Robustheitskonzepte und -untersuchungen für Schätzer konvexer Körper

Dissertation

zur Erlangung des Grades

eines Doktors der Naturwissenschaften

der Universität Dortmund

Dem Fachbereich Statistik

der Universität Dortmund

vorgelegt von

Sebastian Paris Scholz

aus Duisburg

Dortmund 2002

Gutachter: Prof. Dr. U. Gather
Prof. Dr. C. Becker

Tag der mündlichen Prüfung: 16. Dezember 2002

Inhaltsverzeichnis

Kapitel 1

Einleitung

Konvexe Körper, d.h. konvexe und kompakte Teilmengen des d-dimensionalen reellen Raumes $\mathbf{R}^d$, werden beispielsweise zur Strukturerkennung hochdimensionaler Daten benötigt. Zunehmend ist es von Interesse, spezielle konvexe Körper auf Datenbasis zu schätzen und insbesondere Schätzer konvexer Körper bezüglich ihrer Eigenschaften zu untersuchen.
In dieser Arbeit sollen Robustheitsaspekte von Schätzern für konvexe Körper betrachtet werden. Dazu muss der Begriff „Robustheit“ zunächst definiert werden. Dabei wird der Schwerpunkt in der Analyse des Einflusses von Ausreißern auf Schätzungen konvexer Körper im $\mathbf{R}^d$ gelegt. Als Ausreißer werden Beobachtungen bezeichnet, die sich „weit entfernt“ von der Hauptmasse der Daten befinden. In der Praxis sind solche Beobachtungen in vielen Datensätzen zu finden und können somit die Datenanalyse beeinflussen. Es ist daher wünschenswert, damit in Verbindung stehende Effekte auf die Schätzung von konvexen Körpern zu untersuchen, sowie Verfahren zu entwickeln, die ihren Einfluss begrenzen können. Bisher existieren keine Kriterien, die das Bewerten der Schätzungen bei Vorliegen von Ausreißern erlauben.

Im zweiten Kapitel wird daher zur Beurteilung von derartigen Schätzverfahren der von Donoho und Huber (1983) definierte Begriff des finite-sample Bruchpunktes als Robustheitsmaß in geeigneter Weise auf die hier betrachtete Situation übertragen. Die von Donoho und Huber (1983) vorgeschlagene Bruchpunktdefinition für Lokations- und Kovarianzschätzer stellt einen Spezialfall von der in dieser Arbeit eingeführten finite-sample Bruchpunktdefinition für Schätzer konvexer Körper bei geeigneter Parametrisierung dar. Es werden im Weiteren einige theoretische Eigenschaften dieses Bruchpunktkonzeptes hergeleitet. Die in diesem Kapitel vorgestellte Definition ermöglicht die Untersuchung von Schätzern konvexer und kompakter Teilmengen bzgl. ihres Bruchpunktverhaltens.

Ein Beispiel für konvexe Körper sind Zonoide bzw. Liftzonoide von Verteilungen, welche im dritten Kapitel vorgestellt werden. Diese speziellen Körper wurden 1998 von Koshevoy und Mosler „wieder entdeckt". Sie weisen die Besonderheit auf, dass sie dem Erwartungswert eines zufälligen konvexen Körper entsprechen. Sie können somit als Parameter einer Verteilung interpretiert werden. Während Liftzonoide eine Verteilung eindeutig charakterisieren und die Konstruktion von Kontur-Toleranzbreichen erlauben (Koshevoy und Mosler (1997) und (1998)), kann das Volumen des geschätzten Zonoids als Variabilitätsmaß aufgefasst werden, mit dessen Hilfe ein affin äquivarianter multivariater Lokations- bzw. Kovarianzschätzer mit hohem Bruchpunkt konstruiert werden kann (siehe fünftes Kapitel). Die Schätzung von Zonoiden bzw. Liftzonoiden basiert im Allgemeinen auf Polytopen. Unter einem Polytop wird die konvexe Hülle einer endlichen Punktmenge verstanden. Es zeigt sich, dass bei dieser Art der Schätzung schon eine einzelne Beobachtung ausreicht, um die Schätzung im Sinne der im ersten Kapitel vorgestellten Bruchpunktdefinition zusammenbrechen zu lassen.

