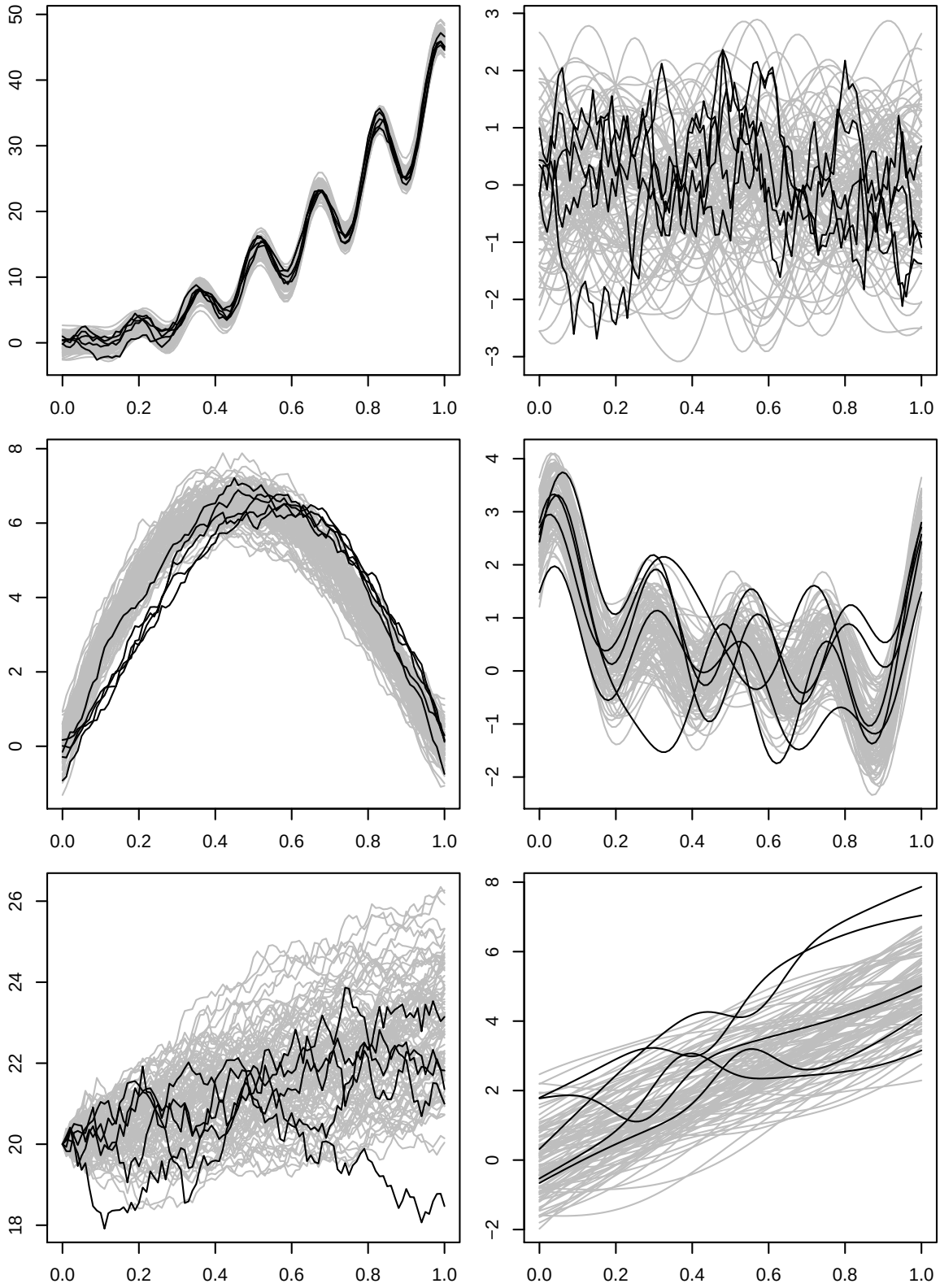


Abbildung 5.9: Beispielplots der sechs Simulationsszenarien im *IID setting* mit regulären Daten (grau) und Ausreißern (schwarz)

6. Gegeben sei ein Gaußprozess $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$ mit $\mu(t) = 5t$ für die regulären Beobachtungen und $\mu(t) = 5t + (-1)^{x_B}(1.2 - 1/(3\sqrt{\pi/50})\exp(-80(t - x_R)^2))$ für die Ausreißer. Dabei ist $x_B$ für jeden Ausreißer eine Realisation von $\mathcal{X}_B \sim \text{Bin}(1, 0.5)$ und $x_R$ eine Realisation von $\mathcal{X}_R \sim R[0.25, 0.75]$. Für beide Populationen gilt $\sigma(s, t) = \exp(-3.125(s - t)^2)$. Ein ähnliches Szenario wurde auch von Arribas-Gil und Romo (2014) analysiert (Abb. 5.9, unten rechts).

Wie Abbildung 5.9 zu entnehmen ist, sind die Ausreißer nicht in allen Situationen mit bloßem Auge von den Nicht-Ausreißern zu unterscheiden. Vornehmlich ist ihre *outlyingness* dadurch motiviert, dass sie mit Hilfe eines abweichenden stochastischen Prozess generiert wurden. Dies erhöht den Schwierigkeitsgrad für die Ausreißeridentifizierer. Des Weiteren sei angemerkt, dass die zugrunde liegenden stochastischen Prozesse teilweise nicht differenzierbar sind. Die resultierenden Datensätze werden aufgefasst als diskretisierte funktionale Daten, die mit oder ohne Messfehler beobachtet wurden. Da die Funktionen auf einem sehr feinen Gitter realisiert wurden, können die Tangenten in den Schnittpunkten durch lineare Interpolation geschätzt werden. Somit existiert das Problem der Nicht-Differenzierbarkeit nur auf der theoretischen Ebene.

### *5.2.2 Auswertung der Simulationsstudie*

Die Ergebnisse der Simulationsstudie werden in den Tabellen 5.3 (mit Ausreißern) und 5.4 (ohne Ausreißer) festgehalten und können aufgeschlüsselt nach Szenario wie folgt zusammengefasst werden:

1. Die Richtig-Positiv-Raten der FMD, HMD und MBD sind sehr klein. Dabei spielt die Zentrierung der Daten keine nennenswerte Rolle. Die Detektionsraten von FUNTA und rFUNTA sind deutlich besser. Besonders rFUNTA erkennt selbst bei $p = 1$ fast alle Ausreißer ($99.6\%$), hat aber auch eine deutlich kleinere Richtig-Negativ-Rate als alle anderen Datentiefen ($96.9\%$). Die anderen Datentiefen liegen mit etwa $99\%$ bzw. $95\%$ nah an den für das jeweilige $p$ erwarteten Werten; FUNTA sogar deutlich darüber. In der Situation ohne

Tabelle 5.3: Simulationsergebnisse des funktionalen *IID settings* mit Ausreißern für unzentrierte (U) und zentrierte (Z) Daten

| | | | FMD | | HMD | | MBD | | FUNTA | | rFUNTA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Szenario | | $p$ | RPR | RNR | RPR | RNR | RPR | RNR | RPR | RNR | RPR | RNR |
| 1 | U | 1 | 0.008 | 0.989 | 0.008 | 0.991 | 0.006 | 0.990 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 0.020 | 0.946 | 0.032 | 0.948 | 0.026 | 0.949 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | 0.002 | 0.991 | 0.006 | 0.993 | 0.004 | 0.992 | 0.218 | 0.999 | 0.996 | 0.969 |
| | Z | 5 | 0.018 | 0.946 | 0.030 | 0.949 | 0.018 | 0.948 | 0.774 | 0.982 | 0.998 | 0.908 |
| 2 | U | 1 | 0.008 | 0.989 | 0.008 | 0.991 | 0.006 | 0.990 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 0.020 | 0.946 | 0.032 | 0.948 | 0.024 | 0.950 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | 0.002 | 0.991 | 0.006 | 0.993 | 0.000 | 0.992 | 0.000 | 0.994 | 0.592 | 1.000 |
| | Z | 5 | 0.018 | 0.947 | 0.030 | 0.950 | 0.020 | 0.948 | 0.000 | 0.953 | 1.000 | 0.973 |
| 3 | U | 1 | 0.154 | 0.994 | 0.476 | 0.995 | 0.148 | 0.996 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 0.608 | 0.960 | 0.842 | 0.957 | 0.628 | 0.965 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | 0.408 | 0.999 | 0.702 | 0.999 | 0.332 | 0.999 | 0.170 | 0.999 | 0.000 | 0.996 |
| | Z | 5 | 0.880 | 0.970 | 0.946 | 0.955 | 0.876 | 0.974 | 0.510 | 0.973 | 0.016 | 0.963 |
| 4 | U | 1 | 0.188 | 0.999 | 0.280 | 1.000 | 0.188 | 1.000 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 0.524 | 0.975 | 0.816 | 0.989 | 0.576 | 0.977 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | 0.184 | 0.999 | 0.278 | 1.000 | 0.196 | 0.999 | 0.082 | 0.994 | 0.196 | 0.993 |
| | Z | 5 | 0.550 | 0.979 | 0.798 | 0.989 | 0.600 | 0.982 | 0.248 | 0.952 | 0.450 | 0.940 |
| 5 | U | 1 | 0.126 | 0.986 | 0.288 | 0.990 | 0.120 | 0.990 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 0.298 | 0.939 | 0.492 | 0.948 | 0.296 | 0.947 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | 0.140 | 0.990 | 0.372 | 0.992 | 0.142 | 0.992 | 0.004 | 0.991 | 0.248 | 1.000 |
| | Z | 5 | 0.332 | 0.946 | 0.622 | 0.948 | 0.352 | 0.951 | 0.024 | 0.947 | 0.818 | 0.992 |
| 6 | U | 1 | 0.068 | 0.987 | 0.050 | 0.994 | 0.052 | 0.989 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 0.206 | 0.935 | 0.228 | 0.958 | 0.204 | 0.944 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | 0.028 | 0.990 | 0.038 | 0.993 | 0.020 | 0.990 | 0.066 | 0.988 | 0.478 | 0.988 |
| | Z | 5 | 0.082 | 0.942 | 0.152 | 0.957 | 0.078 | 0.947 | 0.418 | 0.949 | 0.640 | 0.933 |

Ausreißer zeigt sich ein ähnliches Bild. Die RNR von FUNTA ist allerdings deutlich näher an den erwarteten Werten und die RNR von rFUNTA ist etwas näher an den erwarteten Werten, aber immer noch darunter.

2. Die Ergebnisse der FMD, HMD und MBD unterscheiden sich im Vergleich zu Szenario 1 nur marginal. Die RPR von FUNTA sinkt von $21.8\%$ bzw. $77.4\%$ auf $0\%$. Daraus lässt sich schließen, dass in dieser Situation trotz unterschiedlicher Form der Kovarianzkerne kein Unterschied bzgl. des durchschnittlichen Schnittwinkels besteht. Der maximale paarweise Schnittwinkel ist jedoch im-

Tabelle 5.4: Richtig-Negativ-Raten des funktionalen *IID settings* ohne Ausreißer für unzentrierte (U) und zentrierte (Z) Daten

| Szenario | | $p$ | FMD | HMD | MBD | FUNTA | rFUNTA |
|---|---|---|---|---|---|---|---|
| 1 | U | 1 | 0.990 | 0.993 | 0.991 | / | / |
| | U | 5 | 0.948 | 0.950 | 0.951 | / | / |
| | Z | 1 | 0.992 | 0.992 | 0.993 | 0.991 | 0.974 |
| | Z | 5 | 0.949 | 0.949 | 0.950 | 0.950 | 0.925 |
| 2 | U | 1 | 0.990 | 0.992 | 0.992 | / | / |
| | U | 5 | 0.948 | 0.949 | 0.951 | / | / |
| | Z | 1 | 0.992 | 0.992 | 0.992 | 0.995 | 0.989 |
| | Z | 5 | 0.949 | 0.949 | 0.950 | 0.960 | 0.935 |
| 3 | U | 1 | 0.987 | 0.989 | 0.989 | / | / |
| | U | 5 | 0.940 | 0.939 | 0.943 | / | / |
| | Z | 1 | 0.989 | 0.989 | 0.990 | 0.993 | 0.996 |
| | Z | 5 | 0.942 | 0.939 | 0.944 | 0.955 | 0.965 |
| 4 | U | 1 | 0.994 | 0.999 | 0.996 | / | / |
| | U | 5 | 0.955 | 0.968 | 0.957 | / | / |
| | Z | 1 | 0.994 | 0.997 | 0.996 | 0.991 | 0.987 |
| | Z | 5 | 0.961 | 0.972 | 0.962 | 0.944 | 0.924 |
| 5 | U | 1 | 0.981 | 0.987 | 0.986 | / | / |
| | U | 5 | 0.927 | 0.937 | 0.936 | / | / |
| | Z | 1 | 0.987 | 0.987 | 0.988 | 0.990 | 0.997 |
| | Z | 5 | 0.933 | 0.937 | 0.938 | 0.946 | 0.976 |
| 6 | U | 1 | 0.985 | 0.993 | 0.986 | / | / |
| | U | 5 | 0.928 | 0.951 | 0.938 | / | / |
| | Z | 1 | 0.989 | 0.992 | 0.989 | 0.985 | 0.979 |
| | Z | 5 | 0.941 | 0.951 | 0.945 | 0.935 | 0.920 |

mer noch hilfreich zur Bestimmung der Ausreißer. rFUNTA identifiziert fast $60\%$ der Ausreißer korrekt bei $p = 1$, wobei alle Nicht-Ausreißer korrekt klassifiziert werden. Für $p = 5$ werden sogar alle Ausreißer korrekt klassifiziert, und nur $2.7\%$ der Nicht-Ausreißer falsch zugeordnet. Bei Betrachtung der Situation ohne Ausreißer fallen kaum Besonderheiten auf. Nur bei $p = 5$ ist die RNR für rFUNTA etwas kleiner als der erwartete Wert ($93.5\%$).

3. Die Unterschiede zwischen Ausreißern und Nicht-Ausreißern beziehen sich in diesem Szenario nur auf die Erwartungsfunktion. Die Amplitude und generelle parabolische Form sind jedoch sehr ähnlich. Trotzdem sind die Ergebnisse

für die zentrierten Kurven durchgehend besser als für die unzentrierten. Die besten Ergebnisse liefert die HMD mit bis zu $95\%$ erkannter Ausreißer und Richtig-Negativ-Raten, die etwas besser als erwartet sind. FMD und MBD liefern relativ ähnliche Ergebnisse und können bis zu $88\%$ der Ausreißer identifizieren. FUNTA kann dagegen nur maximal jeden zweiten Ausreißer erkennen. rFUNTA ist in diesem Szenario unbrauchbar, die paarweisen maximalen Schnittwinkel unterscheiden sich also zu wenig. In der Situation ohne Ausreißer fällt auf, dass die Richtig-Negativ-Raten von FMD, HMD und MBD leicht unter dem erwarteten Niveau liegen und die von FUNTA und rFUNTA leicht über dem erwarteten Niveau.

4. Die beste Performanz liefert die HMD, gleichermaßen bei zentrierten und unzentrierten Daten erkennt sie etwa $80\%$ der Ausreißer. Etwas schlechtere Ergebnisse liefern FMD und MBD. Bei $p = 1$ ist die RPR von rFUNTA ähnlich gut, bei $p = 5$ jedoch deutlich schlechter als die von FMD und MBD. Die schlechtesten Ergebnisse liefert FUNTA. In Bezug auf die Datensituationen ohne Ausreißer fällt auf, dass rFUNTA einen zu niedrigen Wert für $p = 5$ aufweist.

5. Die Ausreißererkennung ist in diesem Beispiel bei vorheriger Zentrierung deutlich besser als ohne. Für $p = 1$ hat HMD die besten Ergebnisse ($37.2\%$), gefolgt von rFUNTA, MBD, FMD und FUNTA. Für $p = 5$ ist hingegen rFUNTA deutlich am besten ($81.8\%$), gefolgt von den übrigen Datentiefen in derselben Reihenfolge. Bei den Datensituationen ohne Ausreißer sind nur FUNTA und rFUNTA wie erwartet oder besser, wohingegen FMD, HMD und MBD durchgängig zu niedrige RNR haben.

6. Das letzte Szenario ist das einzige, bei dem die Performanz auf den unzentrierten Beobachtungen deutlich besser ist als auf den zentrierten. rFUNTA besitzt die höchste RPR ($64\%$), mit großem Abstand vor FUNTA, welche wiederum einen großen Abstand vor HMD, FMD und MBD hat. Allerdings ist die RNR von rFUNTA unter den erwarteten Werten, was auch auf die FMD zutrifft. Dies bestätigt sich beim Blick auf die Ausreißer-freie Situation.

Die Simulationsergebnisse der $h$-modal Datentiefe sind bezüglich der in `depth.mode` voreingestellten Argumente (Distanzmaß für Kurven: $L_2$-Norm, $h$: $15\%$-Quantil der paarweisen Distanzen der Kurven, Trimmanteil $25\%$) zu interpretieren. Variationen des Trimmanteils $(1\%)$ und der Normen $(L_1, L_\infty)$ führten nur zu marginalen Veränderungen bzgl. des geschätzten Cutoffs, weshalb zugunsten der Übersichtlichkeit auf die explizite Nennung der Ergebnisse verzichtet wird. Für die modifizierte Bandtiefe wurde der Tuningparameter $J = 2$ gewählt, da für diesen Fall ein schneller Algorithmus zur Verfügung steht (Sun *et al.*, 2012).

## 5.3 AUSREISSERIDENTIFIKATION BEI ABHÄNGIGEN FUNKTIONALEN DATEN

Bisher wurde die Ausreißererkennung im funktionalen Kontext ausschließlich unter der Annahme unabhängiger Beobachtungen thematisiert. Häufig resultieren sie jedoch als abhängige Variablen aus kontrollierten Zufallsexperimenten oder Beobachtungsstudien. Gerade dann ist es von Interesse, Ausreißer in den funktionalen Zielgrößen zu identifizieren, da sie Hinweise auf Messfehler, Änderungen des datengenerierenden Prozesses sowie eine fehlspezifizierte Modell- oder Variablenselektion geben können. Letzteres ist einer der Hauptvorteile funktionaler Ausreißeridentifikation, da ein scheinbarer Ausreißer durch ein anderes Modell oder zusätzliche Kovariablen zu einer unauffälligen Beobachtung werden kann. Dabei wird die generalisierte *function-on-scalar regression* (GFOSR, Kap. 3.2.2) zur Modellierung der Zielgröße verwendet. Teilergebnisse dieses Abschnitts wurden bereits in Kuhnt, Rehage, Becker-Emden, Tillmann und Hussong (2016) vorab veröffentlicht.

BEISPIEL 5.27. Hochgeschwindigkeitsflammspritzen

*Beim Hochgeschwindigkeitsflammspritzen kann die Geschwindigkeit der Partikel im Flug über die Zeit gemessen und somit als funktionale Variable aufgefasst werden. Sie hängt von den Maschinenparametern Kerosin (Ker), Lambda (Lam, dem Sauerstoff-Kerosin-Verhältnis), Distanz (Sod) und Förderrate (FDV) ab. Darauf basierend wird ein zentral zusammengesetzter Plan (CCD) erstellt, der in Tillmann* et al. *(2012) aufgeführt wird, dort jedoch zur Modellierung der Mittelwerte der Partikeleigenschaften genutzt wird. Der CCD besteht aus 30 Läufen und ermöglicht es, sowohl quadratische als auch Zweifach-*

*Interaktionseffekte zu schätzen. Mit Hilfe von generalisierter* function-on-scalar regression *lässt sich die Geschwindigkeit funktional modellieren. Zunächst wird ein passendes Basissystem benötigt. Da die Geschwindigkeit auf den ersten Blick keine klare Periodizität erkennen lässt, wird ein kubisches B-Spline-Basissystem verwendet. Die Anzahl der Basisfunktionen wird per generalisierter Kreuzvalidierung (Craven und Wahba, 1979) geschätzt. Außerdem muss eine Variablenselektion durchgeführt werden. Neben den Haupteffekten stehen die quadratischen Effekte sowie Zweifach-Interaktionseffekte zur Verfügung. Da funktionale Regression häufig anfällig für Überanpassung ist, wird ein restriktives Optimalitätskriterium gewählt, die integrierte vorhergesagte Residuenquadratsumme (PRESS):*

$$PRESS = n^{-1} \sum_{i=1}^{n} \left( (b-a)^{-1} \int_{\mathcal{T}} (y_i(t) - \hat{\mu}_{i,-i}(t))^2 dt \right)^{1/2},$$

*wobei $\mathcal{T} = [a, b]$ und $\hat{\mu}_{i,-i}(t)$ die $i$-te geschätzte Zielvariable bezeichnet, geschätzt basierend auf allen außer der $i$-ten beobachteten Zielgröße. Es wird eine Vorwärtsselektion durchgeführt, beginnend mit den vier Haupteffekten. Generell soll eine passende Linkfunktion und Verteilung im GLM-Kontext* a priori *auf Basis substanzwissenschaftlicher Erkenntnisse gewählt werden. Für das Hochgeschwindigkeitsflammspritzen liegen bisher nur im nichtfunktionalen Fall Erkenntnisse vor (Hoyden, 2011). Daher wird auch die Wahl von Linkfunktion und Verteilung von der PRESS-Statistik abhängig gemacht.*

*Tabelle 5.5 zeigt die Variablenselektion für jeden betrachteten Typus der GFOSR. Das PRESS-minimale Modell für die Geschwindigkeit lautet*

$$\begin{aligned} \hat{\mu}_{Vel}(t) = \exp \Big[ & \hat{\beta}_0 + \hat{\beta}_{Ker}(t) Ker + \hat{\beta}_{Lam}(t) Lam + \hat{\beta}_{FDV}(t) FDV + \hat{\beta}_{Sod}(t) Sod \\ & + \hat{\beta}_{Ker^2}(t) Ker^2 + \hat{\beta}_{KerLam}(t) Ker \cdot Lam + \hat{\beta}_{KerFDV}(t) Ker \cdot FDV \Big] \end{aligned} \tag{5.9}$$

*mit gammaverteilter Zielgröße und $PRESS = 4.833$. Der quadratische Kerosin-Effekt liefert dabei die größte Verbesserung der nicht-linearen Effekte.*

*Nachfolgend wird analysiert, ob Ausreißer in den geschätzten Devianzresiduen auftreten. Dabei werden die Datentiefen und Pseudo-Datentiefen in Verbindung mit dem geglätteten Bootstrap und $p = 1$ verwendet. Bei einem größeren Wert wie $p = 5$ wäre bei der*

Tabelle 5.5: PRESS-minimale Modelle für jeden Typus der GFOSR

| Verteilung | $g(\mu)$ | Variablenselektion | PRESS |
|---|---|---|---|
| Gamma | $\mu$ | Ker, Lam, Sod, FDV, Ker$^2$ | 4.991 |
| | $\mu^{-1}$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·Lam, Ker·FDV | 4.932 |
| | $\log(\mu)$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·Lam, Ker·FDV | **4.833** |
| Normal | $\mu$ | Ker, Lam, Sod, FDV, Ker$^2$ | 5.021 |
| | $\mu^{-1}$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·Lam, Ker·FDV | 5.000 |
| | $\log(\mu)$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·Lam, Ker·FDV | 4.892 |

*Stichprobengröße bereits mehr als ein Ausreißer zu erwarten, was hier nicht erwünscht ist. Tabelle 5.6 zeigt, dass MBD und HMD (auf den unzentrierten Daten) denselben Lauf als Ausreißer deklarieren, wohingegen rFUNTA (auf den zentrierten Daten) einen anderen Lauf findet. Weiterhin konnte die PRESS-Statistik nach dem Entfernen dieser beiden Läufe um* $4.7\%$ *verringert werden.*

Tabelle 5.6: Als Ausreißer identifizierte Läufe im HVOF-Prozess für das PRESS-minimale Modell mit $p = 1$ und PRESS vor und nach Entfernung der Ausreißer

| | | | | | | PRESS | |
|---|---|---|---|---|---|---|---|
| Verteilung | $g(\mu)$ | MBD | HMD | FUNTA | rFUNTA | vorher | nachher |
| Gamma | $\log(\mu)$ | 26 | 26 | ⁄ | 19 | 4.833 | 4.604 |

*Die geschätzten Devianzresiduen der Geschwindigkeit zeigen, dass eine Kurve außerhalb der konvexen Hülle aller anderen Kurven liegt. Diese Kurve ist in Abbildung 5.10 punktiert gekennzeichnet und entstammt dem Lauf 26, der durch die Einstellung (Ker, Lam, Sod, FDV) =* $(0, 0, 0, -2)$ *verursacht wird. Diese Kurve wird von HMD und MBD als Ausreißer identifiziert. Für die zentrierten Kurven ist eine andere Kurve auffällig: Die gestrichelte Kurve verfügt über einen relativ steilen Anstieg um Sekunde 20. Diese Kurve, die von rFUNTA als Ausreißer identifiziert wird, entstammt Lauf 19 und wird durch einen Nulllauf verursacht.*

*Die Vorwärtsselektion wurde für die verbleibenden 28 Läufe erneut durchgeführt und resultiert im PRESS-minimalen Modell*

$$\begin{aligned}\hat{\mu}_{Vel}(t) = \hat{\beta}_0 + \hat{\beta}_{Ker}(t)Ker + \hat{\beta}_{Lam}(t)Lam + \hat{\beta}_{FDV}(t)FDV + \hat{\beta}_{Sod}(t)Sod \\ + \hat{\beta}_{Ker^2}(t)Ker^2 + \hat{\beta}_{KerFDV}(t)Ker \cdot FDV + \hat{\beta}_{KerLam}(t)Ker \cdot Lam \end{aligned} \tag{5.10}$$

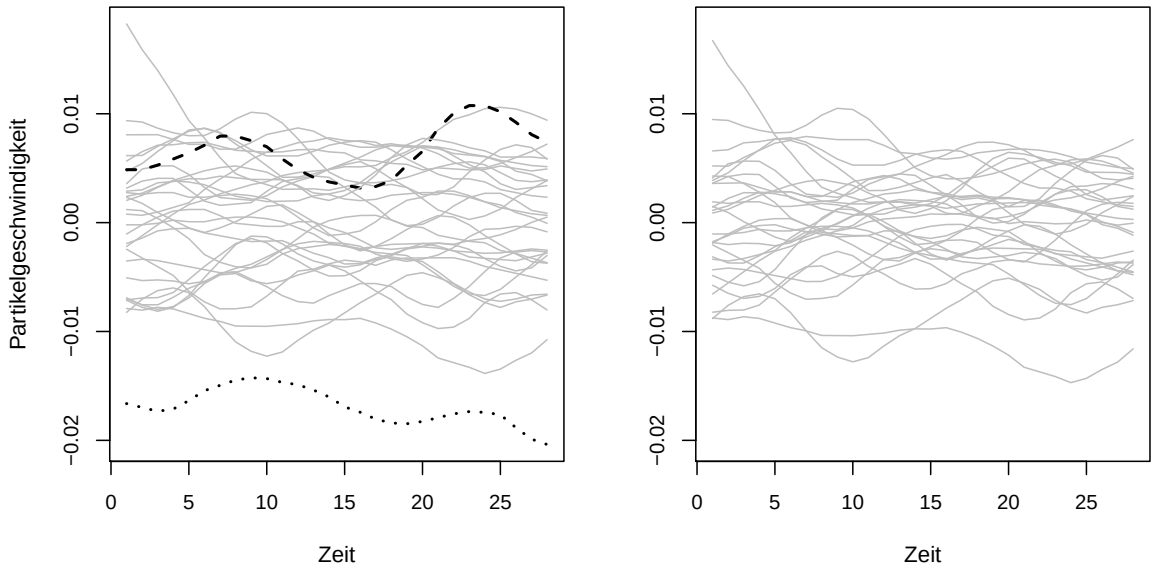



Abbildung 5.10: Geschätzte Devianzresiduen des PRESS-optimalen Modells für die Geschwindigkeit, vor (linke Seite) und nach (rechte Seite) der Ausreißerentfernung. Lauf 19 (gestrichelt) und Lauf 26 (punktiert) sind hervorgehoben.

*mit gammaverteilter Zielgröße. Gleichung* (5.10) *beinhaltet dieselben Kovariablen wie* (5.9)*, allerdings in einer anderen Reihenfolge und mit einer anderen Linkfunktion. Die korrespondierenden geschätzten Devianzresiduen sind auf der rechten Seite von Abbildung 5.10 abgebildet. Die Residuen unterscheiden sich nicht besonders voneinander, wenn man von den beiden entfernten Beobachtungen absieht. Die PRESS-Statistik verringert sich durch die zusätzlich veränderte Linkfunktion sogar auf* $4.506$*, das entspricht einer Reduktion um* $6.8\%$*.*

Tabelle 5.7: PRESS-minimale Modelle für jeden Typus nach der Ausreißerentfernung

| Verteilung | $g(\mu)$ | Variablenselektion | PRESS |
|---|---|---|---|
| Gamma | $\mu$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·FDV, Ker·Lam | **4.506** |
| | $\mu^{-1}$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·Lam, Ker·FDV | 4.780 |
| | $\log(\mu)$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·Lam, Ker·FDV | 4.604 |
| Normal | $\mu$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·FDV, Ker·Lam | 4.549 |
| | $\mu^{-1}$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·Lam | 4.736 |
| | $\log(\mu)$ | Ker, Lam, Sod, FDV, Ker$^2$, Ker·Lam, Ker·FDV | 4.654 |

Die Planung einer Simulationsstudie für funktionale Daten, die von nichtstochastischen, skalaren Kovariablen abhängig sind, wird in Abschnitt 5.3.1 thematisiert und in Abschnitt 5.3.2 ausgewertet. Zusätzlich ist von Interesse, inwieweit

falsch spezifizierte GLM-Familien und Linkfunktionen die Güte der Ausreißererkennung beeinflussen.

### *5.3.1 Planung der Simulationsstudie*

Die Simulationsstudie betrifft die funktionalen geschätzten Residuen der GFOSR. Die Datentiefe der Residuen wird herangezogen, um zu überprüfen, ob funktionale Ausreißer vorliegen. Dabei werden HMD, MBD, FUNTA und rFUNTA dahingehend verglichen, ob sie bei unterschiedlichen Situationen der GFOSR in der Lage sind, die Ausreißer zu identifizieren. Auf die integrierte Datentiefe wird hier verzichtet, da sie zur gleichen Klasse von Datentiefen gehört wie die modifizierte Bandtiefe (Nagy *et al.*, 2016) und in Abschnitt 5.2.2 sehr ähnliche Ergebnisse lieferte. Außerdem ist ihre Bestimmung rechenaufwändiger als die der modifizierten Bandtiefe. Es werden alle Kombinationen der angenommenen Fehlerstruktur (Gamma- und Normalverteilung) und Linkfunktion (Identitätslink, inverser Link, Log-Link) betrachtet. Die wahre Linkfunktion ist dabei der Identitätslink. Die Daten werden jeweils einmal entsprechend der Gamma- und Normalverteilung generiert. Als Residuentyp werden die Devianzresiduen herangezogen.

Da vollfaktorielle Designs typisch für industrielle Anwendungen wie das Hochgeschwindigkeitsflammspritzen sind, wird ein $2^6$ Versuchsplan gewählt. Sei $\mathbf{X} \in \mathbb{R}^{2^k \times k}$ die vollfaktorielle Designmatrix mit Einträgen $x_{i,j} \in \{-1, 1\}$.

Die wahren, regulären Daten werden durch die Formel

$$\Upsilon_i(t) = \phi_0 + \sum_{j=1}^{k/2} \left[\phi_{2j-1} \sin(2\pi jt) x_{i,2j-1} + \phi_{2j} \cos(2\pi jt) x_{i,2j}\right], \quad t \in [0,1], \quad (5.11)$$

generiert, wobei $\phi_0 \sim \aleph(10, 1)$ gewählt wird um negative Werte zu vermeiden und $\phi_j \sim \aleph(0, 1), j = 1, \ldots, k$. Sei $\Upsilon_{i,beo}(t)$ eine Realisation von $\Upsilon_i(t)$. Dann wird $\Upsilon_{i,beo}(t)$ als wahre, zu schätzende Funktion aufgefasst. Die Verteilung der Zufallsvariable $\mathcal{Y}_i(t)$, die zu Realisation $y_i(t), i = 1, \ldots, n, t \in \{0, 0.01, \ldots, 1\}$ korrespondiert, werde so gewählt, dass $\mathsf{E}(\mathcal{Y}_i(t)) = \Upsilon_{i,beo}(t)$ und $\mathsf{Cov}(\mathcal{Y}_i(t)) = 1$: Mit $\mathcal{Y}_i(t) \sim$

$\Gamma(\Upsilon_{i,beo}(t)^2, \Upsilon_{i,beo}(t)^{-1})$ und $\mathcal{Y}_i(t) \sim \aleph(\Upsilon_{i,beo}(t), 1)$ werden zwei der gängigsten stetigen Verteilungsfamilien abgedeckt, die auch Exponentialfamilien sind. Die Modellparameter werden basierend auf der angenommenen Verteilung der Zielgröße und der angenommenen Linkfunktion in R mit der Funktion `pffr` (R-Paket `refund`, Goldsmith *et al.*, 2016) geschätzt. Jede Simulationseinstellung wird 100 Mal wiederholt. Als Basissystem werden kubische B-Splines mit acht Basisfunktionen gewählt, um geschachtelte Optimierungsschritte zu vermeiden. Außerdem wird so verhindert, dass die Schätzungen unrealistisch gut werden, da sich das angenommene (B-Splines) und das tatsächliche (Fourier) Basissystem unterscheiden. Des Weiteren wird für die Schätzung des Grenzwertes im Ausreißeridentifizierer die Bootstrap-Stichprobengröße $B = 100$ gewählt.

Von den 64 Beobachtungen werden $n_o = 4$ Ausreißer generiert, wobei zwei unterschiedliche Konzepte zugrunde gelegt werden. Welcher Lauf zu einem Ausreißer führt, muss in jedem Simulationsschritt randomisiert ausgewählt werden. Die Güte der Ausreißeridentifizierer wird wieder mit der Richtig-Positiv-Rate (RPR) und der Richtig-Negativ-Rate (RNR) gemessen.

1. *Trend-Ausreißer:* Für funktionale Residuen wird im Allgemeinen angenommen, dass sie ohne erkennbares Muster um null streuen. Trend-Ausreißer verfügen über solch ein Muster, da sie linear über die Zeit hinweg von null abweichen. Ein solches Verhalten wurde zum Beispiel von Borowski *et al.* (2014) beschrieben, wenn beim Hochgeschwindigkeitsflammspritzen ein Druckabfall des Fördergases auftritt. Formal wird ein Trend-Ausreißer durch $\Upsilon_i(t) = \phi_0 + \sum_{j=1}^{k/2}[\phi_{2j-1}\sin(j2\pi t)x_{i,2j-1} + \phi_{2j}\cos(j2\pi t)x_{i,2j}] - \phi_{k+1}t, t \in [0,1]$ beschrieben, wobei $\phi_{k+1}$ ein Parameter für die Stärke des Ausreißers ist. Die Ergebnisse sind in Tabelle 5.8 zusammengefasst.