Im vierten Kapitel wird eine Klasse von konvexen Körper vorgestellt, die so genannten Kontur-Toleranzbereiche. Diese Bereiche erlauben Aussagen über Struktur und Variabilität der zugrunde liegenden Verteilung F. Die in der Literatur zu findenden Kontur-Toleranzbereiche unterscheiden sich im Wesentlichen durch das zugrunde gelegte Zentrum der Verteilung F. Die Idee der Konturtoleranz-Bereiche ist auf Tukey (1975) zurückzuführen. Zusätzlich definiert Tukey (1975) erstmals einen speziellen

Kontur-Toleranzbereich, den so genannten Halfspace-Bereich. Weitere Beispiele von Kontur-Toleranzbereichen werden in den Arbeiten von Liu (1990), Liu (1992) und Fraiman, Liu und Meloche (1997) vorgestellt. Anwendung finden diese beispielsweise in der Prozesskontrolle und Qualitätssicherung (Liu (1995)), aber auch in der nicht-parametrischen Statistik. So ist der Begriff des Kontur-Toleranzbereiches eng mit der so genannten Datentiefe-Funktion verbunden. Diese Funktion misst den Abstand eines Punktes $x \in \mathbb{R}^d$ vom Zentrum der Verteilung und erlaubt die Konstruktion von multivariaten Rang- und Ordnungsstatistiken (siehe z.B. Liu, Parelius und Singh (1999) oder Donoho und Gasko (1992)). Eine Zusammenstellung verschiedener Kontur-Toleranzbereiche und zugehöriger Datentiefe-Funktionen kann in den Arbeiten von Liu, Parelius und Singh (1999) und Zuo und Serfling (2000) gefunden werden. In diesem Kapitel wird eine Definition von Kontur-Toleranzintervallen für unimodale Verteilungen vorgeschlagen. Aus dieser resultiert auf natürliche Weise die Datentiefe-Funktion, die die von Zuo und Serfling (2000) geforderten Eigenschaften erfüllt. Im Anschluss werden zwei Kontur-Toleranzbereiche vorgestellt. Liu (1992) schlägt die so genannten Mahalanobis-Bereiche vor. Diese konvexen Körper beruhen auf den ersten beiden Momenten der zugrundeliegenden Verteilung. Eine Schätzung dieser konvexen Körper ergibt sich durch die Schätzung der entsprechenden Momente. Des Weiteren wird ein Kontur-Toleranzbereich eingeführt, der eng mit dem im dritten Kapitel vorgestellten Liftzonoiden verbunden ist, den so genannten zonoiden Zonen (Koshevoy und Mosler (1997)). Eigenschaften und eine erste Schätzung dieser Körper mittels Polytopen werden ebenfalls erläutert. Die Untersuchung bezüglich des Bruchpunktes beschränkt sich bisher bei Kontur-Toleranzbereichen nur auf die Schätzung des Zentrums, also auf einen Punkt im $\mathbb{R}^d$. Es zeigt sich, dass der Bruchpunkt der Schätzung der Mahalanobis-Bereiche von den gewählten Lokations- und Kovarianzschätzern abhängt. Die Schätzung der zonoiden Zonen ist gegenüber Ausreißern sehr anfällig, da schon eine einzelne Beobachtung ausreicht, um die Schätzung im Sinne der im zweiten Kapitel vorgeschlagenen Bruchpunktdefinition zusammen brechen zu lassen.