2. *Form-Ausreißer:* Ein weiteres Problem in industriellen Anwendungen ist das Auftreten von verrauschten Daten. Ungewöhnlich starkes Rauschen kann auf defekte Messgeräte hinweisen und sollte daher erkannt werden. Formal basiert $\mathcal{Y}_i(t)$ auf $\Upsilon_i(t)$ aus Formel (5.11). Die regulären Daten folgen den Verteilungen $\Gamma(\Upsilon_{i,beo}(t)^2, \Upsilon_{i,beo}(t)^{-1})$ bzw. $\aleph(\Upsilon_{i,beo}(t), 1)$, die Ausreißer werden mit

einem AR(1)-Prozess kontaminiert. Aufgrund der diskretisierten Beobachtungen kann $\mathcal{Y}_i(t), t \in [0,1]$ als $\mathcal{Y}_{it}, t = 0, \ldots, T$ geschrieben werden, wobei hier $T = 100$ gelte. Die Ausreißer sind durch $\mathcal{Y}_{it} + Z_{it}$ definiert, wobei $Z_{it} = \varphi Z_{it-1} + \epsilon_{it}$ mit $\varphi = -0.9$ und $\epsilon_{it} \sim \aleph(0, 0.25)$. Die Ergebnisse sind in Tabelle 5.9 zusammengefasst.

### *5.3.2 Auswertung der Simulationsstudie*

Die Ergebnisse der in Abschnitt 5.3.1 geplanten Simulationsstudie wurden vorab in Kuhnt, Rehage, Becker-Emden, Tillmann und Hussong (2016) veröffentlicht. Folgende Erkenntnisse können daraus gezogen werden:

1. *Trend-Ausreißer*

   (a) Die wahre Verteilung der Zielgröße beeinflusst die RPR und RNR kaum.

   (b) Die RNR liegen in fast allen Fällen über den erwarteten Werten $1 - 0.01 = 0.99$ und $1 - 0.05 = 0.95$. Daher kann die Bootstrapprozedur konservativ genannt werden. Dies wird durch die Erkenntnisse von Sguera *et al.* (2016) bestätigt. Allerdings gibt es Ausnahmen für die HMD, vgl. (d).

   (c) Bei Annahme der Normalverteilung und inversem oder Log-Link lassen sich die besten Ergebnisse mit der MBD erzielen, unabhängig davon, welcher Verteilung die Zielgröße tatsächlich folgt. In allen anderen Fällen ist die Performanz der HMD am besten.

   (d) Die HMD tendiert dazu, zu viele Nicht-Ausreißer falsch zu klassifizieren. Für (i) angenommene gammaverteilte Zielvariable mit Identitätslink und (ii) angenommene normalverteilte Zielvariable mit inversem Link ist die RNR deutlich kleiner als (i, ii) $0.95$ und (ii) $0.99$. In solchen Fällen verhalten sich (i) FUNTA und (ii) MBD besser.

   (e) Für FUNTA ist die Wahl der GLM-Familie und der Linkfunktion entscheidend. Für die Kombinationen aus Gammaverteilung und Identitäts- und Log-Link sind die Resultate von FUNTA zufriedenstellend, auch wenn

Tabelle 5.8: Ergebnisse der Simulationsstudie mit Trendausreißern und $\phi_{k+1} = 2$

| Typ | Gewählte GLM-Familie | $g(\mu)$ | $p$ | MBD RPR | MBD RNR | HMD RPR | HMD RNR | FUNTA RPR | FUNTA RNR | rFUNTA RPR | rFUNTA RNR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Gamma | $\mu$ | 1 | 0.135 | 1.000 | 0.818 | 0.991 | 0.150 | 1.000 | 0.125 | 1.000 |
| | | | 5 | 0.775 | 0.999 | 0.985 | 0.930 | 0.807 | 0.999 | 0.587 | 0.998 |
| | | $\mu^{-1}$ | 1 | 0.135 | 1.000 | 0.240 | 1.000 | 0.000 | 0.999 | 0.005 | 0.994 |
| | | | 5 | 0.800 | 0.999 | 0.913 | 0.997 | 0.025 | 0.979 | 0.045 | 0.968 |
| | | $\log(\mu)$ | 1 | 0.125 | 1.000 | 0.357 | 1.000 | 0.087 | 1.000 | 0.058 | 0.997 |
| | | | 5 | 0.815 | 0.999 | 1.000 | 0.991 | 0.547 | 0.997 | 0.285 | 0.982 |
| | Normal | $\mu$ | 1 | 0.152 | 1.000 | 0.185 | 1.000 | 0.007 | 0.999 | 0.000 | 0.999 |
| | | | 5 | 0.817 | 0.999 | 0.930 | 0.999 | 0.068 | 0.980 | 0.010 | 0.986 |
| | | $\mu^{-1}$ | 1 | 0.305 | 1.000 | 0.020 | 0.934 | 0.002 | 0.987 | 0.002 | 0.984 |
| | | | 5 | 0.815 | 0.997 | 0.030 | 0.867 | 0.002 | 0.963 | 0.010 | 0.947 |
| | | $\log(\mu)$ | 1 | 0.195 | 1.000 | 0.018 | 0.991 | 0.000 | 0.998 | 0.002 | 0.990 |
| | | | 5 | 0.835 | 0.999 | 0.235 | 0.960 | 0.000 | 0.975 | 0.010 | 0.959 |
| 2 | Gamma | $\mu$ | 1 | 0.135 | 1.000 | 0.818 | 0.991 | 0.152 | 1.000 | 0.130 | 1.000 |
| | | | 5 | 0.775 | 0.999 | 0.985 | 0.931 | 0.807 | 1.000 | 0.590 | 0.998 |
| | | $\mu^{-1}$ | 1 | 0.143 | 1.000 | 0.230 | 1.000 | 0.000 | 0.999 | 0.005 | 0.994 |
| | | | 5 | 0.802 | 0.999 | 0.915 | 0.997 | 0.028 | 0.980 | 0.042 | 0.968 |
| | | $\log(\mu)$ | 1 | 0.120 | 1.000 | 0.365 | 1.000 | 0.075 | 1.000 | 0.065 | 0.997 |
| | | | 5 | 0.815 | 0.999 | 1.000 | 0.991 | 0.545 | 0.997 | 0.285 | 0.982 |
| | Normal | $\mu$ | 1 | 0.155 | 1.000 | 0.175 | 1.000 | 0.002 | 0.999 | 0.000 | 0.998 |
| | | | 5 | 0.820 | 0.999 | 0.930 | 0.999 | 0.068 | 0.979 | 0.010 | 0.986 |
| | | $\mu^{-1}$ | 1 | 0.305 | 1.000 | 0.022 | 0.934 | 0.002 | 0.987 | 0.002 | 0.985 |
| | | | 5 | 0.807 | 0.997 | 0.030 | 0.867 | 0.002 | 0.963 | 0.010 | 0.947 |
| | | $\log(\mu)$ | 1 | 0.205 | 1.000 | 0.018 | 0.991 | 0.000 | 0.998 | 0.002 | 0.990 |
| | | | 5 | 0.833 | 0.999 | 0.237 | 0.961 | 0.000 | 0.975 | 0.010 | 0.960 |

$\phi_{k+1}$ größer gewählt ist. Jede andere Kombination führt zu einer schlechten RPR.

(f) In allen Fällen ist die Performanz von FUNTA mindestens genauso gut wie die von rFUNTA.

(g) FUNTA und rFUNTA haben eine schlechte RPR für $p = 1$.

(h) Für größer werdendes $\phi_{k+1}$ (nicht abgebildet in Tabelle 5.8) steigt die RPR in den meisten Fällen. Die Ausnahme sind manche Fälle, in denen die RPR auch für $\phi_{k+1} = 2$ schon klein war. Die RNR ist in den meisten Fällen ähnlich oder nur unwesentlich größer.

Tabelle 5.9: Ergebnisse der Simulationsstudie mit Formausreißern mit $\varphi = -0.9$

| Typ | Gewählte GLM-Familie | $g(\mu)$ | $p$ | MBD RPR | MBD RNR | HMD RPR | HMD RNR | FUNTA RPR | FUNTA RNR | rFUNTA RPR | rFUNTA RNR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Gamma | $\mu$ | 1 | 0.158 | 1.000 | 0.232 | 0.983 | 0.045 | 0.999 | 0.148 | 1.000 |
| | | | 5 | 0.323 | 0.989 | 0.390 | 0.937 | 0.198 | 0.993 | 0.460 | 0.992 |
| | | $\mu^{-1}$ | 1 | 0.145 | 1.000 | 0.163 | 0.999 | 0.135 | 1.000 | 0.112 | 0.999 |
| | | | 5 | 0.312 | 0.986 | 0.350 | 0.979 | 0.568 | 0.997 | 0.410 | 0.985 |
| | | $\log(\mu)$ | 1 | 0.150 | 1.000 | 0.210 | 0.999 | 0.138 | 1.000 | 0.170 | 0.999 |
| | | | 5 | 0.315 | 0.988 | 0.380 | 0.973 | 0.590 | 0.997 | 0.463 | 0.989 |
| | Normal | $\mu$ | 1 | 0.150 | 1.000 | 0.172 | 1.000 | 0.058 | 0.999 | 0.148 | 1.000 |
| | | | 5 | 0.320 | 0.989 | 0.362 | 0.987 | 0.305 | 0.989 | 0.625 | 0.998 |
| | | $\mu^{-1}$ | 1 | 0.147 | 1.000 | 0.105 | 0.944 | 0.040 | 0.990 | 0.015 | 0.984 |
| | | | 5 | 0.302 | 0.986 | 0.222 | 0.881 | 0.258 | 0.975 | 0.075 | 0.950 |
| | | $\log(\mu)$ | 1 | 0.147 | 1.000 | 0.145 | 0.996 | 0.120 | 1.000 | 0.082 | 0.995 |
| | | | 5 | 0.320 | 0.988 | 0.300 | 0.964 | 0.557 | 0.996 | 0.330 | 0.974 |
| 2 | Gamma | $\mu$ | 1 | 0.152 | 1.000 | 0.162 | 0.963 | 0.035 | 0.998 | 0.108 | 1.000 |
| | | | 5 | 0.320 | 0.987 | 0.302 | 0.909 | 0.172 | 0.992 | 0.330 | 0.992 |
| | | $\mu^{-1}$ | 1 | 0.130 | 0.999 | 0.172 | 0.999 | 0.117 | 1.000 | 0.117 | 1.000 |
| | | | 5 | 0.307 | 0.988 | 0.365 | 0.980 | 0.573 | 0.998 | 0.407 | 0.985 |
| | | $\log(\mu)$ | 1 | 0.143 | 0.999 | 0.172 | 0.997 | 0.115 | 1.000 | 0.117 | 0.997 |
| | | | 5 | 0.315 | 0.989 | 0.338 | 0.964 | 0.583 | 0.997 | 0.365 | 0.980 |
| | Normal | $\mu$ | 1 | 0.143 | 1.000 | 0.163 | 1.000 | 0.065 | 0.999 | 0.152 | 1.000 |
| | | | 5 | 0.315 | 0.989 | 0.375 | 0.986 | 0.323 | 0.989 | 0.600 | 0.998 |
| | | $\mu^{-1}$ | 1 | 0.130 | 1.000 | 0.145 | 0.970 | 0.098 | 0.997 | 0.055 | 0.989 |
| | | | 5 | 0.297 | 0.989 | 0.305 | 0.920 | 0.400 | 0.985 | 0.160 | 0.960 |
| | | $\log(\mu)$ | 1 | 0.135 | 0.999 | 0.185 | 0.998 | 0.140 | 1.000 | 0.115 | 0.998 |
| | | | 5 | 0.315 | 0.988 | 0.368 | 0.973 | 0.568 | 0.997 | 0.422 | 0.983 |

2. *Form-Ausreißer:*

   (a) Die wahre Verteilung der Zielvariable beeinflusst RPR und RNR ein bisschen stärker als im vorigen Szenario, besonders im Falle der HMD und rFUNTA.

   (b) Die RNR ist wieder in fast allen Fällen über ihrem erwarteten Wert, mit einigen Ausnahmen bezüglich der HMD, siehe (c).

   (c) Die größte RPR wird für $p = 1$ meistens bei der HMD gefunden. Allerdings ist sie für angenommene Gammaverteilung und Identitätslink sowie angenommene Normalverteilung und inversem Link kleiner als $99\%$.

Tabelle 5.10: Ergebnisse der Simulationsstudie mit Formausreißern mit $\varphi = -0.95$

| Typ | Gewählte GLM-Familie | $g(\mu)$ | $p$ | MBD RPR | MBD RNR | HMD RPR | HMD RNR | FUNTA RPR | FUNTA RNR | rFUNTA RPR | rFUNTA RNR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Gamma | $\mu$ | 1 | 0.310 | 1.000 | 0.585 | 0.990 | 0.170 | 1.000 | 0.130 | 1.000 |
| | | | 5 | 0.685 | 0.997 | 0.730 | 0.952 | 0.453 | 0.998 | 0.562 | 0.999 |
| | | $\mu^{-1}$ | 1 | 0.258 | 1.000 | 0.370 | 1.000 | 0.185 | 1.000 | 0.193 | 1.000 |
| | | | 5 | 0.675 | 0.997 | 0.730 | 0.986 | 0.775 | 1.000 | 0.650 | 0.995 |
| | | $\log(\mu)$ | 1 | 0.270 | 1.000 | 0.432 | 1.000 | 0.198 | 1.000 | 0.208 | 1.000 |
| | | | 5 | 0.680 | 0.997 | 0.785 | 0.976 | 0.797 | 1.000 | 0.738 | 0.996 |
| | Normal | $\mu$ | 1 | 0.270 | 1.000 | 0.362 | 1.000 | 0.122 | 1.000 | 0.193 | 1.000 |
| | | | 5 | 0.693 | 0.998 | 0.777 | 0.991 | 0.558 | 0.995 | 0.877 | 1.000 |
| | | $\mu^{-1}$ | 1 | 0.323 | 1.000 | 0.405 | 0.952 | 0.158 | 0.995 | 0.103 | 0.990 |
| | | | 5 | 0.675 | 0.997 | 0.565 | 0.896 | 0.500 | 0.986 | 0.275 | 0.967 |
| | | $\log(\mu)$ | 1 | 0.273 | 1.000 | 0.405 | 0.999 | 0.203 | 1.000 | 0.193 | 0.999 |
| | | | 5 | 0.667 | 0.998 | 0.703 | 0.968 | 0.780 | 1.000 | 0.613 | 0.986 |
| 2 | Gamma | $\mu$ | 1 | 0.333 | 1.000 | 0.498 | 0.969 | 0.165 | 0.999 | 0.117 | 1.000 |
| | | | 5 | 0.690 | 0.996 | 0.650 | 0.924 | 0.410 | 0.997 | 0.460 | 0.997 |
| | | $\mu^{-1}$ | 1 | 0.253 | 1.000 | 0.375 | 1.000 | 0.175 | 1.000 | 0.165 | 1.000 |
| | | | 5 | 0.677 | 0.997 | 0.745 | 0.988 | 0.760 | 1.000 | 0.645 | 0.995 |
| | | $\log(\mu)$ | 1 | 0.258 | 1.000 | 0.415 | 0.999 | 0.190 | 1.000 | 0.193 | 0.999 |
| | | | 5 | 0.682 | 0.997 | 0.735 | 0.971 | 0.777 | 1.000 | 0.667 | 0.991 |
| | Normal | $\mu$ | 1 | 0.253 | 1.000 | 0.388 | 1.000 | 0.133 | 1.000 | 0.230 | 1.000 |
| | | | 5 | 0.690 | 0.997 | 0.790 | 0.992 | 0.583 | 0.995 | 0.880 | 1.000 |
| | | $\mu^{-1}$ | 1 | 0.287 | 1.000 | 0.458 | 0.972 | 0.180 | 0.998 | 0.177 | 0.995 |
| | | | 5 | 0.672 | 0.997 | 0.607 | 0.931 | 0.577 | 0.994 | 0.367 | 0.976 |
| | | $\log(\mu)$ | 1 | 0.268 | 1.000 | 0.410 | 1.000 | 0.208 | 1.000 | 0.193 | 1.000 |
| | | | 5 | 0.682 | 0.997 | 0.770 | 0.975 | 0.790 | 1.000 | 0.695 | 0.995 |

In diesen Fällen ist die MBD vorzuziehen, da ihre RNR größer ist und ihre RPR ähnlich groß.

(d) rFUNTA ist die beste Wahl für die Wahl des Identitätslinks und $p = 5$. Für andere Linkfunktionen hat FUNTA – abgesehen von einer Situation – die höchste RPR.

(e) Die Variation bezüglich der Wahl von Verteilung und Linkfunktion ist bei MBD und HMD deutlich kleiner als bei FUNTA und rFUNTA.

(f) Wird $\varphi = -0.95$ (siehe Tab. 5.10) gewählt, bestätigen sich viele der Punkte (a) bis (e). Alle Richtig-Positiv-Raten sind höher als für $\varphi = -0.9$. Es

Tabelle 5.11: Ergebnisse der Simulationsstudie mit Formausreißern mit $\varphi = -0.8$

| Typ | Gewählte GLM-Familie | $g(\mu)$ | $p$ | MBD RPR | MBD RNR | HMD RPR | HMD RNR | FUNTA RPR | FUNTA RNR | rFUNTA RPR | rFUNTA RNR |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Gamma | $\mu$ | 1 | 0.010 | 0.998 | 0.010 | 0.976 | 0.000 | 0.998 | 0.018 | 0.997 |
| | | | 5 | 0.010 | 0.977 | 0.028 | 0.928 | 0.010 | 0.984 | 0.082 | 0.981 |
| | | $\mu^{-1}$ | 1 | 0.007 | 0.998 | 0.003 | 0.997 | 0.018 | 1.000 | 0.020 | 0.995 |
| | | | 5 | 0.015 | 0.974 | 0.010 | 0.968 | 0.150 | 0.984 | 0.092 | 0.970 |
| | | $\log(\mu)$ | 1 | 0.007 | 0.998 | 0.003 | 0.995 | 0.022 | 1.000 | 0.020 | 0.995 |
| | | | 5 | 0.013 | 0.976 | 0.015 | 0.960 | 0.163 | 0.987 | 0.128 | 0.975 |
| | Normal | $\mu$ | 1 | 0.010 | 0.999 | 0.003 | 0.998 | 0.000 | 0.999 | 0.037 | 0.999 |
| | | | 5 | 0.013 | 0.976 | 0.010 | 0.976 | 0.040 | 0.976 | 0.150 | 0.989 |
| | | $\mu^{-1}$ | 1 | 0.007 | 0.998 | 0.003 | 0.939 | 0.000 | 0.987 | 0.000 | 0.984 |
| | | | 5 | 0.015 | 0.974 | 0.010 | 0.869 | 0.017 | 0.964 | 0.005 | 0.944 |
| | | $\log(\mu)$ | 1 | 0.010 | 0.998 | 0.000 | 0.990 | 0.003 | 1.000 | 0.005 | 0.989 |
| | | | 5 | 0.013 | 0.976 | 0.003 | 0.952 | 0.095 | 0.982 | 0.043 | 0.960 |
| 2 | Gamma | $\mu$ | 1 | 0.007 | 0.998 | 0.005 | 0.958 | 0.000 | 0.997 | 0.003 | 0.997 |
| | | | 5 | 0.015 | 0.975 | 0.015 | 0.900 | 0.010 | 0.987 | 0.015 | 0.980 |
| | | $\mu^{-1}$ | 1 | 0.005 | 0.997 | 0.003 | 0.997 | 0.013 | 1.000 | 0.013 | 0.994 |
| | | | 5 | 0.018 | 0.975 | 0.013 | 0.969 | 0.155 | 0.988 | 0.100 | 0.972 |
| | | $\log(\mu)$ | 1 | 0.010 | 0.998 | 0.000 | 0.992 | 0.010 | 0.999 | 0.007 | 0.993 |
| | | | 5 | 0.015 | 0.976 | 0.007 | 0.952 | 0.115 | 0.986 | 0.062 | 0.963 |
| | Normal | $\mu$ | 1 | 0.005 | 0.998 | 0.003 | 0.999 | 0.003 | 0.999 | 0.025 | 0.999 |
| | | | 5 | 0.015 | 0.977 | 0.010 | 0.974 | 0.050 | 0.978 | 0.148 | 0.989 |
| | | $\mu^{-1}$ | 1 | 0.005 | 0.998 | 0.010 | 0.963 | 0.005 | 0.994 | 0.005 | 0.986 |
| | | | 5 | 0.013 | 0.976 | 0.025 | 0.905 | 0.068 | 0.971 | 0.025 | 0.949 |
| | | $\log(\mu)$ | 1 | 0.007 | 0.998 | 0.000 | 0.995 | 0.020 | 1.000 | 0.010 | 0.994 |
| | | | 5 | 0.015 | 0.976 | 0.015 | 0.959 | 0.117 | 0.983 | 0.077 | 0.966 |

ist festzuhalten, dass die HMD verglichen mit den anderen Datentiefen etwas besser ist als zuvor und rFUNTA etwas schlechter. Allerdings resultieren wiederum einige Kombinationen von GLM-Familie und Linkfunktion zu unerwartet niedrigen RNR-Werten der HMD.

(g) Falls die Ausreißer weniger extrem sind als zuvor, also $\varphi = -0.8$, siehe Tabelle 5.11, ist die Performanz aller Datentiefen deutlich schlechter als für $\varphi = -0.9$. Besonders die HMD findet fast keine tatsächlichen Ausreißer mehr, wobei bei gewissen Kombinationen aus GLM-Familie und Linkfunktion zu viele reguläre Beobachtungen als falsche Ausreißer klassifiziert werden. MBD ist nur für wenige Kombinationen aus GLM-Familie

und Linkfunktion mit $p = 1$ die beste Wahl. Allgemein lassen sich für FUNTA und rFUNTA die akzeptabelsten Ergebnisse finden.

Anhand der Tabellen 5.8, 5.9, 5.10 und 5.11 ist es schwierig, situationsübergreifende Erkenntnisse über die Ausreißeridentifikation in GFOSR festzuhalten. Ein interessanter Fakt betrifft jedoch die Rolle der Linkfunktionen: Sie haben bei der MBD keinen relevanten Einfluss auf die RPR, bei der HMD einen recht kleinen Einfluss auf die RPR und bei FUNTA und rFUNTA einen entscheidenden Einfluss auf die RPR. Des Weiteren ist FUNTA bei Formausreißern und $p = 5$ in den meisten Fällen die beste Wahl. Bei Trendausreißern ist MBD für $p = 5$ zwar nicht immer die beste Wahl, hat aber hohe, sehr stabile Richtig-Positiv-Raten und fast perfekte Richtig-Negativ-Raten und wird somit empfohlen.

## 5.4 $\alpha$-AUSREISSER IN GAUSSPROZESSEN

Bisher wurden in Kapitel 5 Verfahren vorgestellt und verglichen, um Ausreißer in funktionalen Beobachtungen zu identifizieren. Dabei wurden etwaige Informationen über den datengenerierenden Prozess nicht verwendet. Falls dieser jedoch bekannt ist, wird durch die Verwendung von Datentiefen und Resamplingverfahren die verfügbare Information nicht komplett ausgeschöpft. In der Praxis ist der Gaußprozess (siehe Def. 3.11) als unendlichdimensionale Verallgemeinerung der Normalverteilung eine häufige Annahme.

Eine exakte Übertragbarkeit des Konzepts der $\alpha$-Ausreißer ist nicht möglich. Hierfür müsste das Konzept der Wahrscheinlichkeitsdichte auch im Unendlichdimensionalen existieren. Delaigle und Hall (2010) zeigen im Anhang ihres Artikels, dass die Verallgemeinerung einer multivariaten Wahrscheinlichkeitsdichte

$$f(\mathbf{x}) = \lim_{h \searrow 0} (h^r v_r)^{-1} P(\|\boldsymbol{\mathcal{X}} - \mathbf{x}\| \leq h), \qquad \mathbf{x}, \boldsymbol{\mathcal{X}} \in \mathbb{R}^r$$

mit der euklidischen Distanz $\|\cdot\|$ und $v_r$ dem Volumen der $r$-dimensionalen Ein-

heitssphäre auf unendlichdimensionale Räume

$$f(x) = \lim_{h \searrow 0} (\kappa_1(h))^{-1} P(\|\mathcal{X} - x\| \leq h), \qquad x, \mathcal{X} \in \mathbb{B}$$

mit einer geeigneten Funktion $\kappa_1(h)$ und der $L_2$-Norm $\|\cdot\|$ in einen Widerspruch mündet, da kein $\kappa_1(h)$ existiert, so dass $f(x)$ wohldefiniert ist.

Als Alternative schlagen Delaigle und Hall Log-Dichten basierend auf funktionaler Hauptkomponentenanalyse (vgl. Anhang A.2) vor. Mit den Schätzern $\hat{\theta}, \hat{\psi}$ der Eigenwerte $\theta$ und Eigenfunktionen $\psi$ wird die Log-Dichte der gewichteten PC Scores $\xi_j^w = \theta_j^{-1/2} \int \mathcal{X}(t)\psi_j(t)dt$ bestimmt. Die Dichten der Scores können mit Hilfe von Kerndichteschätzern geschätzt werden. Ausreißerregionen einzelner Hauptkomponenten wurden in das R-Paket `alphaOutlier` durch den Befehl `aout.kernel` integriert (siehe Kap. 2.2 sowie Rehage und Kuhnt, 2016). Bei funktionaler Hauptkomponentenanalyse existieren unendlich viele Hauptkomponenten. Die Anzahl der Hauptkomponenten ist in der Praxis aber durch die gewählte Anzahl der Basisfunktionen der Beobachtungen beschränkt. Wie im multivariaten Fall kann der Anteil der zu erklärenden kumulierten Varianz an der Gesamtvarianz festgelegt und entsprechend viele Hauptkomponenten gewählt werden. Die Scores der wichtigsten Hauptkomponenten lassen sich dann auf $\alpha$-Ausreißer untersuchen.

Statt der Kerndichteschätzung lassen sich auch Verteilungsannahmen treffen: Scores eines Gaußprozesses sind normalverteilt, geschätzte Scores sind asymptotisch normalverteilt (Aston *et al.*, 2010). Die Parameter dieser Verteilung lassen sich nicht vollständig herleiten. Während die Eigenfunktionen von Kovarianzkernen deterministisch sind (Rasmussen und Williams, 2006, Kap. 4.3), sind die Eigenfunktionen eines Gaußprozesses stochastisch.

Da Gaußprozesse vollständig durch ihre ersten beiden Momente charakterisiert sind und alle endlichen Teilmengen des Gaußprozesses einer gemeinsamen multivariaten Normalverteilung folgen, lässt sich der „Dichtewert“ einer Realisation des Gaußprozesses approximieren.

KOROLLAR 5.28. *Sei $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$ ein Gaußprozess mit Indexmenge $\mathcal{T} = [a, b]$. Sei zudem angenommen, dass $\mathbf{y} = (y_{t_1}, \ldots, y_{t_T})^\top =: (y_\mathbf{t})^\top$ eine diskretisierte Realisation*

*dieses Gaußprozesses mit $t_1, \ldots, t_T \in \mathcal{T}$ ist. Dann gilt für $(\mathcal{Y}(t_1), \ldots, \mathcal{Y}(t_T))^\top =: \boldsymbol{\mathcal{Y}} \sim \aleph(\mu(\mathbf{t}), \sigma(\mathbf{t}, \mathbf{t})) =: P_{\boldsymbol{\theta}}$.*

*Beweis:* Folgt direkt aus der Definition des Gaußprozesses, vgl. Cuevas (2014). □

Falls $\mathbf{y} \in \mathrm{out}(\alpha, P_{\boldsymbol{\theta}})$ mit $P_{\boldsymbol{\theta}}$ wie in Korollar 5.28 gilt, wird $\mathbf{y}$ als $\alpha$-Ausreißer der projektiven Familie bezeichnet. Die Bestimmung des Grenzwerts $B(\alpha)$ in (2.1) erfolgt mit dem $(1-\alpha)$-Quantil der $\chi^2_T$-Verteilung. Die Angabe einer $\alpha$-Ausreißerregion in geschlossener Form ist für solche Prozesse nicht möglich. Je größer $T$ ist, desto genauer kann das Konzept der $\alpha$-Ausreißer übertragen werden. Allerdings ist die Konditionszahl der Kovarianzmatrix $\sigma(\mathbf{t}, \mathbf{t})$ steigend in $T$, was zur numerischen Singularität führt. Das Verfahren kann folglich im Allgemeinen nicht beliebig genau werden. Bei gewissen Kovarianzkernen (siehe Bem. 5.29) greifen Algorithmen zur Bestimmung der Inversen einer Toeplitz-Matrix (McLeod *et al.*, 2007).

BEMERKUNG 5.29. *Sei $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$ mit stationärer Kovarianzfunktion $\sigma(s,t) = f(|s-t|)$. Dann gilt: $\sigma(\mathbf{t}, \mathbf{t})$ ist für $t_i - t_{i-1} = c \,\forall\, t \in \{2, \ldots, T\}$ eine Toeplitz-Matrix mit den Einträgen*

$$\sigma(\mathbf{t}, \mathbf{t}) = \begin{pmatrix} f(0) & f(c) & f(2c) & \ldots & f((T-1)c) \\ f(c) & f(0) & f(c) & \ldots & f((T-2)c) \\ f(2c) & f(c) & f(0) & \ldots & f((T-3)c) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ f((T-1)c) & f((T-2)c) & f((T-3)c) & \ldots & f(0) \end{pmatrix}.$$

Somit kann beispielsweise bei quadriert-exponentieller Kovarianzfunktion mit äquidistanten Zeitpunkten der Dichtewert einer Beobachtung mit Hilfe der Inversen der Toeplitz-Matrix $\sigma(\mathbf{t}, \mathbf{t})$ bestimmt werden. Dies gilt allerdings nicht uneingeschränkt, da die Kovarianzmatrix $\sigma(\mathbf{t}, \mathbf{t})$ nur bei gewissen Kombinationen aus $\mathbf{t}$ und $\ell$ positiv definit und somit invertierbar ist. Während der Hyperparameter $\ell$ fest mit dem Prozess zusammenhängt, kann in der Auswertung der Daten die Dichtheit von $\mathbf{t}$ künstlich verringert werden. Statt an jedem Zeitpunkt wird die Beobachtung und damit auch der Erwartungswertvektor und die Kovarianzmatrix nur an jedem $k$-ten Eintrag von $\mathbf{t}$ bestimmt, kurz $\mathbf{t}^k$. Um so wenig Informationen wie möglich zu

verlieren, wird das kleinste $k$ ausgewählt, für das $\sigma(\mathbf{t}^k, \mathbf{t}^k)$ invertierbar ist. In Korollar 5.28 ist das äquivalent mit einer anderen Wahl von $\mathbf{t}$, wodurch die Annahme der Normalverteilung erhalten bleibt.

### *5.4.1 Lineare Funktionale von Gaußprozessen*

Die Beschränkung auf die Scores „wichtiger" Hauptkomponenten oder auf eine projektive Familie des Gaußprozesses beinhaltet eine gewisse Willkür. In diesem Abschnitt soll ein systematischerer Ansatz zur Erkennung von $\alpha$-Ausreißern in Gaußprozessen entwickelt werden.

SATZ 5.30. *Sei $\mathcal{Y} = \{\mathcal{Y}(t) : t \in \mathcal{T}\}$ ein Gaußprozess mit $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$, wobei $\mu = \{\mu(t) : t \in \mathcal{T}\}$ und die positiv definite Kovarianzfunktion durch $\sigma = \{\sigma(s,t) : s,t \in \mathcal{T}\}$ definiert ist. Dann gilt: Das lineare Funktional $\int_{\mathcal{T}} \mathcal{Y}(t)dt$ dieses Gaußprozesses ist normalverteilt mit Erwartungswert $\int_{\mathcal{T}} \mu(t)dt$ und Varianz $\int_{\mathcal{T}} \int_{\mathcal{T}} \sigma(s,t)ds\, dt$.*

*Beweis:* Siehe Barrett und Myers (2004, S. 410 f.). □

Mit Hilfe von Satz 5.30 kann die $\alpha$-Ausreißerregion des linearen Funktionals eines Gaußprozesses bestimmt werden, siehe Beispiel 5.31.

BEISPIEL 5.31. *Sei $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$, wobei $\mu(t) \equiv 0$, $t \in \mathcal{T} = [0,1]$ gelte. Als Kovarianzfunktion werde der* quadriert-exponentielle Kovarianzkern $\sigma_{QE}(s,t) = \exp\left(-\frac{(s-t)^2}{2\ell^2}\right)$ *gewählt, wobei $\ell = 1$ gelte. Dann gilt: $\mathsf{E}(\int_{\mathcal{T}} \mathcal{Y}(t)dt) = 0$. Durch zweimaliges Integrieren der Kovarianzfunktion folgt außerdem:*

$$\begin{aligned}
\mathsf{Cov}\left(\int_{\mathcal{T}} \mathcal{Y}(t)dt\right) &= \int_{\mathcal{T}} \int_{\mathcal{T}} \sigma_{QE}(s,t)ds\, dt \\
&= \int_0^1 \left[\sqrt{\frac{\pi}{2}}\,\mathrm{erf}\left(\frac{s-t}{\sqrt{2}}\right)\right]_0^1 dt \\
&= \sqrt{\frac{\pi}{2}} \int_0^1 \left(\mathrm{erf}\left(\frac{1-t}{\sqrt{2}}\right) - \mathrm{erf}\left(-\frac{t}{\sqrt{2}}\right)\right) dt \\
&= \sqrt{2\pi}(2\Phi(1) - 1) - 2 + \frac{2}{\sqrt{e}} =: \sigma_{QE}^2 \approx 0.924.
\end{aligned}$$

*Die $\alpha$-Ausreißerregion von $\int_{\mathcal{T}} \mathcal{Y}(t)dt$ stimmt mit* $\mathrm{out}(\alpha, \aleph(0, \sigma_{QE}^2))$ *überein.*

In der Regel kann das Doppelintegral der Kovarianzfunktion nur aufwändig oder gar nicht analytisch hergeleitet werden. Eine numerische Bestimmung ist jedoch möglich und bietet anders als das in Korollar 5.28 beschriebene Vorgehen eine beliebig große Genauigkeit, ohne dass singuläre Kovarianzmatrizen auftreten, z. B. in R mit der Funktion `adaptIntegrate` (R-Paket `cubature`, Johnson und Narasimhan, 2013). Auch ohne Kenntnis der Kovarianzfunktion ist eine Approximation der Verteilung des linearen Funktionals möglich: In Abbildung 5.11 wird aufgezeigt, wie genau die numerische Varianzschätzung bei unterschiedlichen Kombinationen von $n$ (Anzahl Beobachtungen) und $T$ (Anzahl diskretisierter Zeitpunkte) ist. Dabei wurden je 100 Stichproben des in Beispiel 5.31 genannten Gaußprozesses gezogen und das Doppelintegral der Kovarianzfunktion numerisch approximiert. Zunächst fällt auf, dass die Präzision der Varianzschätzung mit steigendem $n$ deutlich größer wird. Andererseits scheint die Anzahl der diskretisierten Zeitpunkte keinen großen Einfluss zu haben. Vor allem bei $n = 30$ und $n = 100$ wird die Streuung der Schätzung mit steigendem $T$ sogar eher größer als kleiner. Da die Stichprobengröße $n$ in praktischen Situationen oft nicht „beliebig“ groß gewählt werden kann, wird als Empfehlung nur $T = 100$ gegeben, und zwar nur dann, wenn die Anzahl tatsächlich beobachteter Zeitpunkte mindestens 100 ist.