Das in dieser Arbeit untersuchte Bruchpunktverhalten von Schätzern legt nahe, vorhandene Schätzer konvexer Körper soweit zu modifizieren, dass sie einen höheren Bruchpunkt aufweisen. Dieser Schritt erfolgt im fünften Kapitel. Dazu wird aus einer gegebenen Stichprobe eine geeignete Teilstichprobe bestimmt, die unter allen zulässigen Teilstichproben ein festgelegtes Variabilitätsmaß minimiert. Die so erhaltene Teilstichprobe wird zur Schätzung der interessierenden konvexen Körper verwendet. Dieses Vorgehen geht auf Rousseeuw (1985) zurück, der Lokations- und Kovarianzschätzer mit hohem Bruchpunkt basierend auf den so genannten MCD- und MVE-Kriterien vorschlägt. In dieser Arbeit wird ein weiteres Kriterium benutzt, basierend auf dem Volumen des geschätzen Zonoids (MZE-Kriterium). Das Volumen des geschätzen Zonoids kann als Variabilitätsmaß aufgefasst werden und erlaubt die Konstruktion von affin äquivarianten Lokations- bzw. Kovarianzschätzer mit hohem Bruchpunkt. Falls die Schätzung eines konvexen Körpers mittels der auf der Grundlage dieses Kriterium resultierenden Teilstichproben durchgeführt wird, so weist diese ebenfalls einen hohen Bruchpunkt auf.

Im sechsten Kapitel erfolgt eine Zusammenfassung der in dieser Arbeit erzielten Ergebnisse und ein Ausblick auf weitergehende Forschung.

Kapitel 2

Robustifizierungskonzepte für Schätzungen konvexer Körper

Beobachtungen, die „weit entfernt“ von der Hauptmasse der Daten liegen (Ausreißer), können einen erheblichen Effekt auf das Ergebnis einer Schätzung haben. Um die Auswirkungen derartiger Beobachtungen auf Schätzungen konvexer Körper untersuchen zu können, werden geeignete Bewertungskriterien benötigt. In der Literatur existieren bislang noch keine Maßzahlen, die diesen Effekt bewerten. In diesem Kapitel wird daher zur Beurteilung der Robustheitseigenschaften von derartigen Schätzern der von Donoho und Huber (1983) geprägte Begriff des finite-sample Bruchpunktes als Robustheitsmaß in geeigneter Weise auf die hier betrachtete Situation übertragen.

2.1 Annahmen und Notation

Ziel dieses Kapitels ist es, kompakten und konvexen Mengen Maßzahlen zuzuordnen und deren Eigenschaften zu untersuchen. Zur Übertragung von Robustheitsaspekten auf die hier betrachtete Situation werden zunächst einige Annahmen getroffen. Seien hierzu $X_1, ..., X_n$ u.i.v. R^d Zufallsvektoren (ZV) mit Verteilung F. Es wird angenommen, dass F einer Klasse von Verteilungen $\mathcal{F}$ angehört. Diese kann in der folgender

Form geschrieben werden kann:

$$\mathcal{F} = \{F_\theta : \theta \in \Theta\}, \text{ mit geeignetem Parameterraum } \Theta,$$

wobei Θ genauer zu spezifizieren ist.

Beispiel 2.1
1. Die Klasse aller d-dimensionalen Normalverteilungen mit bekannter positiv definiter Kovarianzmatrix $\Sigma \in \mathbf{R}^{d\times d}$ und unbekanntem Erwartungswert $\mu \in \mathbf{R}^d$ kann dargestellt werden durch $\mathcal{N}_1 = \{N_{\mu,\Sigma} : \mu \in \Theta\}$, wobei $\Theta = \mathbf{R}^d$.
2. Wird $\Theta = \mathbf{R}^d \times \{\Sigma\}_{p.d.}$ betrachtet, wobei $\{\Sigma\}_{p.d.}$ die Menge aller positiv definiten (p.d.) $d \times d$- Matrizen bezeichne, dann wird durch $\mathcal{N}_2 = \{N_{\mu,\Sigma} : (\mu, \Sigma) \in \Theta\}$ die Klasse aller d-dimensionalen Normalverteilungen mit unbekannter Kovarianzmatrix $\Sigma \in \mathbf{R}^{d\times d}$ und unbekanntem Erwartungswert $\mu \in \mathbf{R}^d$ eindeutig beschrieben.

In dieser Arbeit erweist es sich als zweckmäßig, eine andere Form der Parametrisierung zu wählen. Während in Beispiel 2.1 wird die interessierende Verteilungsklasse durch ihre Parameter beschrieben wird, kann eine Verteilungsklasse auch durch ihre Konturen dargestellt werden.