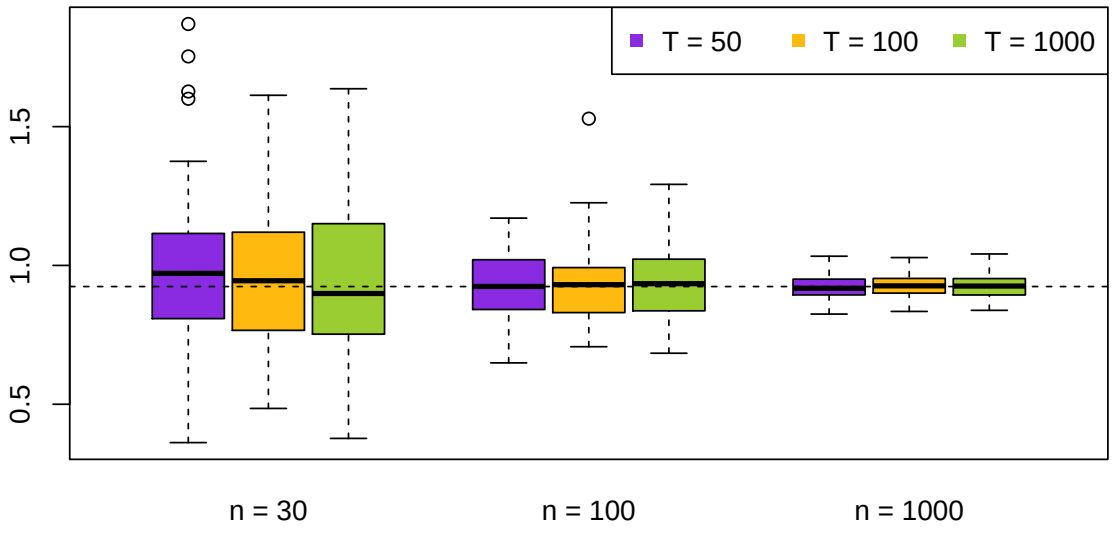



Abbildung 5.11: Numerische Schätzungen der Varianz bei verschiedenen Kombinationen von $n$ und $T$. Gestrichelte Linie: Wahrer Wert für $\sigma^2_{QE}$.

Für die Untersuchung von Ausmaß-Ausreißern ist die Betrachtung des linearen Funktionals $\int \mathcal{Y}(t)dt$ ausreichend. Form-Ausreißer mit einer unverdächtigen Lage werden mit diesem Verfahren jedoch eher nicht erkannt, siehe Beispiel 5.32.

BEISPIEL 5.32. *Betrachtet seien die Funktionen $y_1(t) = t$ und $y_2(t) = t^3, t \in [-1, 1]$. Wegen $\int_{-1}^{1} y_1(t)dt = \int_{-1}^{1} y_2(t)dt = 0$ existieren keine Unterschiede in Bezug auf die linearen Funktionale dieser Funktionen.*

Ein Ansatz, um die Form der funktionalen Beobachtungen besser abzubilden, betrifft die Einbeziehung der Ableitungen. Es gilt, dass die Verteilung von $\mathcal{Y}'$ unter gewissen Bedingungen (siehe Def. 5.13, Kor. 5.14 und Prop. 5.16) hergeleitet werden kann und dann unter Anwendung von Satz 5.30 auch $\int \mathcal{Y}'(t)dt$ eine normalverteilte Zufallsvariable ist. Dieser Ansatz wird nun auf Beispiel 5.32 angewendet.

BEISPIEL 5.33. *Fortsetzung von Beispiel 5.32*
*Für die Ableitungen gilt $y_1'(t) = 1, y_2'(t) = 3t^2, t \in [-1, 1]$. Wegen $\int_{-1}^{1} y_1'(t)dt = \int_{-1}^{1} y_2'(t)dt = 2$ existieren keine Unterschiede in Bezug auf die linearen Funktionale der Ableitungen dieser Funktionen.*

Da auch dieser Ansatz die Unterschiede der Formen von $y_1$ und $y_2$ nicht aufdecken kann, werden die Quadrate der Funktionen und ihrer Ableitungen betrachtet, um zu verhindern, dass sich Positiv- und Negativteile der Funktionen „ausgleichen“.

BEISPIEL 5.34. *Fortsetzung von Beispiel 5.32*
*Für die Quadrate der Funktionen und ihrer Ableitungen gilt $(y_1(t))^2 = t^2, (y_2(t))^2 = t^6, (y_1'(t))^2 = 1, (y_2'(t))^2 = 9t^4, t \in [-1, 1]$. Die linearen Funktionale dieser Werte sind $\int_{-1}^{1}(y_1(t))^2dt = \frac{2}{3}, \int_{-1}^{1}(y_2(t))^2dt = \frac{2}{7}, \int_{-1}^{1}(y_1'(t))^2dt = 2, \int_{-1}^{1}(y_2'(t))^2dt = \frac{18}{5}$. Damit liegt es nahe, die linearen Funktionale der Quadrate der Funktionen und ihrer Ableitungen zur Identifikation von Form-Ausreißern zu betrachten.*

Formal ist die Grundlage für dieses Vorgehen wieder ein Gaußprozess $\mathcal{Y} = \{\mathcal{Y}(t) \mid t \in \mathcal{T}\}$ mit $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$. Zur Vereinfachung sei für alle $t \in \mathcal{T}$ davon ausgegangen, dass $\mu(t) = 0$ und $\sigma(t, t) = 1$. Aus $\mathcal{Y}(t) \sim \aleph(0, 1)$ folgt $(\mathcal{Y}(t))^2 \sim \chi_1^2$. Dann ist $\mathcal{Y}^2 = \{(\mathcal{Y}(t))^2 \mid t \in \mathcal{T}\}$ ein $\chi^2$-Prozess (Adler, 1981, S. 168 f.) mit $n = 1$ Freiheitsgraden. Weiterhin gelte $\mathsf{E}(\mathcal{Y}^2(t)) = \sigma(t, t) = 1$ und $\mathsf{Var}(\mathcal{Y}^2(t)) = 2\sigma(t, t)^2 = 2$. Für das lineare Funktional eines $\chi^2$-Prozesses $\int_{\mathcal{T}} \mathcal{Y}^2(t)dt$ ist nicht klar, ob sich eine zu Satz 5.30

analoge Aussage herleiten lässt. Dasselbe gilt für die Verteilung von $\int_{\mathcal{T}}(\mathcal{Y}'(t))^2dt$ und damit auch für die bivariate Zufallsvariable $(\int_{\mathcal{T}}\mathcal{Y}^2(t)dt, \int_{\mathcal{T}}(\mathcal{Y}'(t))^2dt)^\top$. Als Approximation für eine bivariate $\alpha$-Ausreißerregion kann der Bagplot (Kap. 2.3) zur Ausreißeridentifikation verwendet werden. In (2.14) kann der Faktor 3 analog zu $\alpha$ bzw. $p$ angepasst werden, um die Ausreißerregion des Bagplots zu manipulieren.

### *5.4.2 Simulationsstudie*

In einer Simulationsstudie werden nun Ausreißeridentifizierer basierend auf den in Abschnitt 5.4 vorgestellten Approximationsmöglichkeiten für bekannte Gaußprozesse verglichen, wobei $\mathcal{Y} = \{\mathcal{Y}(t) \mid t \in \mathcal{T}\}$, $\mathcal{Y} \sim \mathcal{GP}(\mu, \sigma)$ mit $\mu = \{\mu(t) \mid t \in \mathcal{T}\}$ und $\sigma = \{\sigma(s,t) \mid s,t \in \mathcal{T}\}$ bekannt. Die folgenden Verfahren werden dabei verwendet:

1. Projektive Familie (PF): Bestimme den größtmöglichen Vektor $\mathbf{t} \in \mathbb{R}^d$, für den $\sigma(\mathbf{t},\mathbf{t})$ invertierbar ist. Bezeichne $\boldsymbol{\mathcal{Y}} := \mathcal{Y}(\mathbf{t})$. Aus Korollar 5.28 folgt $\boldsymbol{\mathcal{Y}} \sim \aleph(\mu(\mathbf{t}), \sigma(\mathbf{t},\mathbf{t}))$. Die Quantile der Standardisierung $(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mathcal{Y}} - \boldsymbol{\mu}) \sim \chi^2_d$, wobei $\boldsymbol{\mu} = \mu(\mathbf{t})$ und $\boldsymbol{\Sigma} = \sigma(\mathbf{t},\mathbf{t})$ gelte, werden für die Bestimmung der $\alpha$-Ausreißerregion genutzt.

2. Bagplot linearer Funktionale quadrierter Funktionen (BLFQF): Die linearen Funktionale der quadrierten Realisation und quadrierten ersten Ableitung werden bestimmt. Da unklar ist, ob die konkrete Verteilung dieser bivariaten Zufallsvariable hergeleitet werden kann, wird der Bagplot als Ausreißeridentifizierer verwendet. Der Faktor, der die Entfernung des *fence* vom *bag* bestimmt, dient als Tuningparameter. Damit die Ergebnisse dieses Ausreißeridentifizierers mit den anderen Verfahren vergleichbar sind, muss der Faktor so eingestellt werden, dass die Richtig-Negativ-Rate etwa bei $(100 - p)\%$ liegt.

3. Als Vergleichswerte dienen die aus Abschnitt 5.1 bekannten funktionalen Datentiefen in Verbindung mit dem geglätteten Bootstrap.

Damit die endlichdimensionale Verteilung gleich ist, wird für die regulären Realisationen die Erwartungsfunktion $\mu_0(t) = 0$ und der quadriert-exponentielle Kova-

rianzkern $\sigma_{QE}(s,t) = \exp\left(-\frac{(s-t)^2}{2}\right)$ gewählt. Fünf von 100 Beobachtungen werden als Ausreißer von einem der folgenden fünf Prozesse generiert:

Szenario 1: $\mathcal{Y} \sim \mathcal{GP}(\mu_0, \sigma_{Mat})$, wobei $\sigma_{Mat}(s,t) = 2|s-t|K_1(2|s-t|)$ die Matérn-Kovarianzfunktion ist, mit der modifizierten Besselfunktion zweiter Art $K_1$ und Ordnung $1$ (Abb. 5.12, oben links).

Szenario 2: $\mathcal{Y} \sim \mathcal{GP}(\mu_0, \sigma_{RQ})$, wobei $\sigma_{RQ}(s,t) = (1 + |s-t|^2)^{-2}$ die rationalquadratische Kovarianzfunktion ist (Abb. 5.12, oben rechts).

Szenario 3: $\mathcal{Y} \sim \mathcal{GP}(\mu_0, \sigma_{WN})$, wobei $\sigma_{WN}(s,t) = 0.01 \cdot \mathbf{1}_{\{t\}}(s)$ die *white noise* Kovarianzfunktion ist (Abb. 5.12, Mitte links).

Szenario 4: $\mathcal{Y} \sim \mathcal{GP}(\mu_0, \sigma_{Per})$, wobei $\sigma_{Per}(s,t) = \exp(-200\sin(|s-t|/2)^2)$ die periodische Kovarianzfunktion ist (Abb. 5.12, Mitte rechts).

Szenario 5: $\mathcal{Y} \sim \mathcal{GP}(\mu_0, \sigma_{QE})$, wobei fünf Beobachtungen mit Wahrscheinlichkeit $5\%$ pro Zeitpunkt kontaminiert sind, das heißt mit dem Faktor 3 multipliziert werden (Abb. 5.12, unten links).

Die wichtigsten Erkenntnisse der Simulationsstudie (Tabelle 5.12) lassen sich wie folgt zusammenfassen:

1. Die PF-Methode schafft es in den ersten vier Szenarien, alle Ausreißer zu identifizieren. Allerdings hält sie das erwartete Niveau konstant nicht ein, ist also zu liberal. Da sich die Richtig-Negativ-Rate für $\alpha = 0.01$ zwischen 0.95 und 0.99 befindet, lässt sich die Methode trotzdem mit den übrigen Verfahren vergleichen. Die Richtig-Positiv-Rate ist perfekt, so dass zu erwarten ist, dass die PF-Methode auch für noch kleineres $\alpha$ immer noch eine sehr gute RPR erzielt. Der Informationsverlust dadurch, dass nur jeder zehnte Zeitpunkt berücksichtigt wird, ist bei Ausreißern, die auf der kompletten Indexmenge auffällig sind, zu vernachlässigen. In der fünften Ausreißersituation ist das nicht der Fall, somit ist die PF-Methode hier nicht mehr perfekt im Sinne der Ausreißererkennung.

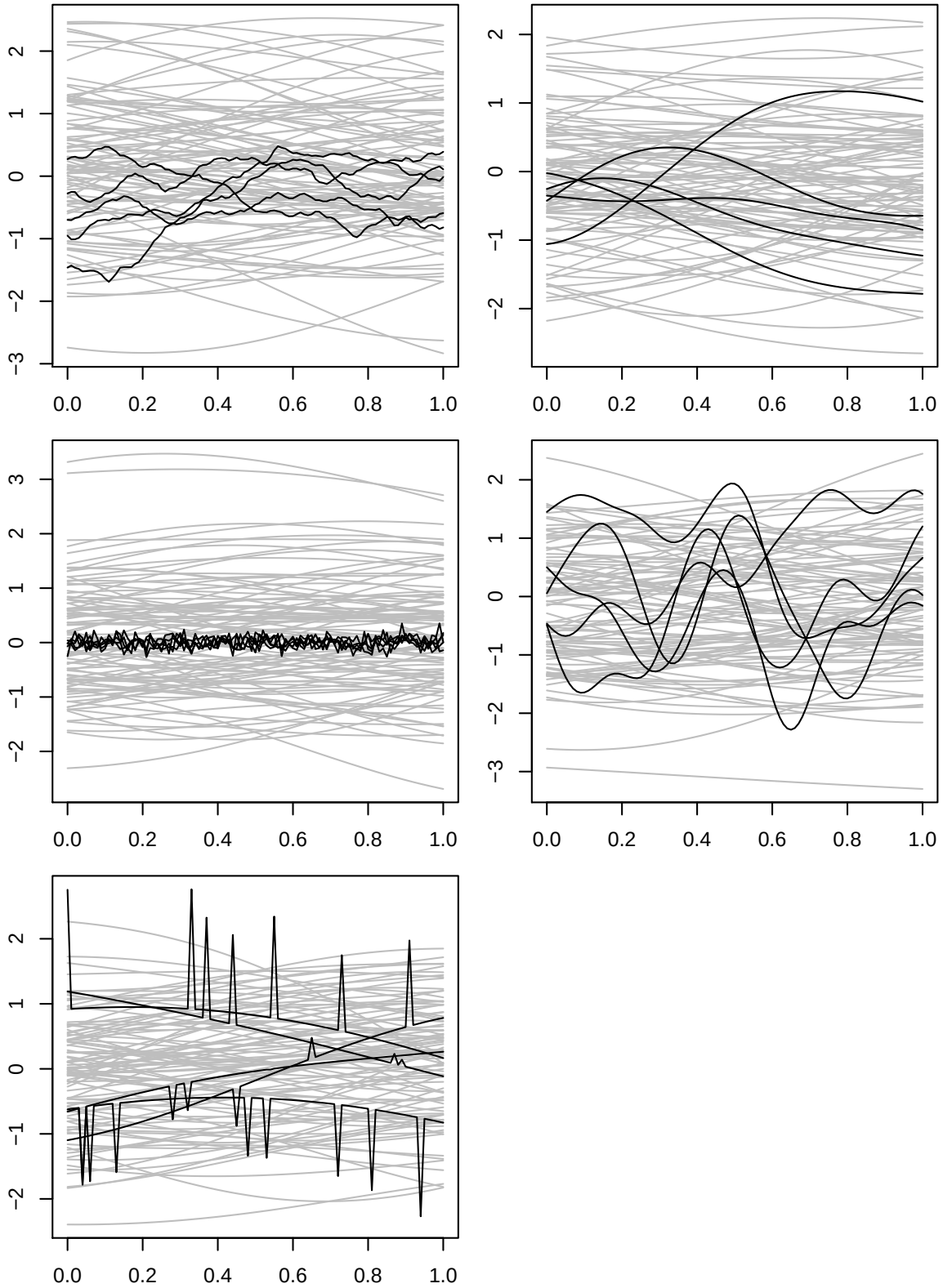



Abbildung 5.12: Beispielplots der Szenarien für die Bewertung der $\alpha$-Ausreißerverfahren bei bekannten Gaußprozessen. Reguläre Daten mit quadriert-exponentieller Kovarianzfunktion (grau) und Ausreißer (schwarz).

2. Das Problem der Bagplot-basierten BLFQF-Methode ist, dass kein explizites $\alpha$ bzw. $p$ eingegeben werden kann, was die Richtig-Negativ-Rate kontrolliert. Der Faktor, der den Abstand zwischen *bag* und *fence* definiert, muss manuell adjustiert werden, um vergleichbare RNR zu erhalten. Dies funktioniert nur in Situationen, in denen die Ausreißer bekannt sind, was die Anwendung der BLFQF-Methode in der Praxis unmöglich macht. Prinzipiell scheint das Verfahren aber vielversprechend zu sein. In drei der ersten vier Situationen hat die BLFQF-Methode RPR über $99\%$ bei einer RNR von $95\%$. Die andere Situation ist mit der rational-quadratischen Kovarianzfunktion die unauffälligste Ausreißersituation. Im letzten Szenario ist die Performanz von BLFQF bei Berücksichtigung der RNR besser als die der PF-Methode.