Beispiel 2.2 Sei $\Theta = \Theta_r = \Big\{K := \{x \in \mathbf{R}^d : (x-\mu)^T\Sigma^{-1}(x-\mu) \le r^2\} : \mu \in \mathbf{R}^d, \Sigma \in \mathbf{R}^{d\times d} \text{ p.d. } \Big\}$, $r > 0$ bekannt. Die Elemente von Θ sind somit Ellipsen. Sie entsprechen konvexen und kompakten Teilmengen des $\mathbf{R}^d$ (konvexen Körpern). Die Klasse der d-dimensionalen Normalverteilungen kann alternativ zu 2.1 auch dargestellt werden als $\mathcal{N}_2 = \Big\{N_K : K \in \Theta\Big\}$. Jedes $K \in \Theta$ entspricht dabei natürlich eineindeutig einem Paar (μ, Σ).

Es bezeichne $\mathcal{K}^d$ die Menge der konvexen Körper im $\mathbf{R}^d$, d.h. die Menge der nichtleeren konvexen und kompakten Teilmengen des $\mathbf{R}^d$. Im Folgenden wird angenommen, dass eine Menge $\Theta \subseteq \mathcal{K}^d$ existiert, so dass die interessierende Verteilungsklasse in folgender Form dargestellt werden kann: $\mathcal{F} = \{F_K : K \in \Theta\}$.
Die Elemente des Parameterraumes Θ entsprechen also konvexen Körpern.
Es wird somit davon ausgegangen, dass ein Funktional existiert, d.h. eine Abbildung :

$$T : \mathcal{F} \longrightarrow \mathcal{K}^d, \text{ so dass } T(F) = K \in \Theta \text{ und } F \in \mathcal{F}. \tag{2.1}$$

Dabei sei $\mathcal{F}$ die Menge aller Verteilungen, für die das Funktional T definiert ist. Weiterhin wird ein Funktional T auf $\mathcal{F}$ Fisher-konsistent genannt, falls gilt:

$$T(F) = T(F_K) = K \text{ für alle } K \in \Theta \subseteq \mathcal{K}^d.$$

Eine naheliegende Schätzung von $T(F) \in \mathcal{K}^d$ ist durch $T(F_n)$ gegeben. Dabei bezeichnet F_n die empirische Verteilung der $\mathbf{R}^d$-ZV $X_1, ..., X_n$, wobei $X_i \sim F$, $i = 1, ..., n$.
Um einen Konvergenzbegriff für Schätzer konvexer Körper einzuführen, wird hier ein geeigneter Abstand definiert. Dies geschieht mittels des so genannten Hausdorff Abstandes. Dieser ist für zwei kompakte Körper K_1 und $K_2 \in \mathcal{K}^d$ wie folgt definiert:

$$d_H(K_1, K_2) = \max\left\{\sup_{x_1 \in K_1}\left\{\inf_{x_2 \in K_2} d_e(x_1, x_2)\right\}, \sup_{x_2 \in K_2}\left\{\inf_{x_1 \in K_1} d_e(x_1, x_2)\right\}\right\}. \qquad (2.2)$$

Dabei entspricht $d_e(x_1, x_2) = ||x_1 - x_2||$ dem euklidischen Abstand zwischen den Punkten x_1 und x_2 im $\mathbf{R}^d$.

Bemerkung 2.1 1. Der Hausdorff Abstand besitzt die folgenden Eigenschaften:

$$\begin{aligned} d_H(K_1, K_1) &= 0, \;\; d_H(K_1, K_2) > 0 \;\; (K_1 \neq K_2), \\ d_H(K_1, K_2) &= d_H(K_2, K_1), \\ d_H(K_1, K_2) &\leq d_H(K_1, K_3) + d_H(K_2, K_3), \end{aligned}$$

wobei $K_i \in \mathcal{K}^d$, $i = 1, 2, 3$. Somit wird die Menge der konvexen Körper $\mathcal{K}^d$ durch den Hausdorff Abstand d_H zu einem metrischen Raum (siehe z.B. Valentine (1964)).
2. Der Abstand zwischen einer konvexen Menge K und dem Ursprung $\{\mathbf{0}\}$ ist gegeben durch:
$d_H(K, \{\mathbf{0}\}) = \max\{||\mathrm{x}|| : \mathrm{x} \in \mathrm{K}\} =: ||\mathrm{K}||$ und wird Norm der konvexen Menge K genannt.
3. Im Spezialfall einelementiger Mengen K_1 und K_2 stimmt der Hausdorff Abstand mit dem euklidischen Abstand überein.