3. Wie zuvor zeigen sich für die FMD und MBD sehr ähnliche Resultate. Die beste Performanz zeigt sich für die modifizierte Bandtiefe im zentrierten Szenario 4 mit $32\%$ erkannten Ausreißern. In den übrigen Situationen ist die RPR zwischen $0\%$ und $28\%$. Die Performanz in den zentrierten Datensituationen ist bis auf Szenario 5 besser als in den unzentrierten.

4. Die HMD identifiziert deutlich mehr Ausreißer korrekt als MBD und FMD. Ihre Performanz ist bei vorheriger Zentrierung des Datensatzes klar besser. Bei einer RNR von $95\%$ erkennt sie mindestens $50\%$ der Ausreißer und ist mit $67.2\%$ der Ausreißer in Situation 5 sogar besser als die PF- und BLFQF-Methode. Die Ausnahme ist Szenario 3, in der wie bei FMD und MBD kein Ausreißer gefunden wird.

5. Die Performanz von FUNTA und rFUNTA in den fünf Szenarien ist sehr ähnlich. Bei $p = 1$ werden jeweils etwa $20\%$ der Ausreißer erkannt, bei $p = 5$ zwischen $47.4\%$ und $100\%$ der Ausreißer. Dabei ist die RNR bis auf einmal klar oberhalb des erwarteten Werts. FUNTA ist meist ähnlich gut wie rFUNTA, ansonsten etwas schlechter. In Szenario 5 ist für $p = 1$ FUNTA nach der $h$-modal Datentiefe das zweitbeste Verfahren, allerdings ist für $p = 5$ die RNR mit $59.6\%$ so niedrig, dass die Schätzung des Quantils durch den Bootstrap ungeeignet ist. Für $p = 5$ ist rFUNTA das beste Verfahren im Vergleich.

Tabelle 5.12: Simulationsergebnisse bei Ausnutzung der bekannten Parameter des Gaußprozesses. Für die PF-Methode gilt $p = \alpha$. Bei BLFQF ist $p$ nicht exakt anzugeben. Stattdessen: Für die Szenarien 1, 2 und 3 wird der Faktor als $3.5$ ($p = 1$) und $2.2$ ($p = 5$) gewählt, für Szenario 4 wird $11.5$ ($p = 1$) und $8.3$ ($p = 5$) gewählt und für Szenario 5 wird $100$ ($p = 1$) und $40.2$ ($p = 5$) gewählt.

| | | | PF | | BLFQF | | FMD | | HMD | | MBD | | FUNTA | | rFUNTA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Szenario | | $p$ | RPR | RNR | RPR | RNR | RPR | RNR | RPR | RNR | RPR | RNR | RPR | RNR | RPR | RNR |
| 1 | U | 1 | 1.000 | 0.980 | 0.918 | 0.987 | 0.002 | 0.985 | 0.006 | 0.990 | 0.000 | 0.985 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 1.000 | 0.920 | 0.994 | 0.951 | 0.036 | 0.934 | 0.100 | 0.950 | 0.030 | 0.938 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | ⁄ | ⁄ | ⁄ | ⁄ | 0.066 | 0.990 | 0.418 | 0.994 | 0.052 | 0.991 | 0.218 | 1.000 | 0.212 | 1.000 |
| | Z | 5 | ⁄ | ⁄ | ⁄ | ⁄ | 0.192 | 0.947 | 0.706 | 0.960 | 0.180 | 0.951 | 0.994 | 1.000 | 1.000 | 1.000 |
| 2 | U | 1 | 1.000 | 0.979 | 0.190 | 0.990 | 0.004 | 0.987 | 0.022 | 0.991 | 0.004 | 0.987 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 1.000 | 0.918 | 0.372 | 0.950 | 0.030 | 0.932 | 0.068 | 0.956 | 0.026 | 0.939 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | ⁄ | ⁄ | ⁄ | ⁄ | 0.126 | 0.992 | 0.242 | 0.995 | 0.112 | 0.993 | 0.178 | 0.999 | 0.184 | 0.999 |
| | Z | 5 | ⁄ | ⁄ | ⁄ | ⁄ | 0.238 | 0.952 | 0.496 | 0.961 | 0.242 | 0.956 | 0.474 | 0.978 | 0.550 | 0.981 |
| 3 | U | 1 | 1.000 | 0.978 | 0.898 | 0.988 | 0.000 | 0.987 | 0.000 | 0.991 | 0.000 | 0.987 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 1.000 | 0.918 | 0.996 | 0.950 | 0.000 | 0.933 | 0.000 | 0.953 | 0.000 | 0.938 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | ⁄ | ⁄ | ⁄ | ⁄ | 0.000 | 0.984 | 0.000 | 0.989 | 0.000 | 0.987 | 0.218 | 1.000 | 0.260 | 1.000 |
| | Z | 5 | ⁄ | ⁄ | ⁄ | ⁄ | 0.000 | 0.935 | 0.000 | 0.943 | 0.000 | 0.939 | 1.000 | 0.933 | 1.000 | 1.000 |
| 4 | U | 1 | 1.000 | 0.978 | 0.992 | 0.990 | 0.000 | 0.982 | 0.088 | 0.992 | 0.000 | 0.985 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 1.000 | 0.915 | 0.996 | 0.951 | 0.000 | 0.929 | 0.362 | 0.958 | 0.000 | 0.936 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | ⁄ | ⁄ | ⁄ | ⁄ | 0.050 | 0.994 | 0.998 | 0.991 | 0.042 | 0.996 | 0.230 | 1.000 | 0.226 | 1.000 |
| | Z | 5 | ⁄ | ⁄ | ⁄ | ⁄ | 0.278 | 0.953 | 1.000 | 0.946 | 0.320 | 0.957 | 0.996 | 0.999 | 1.000 | 1.000 |
| 5 | U | 1 | 0.392 | 0.981 | 0.364 | 0.991 | 0.014 | 0.986 | 0.082 | 0.990 | 0.010 | 0.987 | ⁄ | ⁄ | ⁄ | ⁄ |
| | U | 5 | 0.422 | 0.919 | 0.574 | 0.946 | 0.070 | 0.937 | 0.192 | 0.954 | 0.072 | 0.940 | ⁄ | ⁄ | ⁄ | ⁄ |
| | Z | 1 | ⁄ | ⁄ | ⁄ | ⁄ | 0.010 | 0.984 | 0.540 | 0.992 | 0.008 | 0.986 | 0.454 | 1.000 | 0.340 | 1.000 |
| | Z | 5 | ⁄ | ⁄ | ⁄ | ⁄ | 0.056 | 0.936 | 0.672 | 0.955 | 0.050 | 0.940 | 0.978 | 0.596 | 0.798 | 1.000 |

Abschließend ist festzuhalten, dass es sehr nützlich ist, den wahren Gaußprozess zur Ausreißeridentifikation zu kennen. Die PF-Methode ist sinnvoll, wenn erwartet wird, dass die Ausreißer über der kompletten Indexmenge gleichmäßig auffällig sind. Allerdings wird empfohlen, das Niveau $\alpha$ konservativ zu wählen. Auch die BLFQF-Methode ist in den meisten Fällen sehr gut darin, Ausreißer zu erkennen. Allerdings ist sie nicht sehr praktikabel, weil der Faktor zur Kontrolle der RNR mühsam adjustiert werden muss. Falls wie in Szenario 5 nur wenige auffällige Zeitpunkte vorliegen, ist die robustifizierte FUNTA eine mächtige Alternative, da sie dazu intendiert ist, ein hohes Gewicht auf ungewöhnliche Merkmale der Beobachtungen zu legen.

# 6. Zusammenfassung und Ausblick

In dieser Arbeit wurden Verfahren zur Ausreißeridentifikation in Kontingenztafeln und funktionalen Beobachtungen, insbesondere als Zielgrößen im generalisierten linearen Modell, entwickelt. Zunächst wurde das objektive Konzept der $\alpha$-Ausreißer, das in dieser Arbeit intensiv genutzt wird, um einige Aspekte erweitert und Parallelen zu Boxplots und Bagplots aufgezeigt. Die entwickelten Verfahren sind der OMPC- und OMPCL1-Algorithmus für Kontingenztafeln, die Pseudo-Datentiefen FUNTA und rFUNTA (in Verbindung mit dem geglätteten Bootstrap) im funktionalen Fall bei unbekannter Verteilung des Prozesses sowie die Verfahren PF und BLFQF bei bekannter Verteilung des Prozesses. Sie wurden anhand von Robustheitskriterien, Simulationsstudien und realen Datensätzen mit bestehenden Identifikationsverfahren verglichen.

Als Robustheitskriterien wurden im Wesentlichen diverse Bruchpunktkonzepte hinzugezogen. Hier ist die wichtige Unterscheidung zwischen Masking und Swamping sowie Implosions- und Explosionsfall hervorzuheben. Dabei hat rFUNTA den bestmöglichen Finite-Sample-Bruchpunkt im Implosions- und Explosionsfall.

In Bezug auf die Simulationsstudien fällt auf, dass die vorgeschlagenen Minimalmuster-basierten Identifizierer und Pseudo-Datentiefen künstlich hinzugefügte Ausreißer in den meisten Fällen besser erkennen als die zum Vergleich hinzugezogenen Standard-Verfahren. Auch in realen Datensätzen liefern die vorgeschlagenen Methoden plausible Ergebnisse. Bei bekannter Verteilung des zugrunde liegenden stochastischen Prozesses und Ausreißern, die über den gesamten Träger auffällig sind, ist die projektive Familie in Verbindung mit dem $\alpha$-Ausreißerkonzept eine mächtige Alternative zu Datentiefen in Verbindung mit dem geglätteten Bootstrap.

Minimalmuster-basierte Verfahren benötigen deutlich mehr Rechenzeit als der OL1-

Identifizierer. Die Implementierung von C-Code zur Bestimmung der Schätzer der einzelnen Teilmengen der Kontingenztafel, um diese Diskrepanz zu verringern, steht noch aus. Im Anschluss sollen die Minimalmuster-basierten Identifizierer innerhalb eines R-Paketes zugänglich gemacht werden.

Das generalisierte lineare Modell mit funktionalen Zielgrößen wurde bislang nicht auf multi-funktionale Zielgrößen $y : \mathcal{T} \to \mathbb{R}^d$ erweitert. Die Erforschung des Verhaltens der multivariaten FUNTA (Kuhnt und Rehage, 2016) in solchen Modellen im Vergleich mit anderen multivariaten Datentiefen wie der multivariaten funktionalen Halbraumtiefe (Claeskens *et al.*, 2014) ist deshalb ein bisher nicht betrachteter Aspekt. Andererseits können funktionale Beobachtungen auch in der Form $y : \mathcal{T}^d \to \mathbb{R}$ vorliegen. Schon für $d = 2$ ist die Verallgemeinerung des Schnittwinkelprinzips auf solche Strukturen nicht trivial, da statt Schnittpunkten Schnitt*kurven* vorliegen. Für jeden Punkt in der Schnittkurve ist der Schnittwinkel zu bestimmen. Eine Implementierung dieser Verallgemeinerung für FUNTA und rFUNTA steht noch aus.

Eine weitere interessante Anwendung besteht darin, FUNTA für alle Funktionenpaare zu bestimmen, die so erhaltene Matrix als Ähnlichkeitsmatrix aufzufassen und über $1 - d^{\mathrm{FUNTA}}$ ein Distanzmatrix für hierarchische funktionale Clusteranalyse zu erhalten. Erste eigene Untersuchungen zeigen diesbezüglich, dass das Linkage-Verfahren von Ward (1963) hier zu bevorzugen ist. Die robustifizierte FUNTA Pseudo-Datentiefe ist in diesem Kontext nicht geeignet, weil sie explizit dazu ausgelegt ist, mehr als zwei Kurven zu vergleichen, da die Bestimmung des Medians der paarweisen Maximalschnittwinkel ein zentraler Aspekt der rFUNTA ist.

Zuletzt ist auch die intensivere Nutzung robuster Kerndichteschätzer ein Forschungsansatz für die Zukunft. Die Verteilung der Datentiefen kann somit approximiert werden und das $\alpha$-Ausreißerprinzip den geglätteten Bootstrap zur Ausreißeridentifikation ersetzen. Robuste bivariate Kerndichteschätzung und das $\alpha$-Ausreißerprinzip könnte außerdem in der BLFQF-Methode statt des Bagplots zum Einsatz kommen.

# A. Appendix

## A.1 R-Code: Robuster Kerndichteschätzer

```
> # code originally written in Matlab by JooSeuk Kim
> # translated to R by Andre Rehage
> bandwidth_select <- function(x, b_type = 1, sigma = NA){
+  # output: h_opt: optimal bandwidth
+  # input: x: n vector
+  # b_type: bandwidth type: 1 -> lscv, 2 -> lkcv, 3 -> jakkola heuristic
+  # sigma: bandwidth array
+  if (is.na(sigma)) sigma <- 10^seq(-2, 1, length = 50)
+  l <- length(sigma)
+  n <- length(x)
+  ei <- rep(1, n)
+  # construct matrix of squared distances
+  X <- matrix(0, ncol = n, nrow = n)
+
+  X <- X + (ei %*% t(x) - x %*% t(ei))^2
+  if (b_type == 1){ #least squares cross validation
+    Jmin <- Inf
+    for (i in 1:l){
+       h <- sigma[i]
+       K1 <- (4*pi*h^2)^(-1/2)*exp(-X/(4*h^2))
+       K2 <- (2*pi*h^2)^(-1/2)*(exp(-X/(2*h^2)) - diag(n))
+       J <- sum(K1)/(n^2) - 2/(n*(n-1))*sum(K2)
+       if(J < Jmin){
+         h_opt <- h
+         Jmin <- J
+       }
+     }
+   }
+  if (b_type == 2){ # log-likelihood cross validation
+     Jmax <- -Inf
+     for(i in 1:l){
+       h <- sigma[i]
+       K <- (2*pi*h^2)^(-1/2) * (exp(-X/(2*h^2)) - diag(n))
+       J <- sum(log(colSums(K)/(n - 1)))/n
+       if (J > Jmax){
+         h_opt <- h
+         Jmax <- J
```

```
+       }
+     }
+   }
+   return(h_opt)
+ }
 rho <- function(x, type, a, b, c){
+   # J: \frac{1}{n} sum_{i=1}^n \rho(x_i)
+   # x: input
+   # type: type of loss function, 1-> Huber, 2-> Hampel
+   # a, b, c: parameters
+
+   n <- length(x)
+
+   if(type == 1){ # Huber
+     in1 <- which(x <= a)
+     in2 <- which(x > a)
+     J <- sum(1/2 * x[in1]^2) + sum(a*(x[in2] - a) + 1/2 * a^2)
+   }
+   if(type == 2){ # Hampel
+     in1 <- which(x <= a)
+     in2 <- which(a < x & x <= b)
+     in3 <- which(b < x & x <= c)
+     in4 <- which(c < x)
+     p <- -a/(c - b)
+     q <- a*c/(c - b)
+     r <- a*b - 1/2*a^2 - 1/2*p*b^2-q*b
+     temp <- numeric(n)
+     temp[in1] <- 1/2 * x[in1]^2
+     temp[in2] <- a * (x[in2] - a) + 1/2 * a^2
+     temp[in3] <- 1/2 * p * x[in3]^2 + q * x[in3] + r
+     temp[in4] <- 1/2 * p * c^2 + q * c + r
+     J <- sum(temp)
+   }
+   J <- J/n
+   return(J)
+ }
 psi <- function(input, type, a, b, c){
+   #out: psi function evaluated at input
+   #input: input
+   #type: type of loss function, 1-> Huber, 2-> Hampel
+   #a, b, c: parameters
+   n <- nrow(input)
+   m <- ncol(input)
+   out <- matrix(0, n, m)
+
+   if(type == 1){ # Huber
+     out <- apply(cbind(input, a), 1, min)
+   }
+   if(type == 2){ # Hampel
+     i1 <- which(input < a)
+     i2 <- which(input >= a & input < b)
+     i3 <- which(input >= b & input < c)
```

```
+     i4 <- which(input >= c)
+     out[i1] <- input[i1]
+     out[i2] <- a
+     out[i3] <- a * (c - input[i3])/(c - b)
+     out[i4] <- 0
+   }
+   return(out)
+ }
 parameter_select <- function(K, type){
+   # first find median
+   n <- nrow(K)
+
+   # initial weights
+   w <- rep(1, n)/n
+   tol <- 10^(-8)
+
+   norm2mu <- t(w) %*% K %*% w
+   normdiff <- sqrt(diag(K) + norm2mu - 2 * K %*% w)
+   # rho = abs(x)
+   J <- sum(normdiff)/n
+
+   repeat{
+     J_old <- J
+     w <- 1/normdiff
+     w <- w/sum(w)
+     norm2mu <- t(w) %*% K %*% w
+     normdiff <- sqrt(diag(K) + norm2mu - 2 * K %*% w)
+     J <- sum(normdiff)/n
+     if(abs(J - J_old) < J_old*tol) break
+   }
+
+   sort_norm <- sort(normdiff, decreasing = TRUE)
+
+   if(type == 1){
+     a1 <- sort_norm[floor(n/2)]
+     a2 <- a3 <- 0
+   }
+
+   if(type == 2){
+     a1 <- sort_norm[floor(n/2)]
+     a2 <- sort_norm[floor(n/20)]
+     a3 <- max(normdiff)
+   }
+   return(c(a1, a2, a3))
+ }
 gauss_kern <- function(dist, h){
+   K <- exp(-dist/(2*h^2))/((2*pi*h^2)^(1/2))
+   return(K)
+ }
 robkde <- function(x, h, type = 2){
+   # output: w: weights for RKDE
+   # a,b,c: paramters
```

```
+   # input: x: n vector
+   # h: bandwidth
+   # type: type of loss function, 1-> Huber, 2-> Hampel
+   n <- length(x)
+
+   # construct kernel matrix(Gaussian kernel with bandwidth h)
+   X <- matrix(0, nrow = n, ncol = n)
+   ei <- rep(1, n)
+   X <- X + (ei %*% t(x) - x %*% t(ei))^2
+   # evaluate gaussian kernel of dataset based on squared
+   # distances to kernel centers;
+   # K: Gaussian kernel
+   # dist: matrix of squared distances
+   # h: bandwidth
+   K <- gauss_kern(X, h)
+
+   # find median absolute deviation
+   a <- parameter_select(K, type)
+
+   # initial weights
+   w <- rep(1, n)/n
+   tol <- 10^(-8)
+   norm2mu <- t(w) %*% K %*% w
+   normdiff <- sqrt(diag(K) + norm2mu - 2 * K %*% w)
+   J <- rho(normdiff, type, a[1], a[2], a[3])
+   repeat{
+     J_old <- J
+     w <- psi(normdiff, type, a[1], a[2], a[3])/normdiff
+     w <- w/sum(w)
+
+     norm2mu <- t(w) %*% K %*% w
+     normdiff <- sqrt(diag(K) + norm2mu - 2 * K %*% w)
+
+     J <- rho(normdiff, type, a[1], a[2], a[3])
+     if(abs(J - J_old) < J_old*tol) break
+   }
+   return(list(w = w, a = a))
+ }
```