Daraus lässt sich folgender Konvergenzbegriff für Schätzer konvexer Körper ableiten.

Definition 2.1 *Für $\mathbb{R}^d$ ZV $X_i \sim F$, $i = 1, ..., n$ heißt $T(F_n)$ konvergent in Wahrscheinlichkeit bezüglich des Hausdorff Abstandes gegen $T(F)$ genau dann, wenn*

$$\lim_{n\to\infty} P\Big(d_H\Big(T(F_n), T(F)\Big) = 0\Big) = 1. \tag{2.3}$$

Die in diesem Abschnitt vorgestellten Begriffe sollen anhand eines Beispiels erläutert werden.

Beispiel 2.3 Sei Θ wie in Beispiel 2.2 mit $r = 1$ und $\mathcal{N}_2 = \Big\{N_K : K \in \Theta\Big\}$ bezeichne die Klasse der multivariaten Normalverteilungen. Gegeben seien $X_1, ..., X_n$ u.i.v. $\mathbb{R}^d$-ZV mit $X_i \sim F \in \mathcal{N}_2$. Sei F_n die empirische Verteilung von $X_1, ..., X_n$. Es sei der unbekannte konvexe Körper $K = \{x \in \mathbb{R}^d : (x-\mu)^T\Sigma^{-1}(x-\mu) \leq 1\}$ zu schätzen. Dann ist eine Schätzung von K gegeben durch: $T(F_n) = \Big\{x \in \mathbb{R}^d : (x-\mu_n)^T C_n^{-1}(x-\mu_n) \leq 1\Big\}$, wobei $\mu_n = \frac{1}{n}\sum_{i=1}^n X_i$ ein Schätzer für μ und $C_n = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_n)(X_i - \mu_n)^T$ ein Schätzer für Σ ist.

Im Folgenden werden nur noch Schätzer T für konvexe Körper betrachtet, die Formel (2.3) erfüllen, also konvergent in Wahrscheinlichkeit gegen den wahren Wert $T(F)$ für alle $F \in \mathcal{F}$ sind.
Da die in dieser Arbeit untersuchten Schätzer im gleichen Maße von der empirischen Verteilung der Stichprobe wie von der Stichprobe selbst abhängen, wird folgende Schreibweise festgelegt: $T(\tilde{X}_n) := T(F_n) \in \mathcal{K}^d$. Weiterhin wird als Kurzschreibweise für einelementige Mengen $b := \{b\} \in \mathbb{R}^d$ eingeführt.

Die Äquivarianz gegenüber affinen Transformationen stellt eine natürliche Forderung an Schätzer dar. Oft erscheint es jedoch sinnvoll, schwächere Forderungen wie die Invarianz gegenüber Translationen zu fordern. Für Schätzer konvexer Körper ist Translationsäquivarianz wie folgt definiert:

Definition 2.2 *Sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine beliebige Stichprobe vom Umfang n mit $x_i \in \mathbb{R}^d$, $i = 1, ..., n$ aus einer Verteilung $F \in \mathcal{F}$. Für einen Vektor $b \in \mathbb{R}^d$ bezeichne*