## A.2 FUNKTIONALE HAUPTKOMPONENTENANALYSE

DEFINITION A.1. Karhunen-Loève-Entwicklung *(Hall* et al.*, 2008, Formel (16))* *Sei* $\mathcal{Y} = \{\mathcal{Y}(t) \mid t \in \mathcal{T}\}$ *ein stochastischer Prozess mit* $\mathsf{E}(\mathcal{Y}(t)) = \mu(t)$ *und* $\mathsf{Cov}(\mathcal{Y}(s), \mathcal{Y}(t)) = \sigma(s,t)$*. Dann gilt:* $\mathcal{Y}(t) - \mu(t) = \sum_{j=1}^{\infty} \xi_j \psi_j(t)$ *ist die Karhunen-Loève-Entwicklung, wobei* $\psi_j$ *als Lösung von* $\int \mathsf{Cov}(\mathcal{Y}(s), \mathcal{Y}(t))\psi_j(t)ds = \theta_j\psi_j(t)$ *die zum Eigenwert* $\theta_j$ *gehörende* $j$*-te orthonormale Eigenfunktion von* $\sigma(s,t)$ *ist und* $\xi_j = \int \mathcal{Y}(t)\psi_j(t)dt$ *der* $j$*-te Hauptkomponenten-Score (PC Score).*

Dabei sind die Eigenwerte $\theta_j$ absteigend geordnet und es gilt: $\sum_{j=1}^{\infty} \theta_j < \infty$. Für die Eigenfunktionen $\psi_j(t)$ wird angenommen, dass sie zweimal stetig differenzierbar sind, siehe Chiou und Müller (2007).

## A.3 ZEILENTENSORPRODUKT

Sei $\odot$ das Zeilentensorprodukt: $\mathbf{A} \odot \mathbf{B} = (\mathbf{A} \otimes \mathbb{1}_b^\top) \circ (\mathbb{1}_a^\top \otimes \mathbf{B})$ für $\mathbf{A} \in \mathbb{R}^{m \times a}$ und $\mathbf{B} \in \mathbb{R}^{m \times b}$. Nach Scheipl *et al.* (2016) gilt für das generalisierte additive gemischte Modell, in dem sowohl Kovariablen als auch Zielgrößen funktional sein können:

$$\mathbf{Z}_j(\mathbf{X}) = \mathbf{\Phi}_{xj} \odot \mathbf{\Phi}_{tj}, \tag{A.1}$$

wobei $\mathbf{\Phi}_{xj}$ die Evaluationen der gewählten marginalen Basis an den Stellen $x_{1j}, \ldots, x_{nj}$ bezeichnet sowie $\mathbf{\Phi}_{tj}$ deren Evaluationen an den Stellen $t_1, \ldots, t_T$. Im Fall der generalisierten *function-on-scalar regression* vereinfacht sich (A.1) zu

$$\mathbf{Z}_j(\mathbf{X}) = (\mathbf{x}_{\bullet j} \otimes \mathbf{1}_T) \odot (\mathbf{1}_n \otimes [\phi_k(t_l)]_{k=1,\ldots,K}^{l=1,\ldots,T})$$

$$
\begin{aligned}
&= \begin{pmatrix} x_{1j} \\ \vdots \\ x_{1j} \\ \vdots \\ x_{nj} \\ \vdots \\ x_{nj} \end{pmatrix} \odot \begin{pmatrix} [\phi_k(t_l)]_{k=1,\dots,K}^{l=1,\dots,T} \\ \vdots \\ [\phi_k(t_l)]_{k=1,\dots,K}^{l=1,\dots,T} \end{pmatrix} = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{1j} \\ \vdots \\ x_{nj} \\ \vdots \\ x_{nj} \end{pmatrix} \odot \begin{pmatrix} \phi_1(t_1) & \cdots & \phi_K(t_1) \\ \vdots & & \vdots \\ \phi_1(t_T) & \cdots & \phi_K(t_T) \\ \vdots & & \vdots \\ \phi_1(t_1) & \cdots & \phi_K(t_1) \\ \vdots & & \vdots \\ \phi_1(t_T) & \cdots & \phi_K(t_T) \end{pmatrix} \\
&= \begin{pmatrix} \begin{pmatrix} x_{1j} \\ \vdots \\ x_{1j} \\ \vdots \\ x_{nj} \\ \vdots \\ x_{nj} \end{pmatrix} \mathbf{1}_K^\top \end{pmatrix} \circ \begin{pmatrix} \phi_1(t_1) & \cdots & \phi_K(t_1) \\ \vdots & & \vdots \\ \phi_1(t_T) & \cdots & \phi_K(t_T) \\ \vdots & & \vdots \\ \phi_1(t_1) & \cdots & \phi_K(t_1) \\ \vdots & & \vdots \\ \phi_1(t_T) & \cdots & \phi_K(t_T) \end{pmatrix} \\
&= \begin{pmatrix} x_{1j}\phi_1(t_1) & \cdots & x_{1j}\phi_K(t_1) \\ \vdots & & \vdots \\ x_{1j}\phi_1(t_T) & \cdots & x_{1j}\phi_K(t_T) \\ \vdots & & \vdots \\ x_{nj}\phi_1(t_1) & \cdots & x_{nj}\phi_K(t_1) \\ \vdots & & \vdots \\ x_{nj}\phi_1(t_T) & \cdots & x_{nj}\phi_K(t_T) \end{pmatrix}.
\end{aligned}
$$

## A.4 ZUSÄTZLICHE SIMULATIONSERGEBNISSE

Tabelle A.1: Simulationsergebnisse aus Kuhnt *et al.* (2014) mit $\zeta = 0.5$

| | | $n_{11} = 62$ | | $n_{11} = 141$ | |
|---|---|---|---|---|---|
| | | RPR | RNR | RPR | RNR |
| Sz. 1 | OL1 | 0.320 | 0.963 | 0.480 | 0.974 |
| | OLTCS | 0.480 | 0.946 | 0.530 | 0.950 |
| | OMPC | 0.680 | 0.754 | 0.820 | 0.773 |

| | | $n_{11} = 39$ | | $n_{11} = 105$ | |
|---|---|---|---|---|---|
| | | RPR | RNR | RPR | RNR |
| Sz. 2 | OL1 | 0.620 | 0.979 | 0.620 | 0.989 |
| | OLTCS | 0.740 | 0.943 | 0.680 | 0.937 |
| | OMPC | 0.890 | 0.899 | 0.900 | 0.909 |
| | OMPCL1 | 0.910 | 0.877 | 0.930 | 0.909 |

| | | 2 Antitypen | | 1 Typ, 1 Antityp | | 2 Typen | |
|---|---|---|---|---|---|---|---|
| | | $n_{11} = 39, n_{12} = 42$ | | $n_{11} = 39, n_{12} = 110$ | | $n_{11} = 105, n_{12} = 110$ | |
| | | RPR | RNR | RPR | RNR | RPR | RNR |
| Sz. 3 | OL1 | 0.035 | 0.960 | 0.725 | 0.986 | 0.200 | 0.983 |
| | OLTCS | 0.295 | 0.896 | 0.680 | 0.937 | 0.295 | 0.908 |
| | OMPC | 0.435 | 0.868 | 1.000 | 0.878 | 0.470 | 0.901 |
| | OMPCL1 | 0.615 | 0.833 | 1.000 | 0.846 | 0.720 | 0.874 |

| | | 2 Antitypen | | 1 Typ, 1 Antityp | | 2 Typen | |
|---|---|---|---|---|---|---|---|
| | | $n_{11} = 39, n_{22} = 23$ | | $n_{11} = 39, n_{22} = 79$ | | $n_{11} = 105, n_{22} = 79$ | |
| | | RPR | RNR | RPR | RNR | RPR | RNR |
| Sz. 4 | OL1 | 0.740 | 0.976 | 0.495 | 0.980 | 0.635 | 0.984 |
| | OLTCS | 0.795 | 0.948 | 0.670 | 0.923 | 0.695 | 0.921 |
| | OMPC | 0.975 | 0.804 | 0.840 | 0.834 | 0.965 | 0.857 |
| | OMPCL1 | 0.975 | 0.692 | 0.885 | 0.776 | 0.975 | 0.818 |

| | | 2 Antitypen | | 1 Typ, 1 Antityp | | 2 Typen | |
|---|---|---|---|---|---|---|---|
| | | $n_{11} = 27, n_{12} = 29$ | | $n_{11} = 27, n_{12} = 128$ | | $n_{11} = 124, n_{12} = 128$ | |
| | | RPR | RNR | RPR | RNR | RPR | RNR |
| Sz. 5 | OL1 | 0.140 | 0.896 | 0.980 | 0.987 | 0.450 | 0.969 |
| | OLTCS | 0.370 | 0.870 | 0.835 | 0.936 | 0.430 | 0.898 |
| | OMPC | 0.880 | 0.771 | 1.000 | 0.643 | 0.855 | 0.829 |
| | OMPCL1 | 0.975 | 0.725 | 1.000 | 0.653 | 0.950 | 0.806 |

| | | 2 Antitypen | | 1 Typ, 1 Antityp | | 2 Typen | |
|---|---|---|---|---|---|---|---|
| | | $n_{11} = 18, n_{23} = 9$ | | $n_{11} = 18, n_{23} = 49$ | | $n_{11} = 67, n_{23} = 49$ | |
| | | RPR | RNR | RPR | RNR | RPR | RNR |
| Sz. 6 | OL1 | 0.963 | 0.991 | 0.936 | 0.992 | 0.935 | 0.992 |
| | OLTCS | 0.850 | 0.929 | 0.855 | 0.933 | 0.815 | 0.932 |
| | OMPC | 0.990 | 0.940 | 0.990 | 0.953 | 1.000 | 0.956 |
| | OMPCL1 | 0.910 | 0.995 | 0.895 | 0.997 | 0.940 | 0.997 |

## A.5 DESIGNMATRIZEN

Designmatrix mit Effektkodierung im Beispiel einer $3 \times 3$ Kontingenztafel im Unabhängigkeitsmodell:

$$\widetilde{\mathbf{X}}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 0 & -1 & 1 & 0 & -1 & 1 & 0 & -1 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \end{pmatrix} =: \widetilde{\mathbf{X}}_{3\times 3}^\top. \tag{A.2}$$

Designmatrix mit Effektkodierung im Beispiel einer $4 \times 4$ Kontingenztafel im Unabhängigkeitsmodell:

$$\mathbf{X}_{4\times 4}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 & 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 & 1 & -1 \end{pmatrix}. \tag{A.3}$$

Designmatrix im Beispiel sozialer Netzwerke junger Mütter:

$$\mathbf{X}^\top = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{A.4}$$

# B. SYMBOLVERZEICHNIS

Tabelle B.1: Allgemeine Bezeichnungen

| Symbol | Bedeutung |
|---|---|
| $\mathbb{B}$ | Banachraum |
| $\mathcal{C}(\mathcal{T})$ | Raum der reellen, stetigen Funktionen auf $\mathcal{T}$ |
| $\mathbb{N}$ | Menge der natürlichen Zahlen |
| $\mathbb{N}_0$ | $\mathbb{N} \cup \{0\}$ |
| $\mathbb{R}$ | Menge der reellen Zahlen |
| $\mathbb{R}^+$ | Menge der echt positiven reellen Zahlen |
| $\mathbb{R}_0^+$ | $\mathbb{R}^+ \cup \{0\}$ |
| $\lfloor x \rfloor$ | Gaußklammer: größte ganze Zahl, die kleiner oder gleich $x$ ist |
| $[a, b]$ | Menge aller Elemente $x \in \mathbb{R} : a \leq x \leq b$ |
| $]a, b[$ | Menge aller Elemente $x \in \mathbb{R} : a < x < b$ |
| $\bar{\mathbb{O}}, \mathbb{O}^\circ, \partial\mathbb{O}$ | Abschluss der Menge $\mathbb{O}$, Inneres der Menge $\mathbb{O}$, Rand der Menge $\mathbb{O}$ |
| $\text{diag}(\mathbf{x})$ | Diagonalmatrix $\in \mathbb{R}^{k \times k}$ mit Elementen $\mathbf{x} \in \mathbb{R}^k$ auf der Hauptdiagonale |
| $\mathbb{1}_k$ | Einsvektor: $(1, \ldots, 1)^\top \in \mathbb{N}^k$ |
| $\mathbf{I}_k$ | Einheitsmatrix: $\mathbf{I}_k = \text{diag}(\mathbb{1}_k) \in \{0, 1\}^{k \times k}$ |
| $\mathbf{e}_i$ | $i$-ter kanonischer Einheitsvektor |
| $\mathbf{1}_{\mathbb{O}}(x)$ | Indikatorfunktion: $\mathbf{1}_{\mathbb{O}}(x) = 1$ falls $x \in \mathbb{O}$, sonst 0 |
| $\text{vec}, \wp$ | Vec-Operator für Matrizen, Arrays |
| $\top$ | Transponiertzeichen |
| $\boldsymbol{\theta} \subset \boldsymbol{\Theta}$ | unbekannter Parametervektor der Verteilung |
| $f(\cdot, \boldsymbol{\theta})$ | Wahrscheinlichkeitsdichte |
| $F(\cdot, \boldsymbol{\theta})$ | Verteilungsfunktion |
| $P_{\boldsymbol{\theta}}, \mathcal{P}$ | Verteilung, Verteilungsfamilie |
| $Bin(n, \pi)$ | Binomialverteilung mit $n \in \mathbb{N}$ und $\pi \in [0, 1]$ |
| $\mathcal{GP}(\mu, \sigma)$ | Gaußprozess mit $\mu = \{\mu(t) \vert t \in \mathcal{T}\}$ und $\sigma = \{\sigma(s, t) \vert s, t \in \mathcal{T}\}$ |
| $Mult(n, \boldsymbol{\theta})$ | Multinomialverteilung mit $n \in \mathbb{N}$ und $\boldsymbol{\theta} \in [0, 1]^k$ |
| $\aleph(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ | Normalverteilung mit Erwartungswert $\boldsymbol{\mu} \in \mathbb{R}^k$ und Kovarianz $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ |
| $\Phi$ | Verteilungsfunktion der Standardnormalverteilung |
| $Poi(\lambda)$ | Poissonverteilung mit Parameter $\lambda \in \mathbb{R}^+$ |
| $\mathcal{X}, \mathcal{Y}$ | univariate Zufallsvariablen |
| $\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}$ | vektor- oder matrixwertige Zufallsvariablen |
| $\mathsf{E}, \mathsf{Cov}$ | Erwartungswertoperator, Kovarianzoperator |

Tabelle B.2: Bezeichnungen in Kapitel 2

| Symbol | Bedeutung |
|---|---|
| $\mathrm{out}(\alpha, P_{\boldsymbol{\theta}})$ | $\alpha$-Ausreißerregion der Verteilung $P_{\boldsymbol{\theta}}$ zum Niveau $\alpha \in ]0,1[$ |
| $B(\alpha)$ | Schwellenwert der Dichte einer $\alpha$-Ausreißerregion |
| OI, **OI** | Ausreißeridentifizierer einer Beobachtung / eines Vektors |
| $\hat{f}_k(\cdot, h)$ | Kerndichteschätzer basierend auf $k$ Beobachtungen mit Bandbreite $h \in \mathbb{R}^+$ |
| $K(\cdot)$ | Kern |
| $D(\cdot), d(\cdot)$ | Populations- und Stichprobendatentiefe einer Beobachtung |

Tabelle B.3: Bezeichnungen in Kapitel 3

| Symbol | Bedeutung |
|---|---|
| $\mathcal{Y}$ | Zielvariable |
| $\mathbf{y} = (y_1, \ldots, y_k)$ | Realisationen der Zielvariable |
| $\mathbf{X}, \mathbf{X}_{i\bullet}, \mathbf{X}_{\bullet j}$ | Designmatrix, $i$-te Zeile, $j$-te Spalte der Designmatrix |
| $a, b, c$ | Funktionen der Exponentialfamilie |
| $\boldsymbol{\beta} \in \mathbb{R}^p$ | unbekannter Parametervektor im GLM |
| $\eta$ | Linearer Prädiktor im GLM |
| $g, h$ | Link-, Responsefunktion |
| $T$ | Schätzfunktion |
| $\widehat{\boldsymbol{\beta}}^{\mathrm{ML}}, \widehat{\boldsymbol{\beta}}^{L_1}, \widehat{\boldsymbol{\beta}}^{\mathrm{LTCS}}$ | ML-Schätzer, $L_1$-Schätzer, LTCS-Schätzer |
| $L, l$ | Likelihood, Log-Likelihood |
| $\varepsilon^*$ | Finite-Sample-Bruchpunkt |
| $\mathcal{F}$ | wahres Signal |
| $\boldsymbol{\Delta}, \boldsymbol{\delta}$ | Koeffizientenmatrix und -vektor der Fourierdarstellung |
| $\phi, \varphi$ | Basisfunktionen der Fourierdarstellung |

Tabelle B.4: Bezeichnungen in Kapitel 4

| Symbol | Bedeutung |
|---|---|
| $\mathbb{M} = \{\mathbb{M}_1, \ldots, \mathbb{M}_V\}$ | Menge der Minimalmuster |
| $\mathbb{S}$ | Menge der strengen Minimalmuster |
| $\xi^{\mathfrak{M}}, \varepsilon^{\mathfrak{M}}$ | Masking-Punkt, Masking-Bruchpunkt eines **OI** |
| $\xi^{\mathfrak{S}}, \varepsilon^{\mathfrak{S}}$ | Swamping-Punkt, Swamping-Bruchpunkt eines **OI** |
| $\zeta$ | Tuningparameter im OMPC-Algorithmus |

Tabelle B.5: Bezeichnungen in Kapitel 5

| Symbol | Bedeutung |
|---|---|
| $w_1, \ldots, w_m$ | Schnittwinkel |
| $\gamma_u(\cdot,\cdot), \Gamma(\cdot)$ | unvollständige untere Gammafunktion, Gammafunktion |
| $\mathrm{erf}(\cdot)$ | Fehlerfunktion |

# LITERATURVERZEICHNIS


Adler, R. J. (1981). *The Geometry of Random Fields*. John Wiley & Sons, Chichester.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Hoboken, 2. Auflage.

Arribas-Gil, A. und Romo, J. (2014). Shape outlier detection and visualization for functional data: the outliergram. *Biostatistics*, **15**(4), 603–619.

Aston, J. A. D., Chiou, J.-M. und Evans, J. P. (2010). Linguistic pitch analysis using functional principal component mixed effect models. *Journal of the Royal Statistical Society, Series C*, **59**(2), 297–317.

Barnett, V. (1976). The ordering of multivariate data. *Journal of the Royal Statistical Society, Series A*, **139**(3), 318–355.

Barnett, V. und Lewis, T. (1994). *Outliers in Statistical Data*. Wiley, New York, 3. Auflage.

Barrett, H. H. und Myers, K. J. (2004). *Foundations of Image Science*. Wiley Series in Pure and Applied Optics. Wiley, Hoboken.

Bishop, Y. M. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics*, **25**(2), 383–399.

Borowski, M., Rudak, N., Hussong, B., Wied, D., Kuhnt, S. und Tillmann, W. (2014). On- and offline detection of structural breaks in thermal spraying processes. *Journal of Applied Statistics*, **41**(5), 1073–1090.

Brent, R. P. (1973). *Algorithms for Minimization Without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

Brown, M. B. (1974). Identification of the sources of significance in two-way contingency tables. *Journal of the Royal Statistical Society, Series C*, **23**(3), 405–413.

Brown, P. J., Fearn, T. und Vannucci, M. (2001). Bayesian wavelet regression on curves with applications to a spectroscopic calibration problem. *Journal of the American Statistical Association*, **96**(454), 398–408.

Bruffaerts, C., Verardi, V. und Vermandele, C. (2014). A generalized boxplot for skewed and heavy-tailed distributions. *Statistics & Probability Letters*, **95**, 110–117.

Büning, H. und Trenkler, G. (1994). *Nichtparametrische statistische Methoden*. De Gruyter Lehrbuch. de Gruyter, Berlin, 2. Auflage.

Byrd, R. H., Lu, P., Nocedal, J. und Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, **16**(5), 1190–1208.

Cantoni, E. und Ronchetti, E. M. (2001). Robust inference for generalized linear models. *Journal of the American Statistical Association*, **96**(455), 1022–1030.

Chiou, J.-M. und Müller, H.-G. (2007). Diagnostics for functional regression via residual processes. *Computational Statistics & Data Analysis*, **51**(10), 4849–4863.

Christmann, A. (1992). *Ausreißeridentifikation und robuste Schätzer im logistischen Regressionsmodell*. Dissertation, Universität Dortmund, Dortmund.

Christmann, A. (1998). *On positive breakdown point estimators in regression models with discrete response variables*. Habilitationsschrift, Universität Dortmund, Dortmund.

Claeskens, G., Hubert, M., Slaets, L. und Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, **109**(505), 411–423.

Cramér, H. und Leadbetter, M. R. (1967). *Stationary and Related Stochastic Processes*. John Wiley & Sons, New York.

Craven, P. und Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**(4), 377–403.

Cuesta-Albertos, J. A., del Barrio, E., Fraiman, R. und Matrán, C. (2007). The random projection method in goodness of fit for functional data. *Computational Statistics & Data Analysis*, **51**(10), 4814–4831.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference*, **147**, 1–23.

Cuevas, A., Febrero, M. und Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis*, **51**(2), 1063–1074.

Cuevas, A., Febrero, M. und Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, **22**(3), 481–496.

Davies, P. L. und Gather, U. (1989). The identification of multiple outliers. Forschungsbericht Nr. 89/1, Fachbereich Statistik, Universität Dortmund.

Davies, P. L. und Gather, U. (1993). The identification of multiple outliers. *Journal of the American Statistical Association*, **88**(423), 784–792.

Delaigle, A. und Hall, P. (2010). Defining probability density for a distribution of random functions. *The Annals of Statistics*, **38**(2), 1171–1193.

Donoho, D. L. und Gasko, M. (1992). Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *The Annals of Statistics*, **20**(4), 1803–1827.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1), 1–26.

Fahrmeir, L. und Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York.

Faraway, J. J. (1997). Regression analysis for a functional response. *Technometrics*, **39**(3), 254–261.

Febrero, M., Galeano, P. und González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal $NO_x$ levels. *Environmetrics*, **19**(4), 331–345.

Febrero-Bande, M. und Oviedo de la Fuente, M. (2012). Statistical computing in functional data analysis: The R package fda.usc. *Journal of Statistical Software*, **51**(4), 1–28.

Finucan, H. M. (1964). The mode of a multinomial distribution. *Biometrika*, **51**(3/4), 513–517.

Fraiman, R. und Muniz, G. (2001). Trimmed means for functional data. *Sociedad de Estadistica e Investigacion Operativa Test*, **10**(2), 419–440.

Fried, R., Liboschik, T., Elsaied, H., Kitromilidou, S. und Fokianos, K. (2014). On outliers and interventions in count time series following GLMs. *Austrian Journal of Statistics*, **43**(3-4), 181–193.

Gather, U., Kuhnt, S. und Pawlitschko, J. (2003). Concepts of outlyingness for various data structures, in: J. C. Misra (Hrsg.), *Industrial Mathematics and Statistics*, 545–585, Narosa Publishing House, New Delhi.

Glass, D. V. und Berent, J. (1954). *Social Mobility in Britain*. International Library of Sociology and Social Reconstruction. Routledge & Paul.

Goldsmith, J., Zipunnikov, V. und Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, **71**(2), 344–353.

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M. W., Swihart, B., Xiao, L., Crainiceanu, C. und Reiss, P. T. (2016). *refund: Regression with Functional Data*. R package version 0.1-16.

Goodman, L. A. (1971). A simple simultaneous test procedure for quasi-independence in contingency tables. *Journal of the Royal Statistical Society, Series C*, **20**(2), 165–177.

Guo, W. (2004). Functional data analysis in longitudinal settings using smoothing splines. *Statistical Methods in Medical Research*, **13**(1), 49–62.

Hall, P., Müller, H.-G. und Yao, F. (2008). Modeling sparse generalized longitudinal observations with latent Gaussian processes. *Journal of the Royal Statistical Society, Series B*, **70**(4), 703–723.

Hoaglin, D. (2006). Summarizing shape numerically: The $g$-and-$h$ distributions, in: D. Hoaglin, F. Mosteller, und J. Tukey (Hrsg.), *Exploring Data Tables, Trends, and Shapes*, 461–513, Wiley, Hoboken, 2. Auflage.

Hoerl, A. E. und Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Horváth, L. und Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer, New York.

Hoyden, L. (2011). *Wahl einer geeigneten Linkfunktion im generalisierten linearen Modell mit Anwendung auf einen thermokinetischen Spritzprozess*. Bachelorarbeit, TU Dortmund, Dortmund.

Hsing, T. und Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley & Sons, Chichester.

Hubert, M. (1997). The breakdown value of the $L_1$ estimator in contingency tables. *Statistics & Probability Letters*, **33**, 419–425.

Hubert, M. und Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, **52**(12), 5186–5201.

Iacus, S. M. (2008). *Simulation and inference for stochastic differential equations: With R examples*. Springer Series in Statistics. Springer, New York.

James, G. M. (2002). Generalized linear models with functional predictors. *Journal of the Royal Statistical Society, Series B*, **64**(3), 411–432.

Johnson, S. G. und Narasimhan, B. (2013). *cubature: Adaptive multivariate integration over hypercubes*. R package version 1.1-2.

Kaballo, W. (2000). *Einführung in die Analysis I*. Spektrum, Heidelberg, 2. Auflage.

Kedem, B. und Fokianos, K. (2002). *Regression Models for Time Series Analysis*. Wiley, Hoboken.

Kim, J. (2011). MATLAB-Code in rkde_code.zip. `http://web.eecs.umich.edu/~cscott/code/rkde_code.zip`. Zuletzt abgerufen am 15.12.2016.

Kim, J. und Scott, C. D. (2012). Robust kernel density estimation. *Journal of Machine Learning Research*, **13**(1), 2529–2565.

Königsberger, K. (2004). *Analysis 1*. Springer, Berlin, 6. Auflage.

Koshevoy, G. und Mosler, K. (1997). Zonoid trimming for multivariate distributions. *The Annals of Statistics*, **25**(5), 1998–2017.

Kuhnt, S. (2000). *Ausreißeridentifikation im Loglinearen Poissonmodell für Kontingenztafeln unter Einbeziehung robuster Schätzer*. Dissertation, Universität Dortmund, Dortmund.

Kuhnt, S. (2004). Outlier identification procedures for contingency tables using maximum likelihood and $L_1$ estimates. *Scandinavian Journal of Statistics*, **31**(3), 431–442.

Kuhnt, S. (2010). Breakdown concepts for contingency tables. *Metrika*, **71**, 281–294.

Kuhnt, S. und Rehage, A. (2013). The concept of $\alpha$-outliers in structured data situations, in: C. Becker, R. Fried, und S. Kuhnt (Hrsg.), *Robustness and Complex Data Structures*, 91–108, Springer, Berlin.

Kuhnt, S. und Rehage, A. (2016). An angle-based multivariate functional pseudo-depth for shape outlier detection. *Journal of Multivariate Analysis*, **146**, 325–340.

Kuhnt, S., Rapallo, F. und Rehage, A. (2014). Outlier detection in contingency tables based on minimal patterns. *Statistics and Computing*, **24**(3), 481–491.

Kuhnt, S., Rehage, A., Becker-Emden, C., Tillmann, W. und Hussong, B. (2016). Residual analysis in generalized function-on-scalar regression for an HVOF spraying process. *Quality and Reliability Engineering International*, **32**(6), 2139–2150.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, **18**(1), 405–414.

Lopez-Pintado, S. und Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, **104**(486), 718–734.

Malfait, N. und Ramsay, J. O. (2003). The historical functional linear model. *The Canadian Journal of Statistics*, **31**(2), 115–128.

McKinlay, J. B. (1973). Social networks, lay consultation and help-seeking behavior. *Social Forces*, **51**(3), 275–292.

McLeod, A. I., Yu, H. und Krougly, Z. (2007). Algorithms for linear time series analysis: With r package. *Journal of Statistical Software*, **23**(5).

Mosler, K. (2013). Depth statistics, in: C. Becker, R. Fried, und S. Kuhnt (Hrsg.), *Robustness and Complex Data Structures*, 17–34, Springer, Berlin.

Mosler, K. und Polyakova, Y. (2012). General notions of depth for functional data. `http://arxiv.org/abs/1208.1981v1`. Preprint.

Müller, H. G. und Stadtmüller, U. (2005). Generalized functional linear models. *The Annals of Statistics*, **33**(2), 774–805.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, **9**(1), 141–142.

Nagy, S., Gijbels, I., Omelka, M. und Hlubinka, D. (2016). Integrated depth for functional data: Statistical properties and consistency. *ESAIM: Probability and Statistics*, **20**, 95–130.

Nelder, J. A. und Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**(3), 370–384.

Nieto-Reyes, A. und Battey, H. (2016). A topologically valid definition of depth for functional data. *Statistical Science*, **31**(1), 61–79.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, **1**(6), 327–332.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**(3), 1065–1076.

Pearson, K. (1904). *On the theory of contingency and its relation to association and normal correlation*. Draper's Company Memoires, London.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O. und Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York, 2. Auflage.

Ramsay, J. O., Wickham, H., Graves, S. und Hooker, G. (2014). *fda: Functional Data Analysis*. R package version 2.4.4.

Rasmussen, C. E. und Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge.

Rehage, A. (2016). *FUNTA: Functional Tangential Angle Pseudo-Depth*. R package version 0.1.0.

Rehage, A. und Kuhnt, S. (2014). An angle-based functional depth measure for outlier detection, in: E. G. Bongiorno, E. Salinelli, A. Goia, und P. Vieu (Hrsg.), *Contributions in infinite-dimensional statistics and related topics*, 233–238, Società Editrice Esculapio, Bologna.

Rehage, A. und Kuhnt, S. (2016). *alphaOutlier: Obtain Alpha-Outlier Regions for Well-Known Probability Distributions*. R package version 1.2.0.

Reiss, P. T., Huang, L. und Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. *The International Journal of Biostatistics*, **6**(1).

Rousseeuw, P. J. und Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.

Rousseeuw, P. J., Ruts, I. und Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, **53**(4), 382–387.

Scheipl, F., Gertheiss, J. und Greven, S. (2016). Generalized functional additive mixed models. *Electronic Journal of Statistics*, **10**, 1455–1492.

Sguera, C., Galeano, P. und Lillo, R. E. (2016). Functional outlier detection by a local depth with application to $NO_x$ levels. *Stochastic Environmental Research and Risk Assessment*, **30**(4), 1115–1130.

Shane, K. V. und Simonoff, J. S. (2001). A robust approach to categorical data analysis. *Journal of Computational and Graphical Statistics*, **10**(1), 135–157.

Shang, H. L. und Hyndman, R. J. (2013). *fds: Functional data sets*. R package version 1.7.

Sheather, S. J. und Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**(3), 683–690.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, **62**(318), 626–633.

Steinwart, I. und Christmann, A. (2008). *Support Vector Machines*. Springer, New York.

Sun, Y., Genton, M. G. und Nychka, D. W. (2012). Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked? *Stat*, **1**(1), 68–74.

Terbeck, W. und Davies, P. L. (1998). Interactions and outliers in the two-way analysis of variance. *The Annals of Statistics*, **26**, 1279–1305.

Tillmann, W., Kuhnt, S., Hussong, B., Rehage, A. und Rudak, N. (2012). Einführung eines Tageseffekt-Schätzers zur Verbesserung der Vorhersage von Partikeleigenschaften in einem HVOF-Spritzstrahl. *Thermal Spray Bulletin*, **5**(2), 132–139.

Tillmann, W., Hussong, B., Priggemeier, T., Kuhnt, S., Rudak, N. und Weinert, H. (2013). Influence of parameter variations on WC-Co splat formation in an HVOF process using a new beam-shutter device. *Journal of Thermal Spray Technology*, **22**(2-3), 250–262.

Tukey, J. W. (1975). Mathematics and the picturing of data, in: R. D. James (Hrsg.), *Proceedings of the International Congress of Mathematicians*, Band 2, 523–531.

Tukey, J. W. (1977). Modern techniques in data analysis, in: *NSF-sponsored regional research conference at Southeastern Massachusetts University*, North Dartmouth.

Upton, G. J. G. (1980). Contingency table analysis: Log-linear models. *Quality and Quantity*, **14**(1), 155–180.

Upton, G. J. G. und Guillen, M. (1995). Perfect cells, direct models and contingency table outliers. *Communications in Statistics - Theory and Methods*, **24**(7), 1843–1862.

Venables, W. N. und Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, 4. Auflage.

von Eye, A. (2002). *Configural Frequency Analysis: Methods, Models, and Applications*. Lawrence Erlbaum Associates, Mahwah.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, **58**(301), 236–244.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, **26**, 359–372.

Wolf, H. P. (2014). *aplpack: Another Plot PACKage: stem.leaf, bagplot, faces, spin3R, plotsummary, plothulls, and some slider functions*. R package version 1.3.0.

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society, Series B*, **73**(1), 3–36.

Wood, S. N., Pya, N. und Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Erscheint in: Journal of the American Statistical Association*. doi: 10.1080/01621459.2016.1180986.

Xu, Y., Iglewicz, B. und Chervoneva, I. (2014). Robust estimation of the parameters of $g$-and-$h$ distributions, with applications to outlier detection. *Computational Statistics & Data Analysis*, **75**, 66–80.

Zuo, Y. und Serfling, R. (2000). General notions of statistical depth function. *The Annals of Statistics*, **28**(2), 461–482.