$\tilde{X}_n + b$ *die Stichprobe* $\{x_1 + b, ..., x_n + b\}$. *Dann heißt* $T(\tilde{X}_n)$ *translationsäquivariant, falls für alle* $b \in \mathbb{R}^d$ *und jede Stichprobe* $\tilde{X}_n$ *gilt:*

$$T(\tilde{X}_n + b) = T(\tilde{X}_n) \oplus b.$$

Dabei bezeichnet $\oplus$ *die so genannte Minkowski-Summe. Diese ist für beliebige Teilmengen* K_1 *und* K_2 *des* $\mathbb{R}^d$ *erklärt durch:* $K_1 \oplus K_2 = \{x + y : x \in K_1, y \in K_2\}$.
Ein Schätzer $T(\tilde{X}_n)$ *heißt affin äquivariant, falls für jede nichtsinguläre Matrix* $A \in \mathbb{R}^{d \times d}$, *einen Vektor* $b \in \mathbb{R}^d$ *und die Stichprobe* $A\tilde{X}_n + b := \{Ax_1 + b, ..., Ax_n + b\}$ *gilt:*

$$T(A\tilde{X}_n + b) = AT(\tilde{X}_n) \oplus b, \text{ wobei } AT(\tilde{X}_n) := \{Ax : x \in T(\tilde{X}_n)\}.$$

Die affine Äquivarianz impliziert somit die Translationsäquivarianz eines Schätzers.

2.2 Finite-sample Bruchpunkte für Schätzer konvexer Körper

Für die Beurteilung der Robustheit eines Schätzers existiert unter anderem das von Donoho und Huber (1983) eingeführte Kriterium des finite-sample Bruchpunktes. Diese Maßzahl gibt an, welcher Anteil der Beobachtungen einer Stichprobe durch beliebige Werte ersetzt werden muss, um den Zusammenbruch eines Schätzers im geeigneten Sinne herbeiführen zu können. Im Folgenden sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine Stichprobe vom Umfang n, mit $x_i \in \mathbb{R}^d$. Für Lokationsschätzer definieren Donoho und Huber (1983) den finite-sample Bruchpunkt wie folgt:

Definition 2.3 *Sei* $\tilde{X}_n = \{x_1, ..., x_n\}$, $x_i \in \mathbb{R}^d$ *eine Stichprobe von* n *Beobachtungen aus einer Verteilung* $F_\theta \in \mathcal{F} = \{F_\theta : \theta \in \Theta\}$ *mit* $\Theta = \mathbb{R}^d$. *Sei* $\tilde{Y}_{n,k} := \{x_{i_1}, ..., x_{i_N}, y_1, ..., y_k\}$, $y_j \in \mathbb{R}^d$, $j = 1, ..., k$, $N = n - k$, *eine durch Austausch von* k *Beobachtungen von* $\tilde{X}_n$ *durch beliebige Punkte des* $\mathbb{R}^d$ *entstandene Stichprobe. Der finite-sample Bruchpunkt eines Lokationsschätzers* $T(\tilde{X}_n) \in \mathbb{R}^d$ *für den Parameter* $\theta \in \mathbb{R}^d$, *ist wie folgt definiert:*

$$\epsilon_L(\tilde{X}_n, T) = \min \left\{ \frac{k}{n} : \sup_{\tilde{Y}_{n,k}} d_e\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big) = \infty \right\},$$

wobei $d_e\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big) = ||T(\tilde{X}_n) - T(\tilde{Y}_{n,k})||.$

Definition 2.3 ist nur sinnvoll, wenn der betrachtete Schätzer beliebige Werte im $\mathbb{R}^d$ annehmen kann. Eine Übertragung der Definition auf andere Schätzer ist möglich, zum Beispiel auf Kovarianzschätzer (Lopuhäa und Rouseeuw (1991)).

Definition 2.4 *Sei* $\tilde{X}_n = \{x_1, ..., x_n\}$, $x_i \in \mathbb{R}^d$ *eine Stichprobe von* n *Beobachtungen aus einer Verteilung* $F_\theta \in \mathcal{F} = \{F_\theta : \theta \in \Theta\}$ *mit* $\Theta = \mathbb{R}^d \times \{\Sigma\}_{p.d.}$. *Zu einer symmetrischen Matrix* $A \in \mathbb{R}^{d\times d}$ *seien die der Größe nach geordneten Eigenwerte gegeben durch* $\lambda_d(A) \geq ... \geq \lambda_1(A)$. *Sei* $\tilde{Y}_{n,k}$ *wie in Definition 2.3 und* $C(\tilde{X}_n) \in \mathbb{R}^{d\times d}$ *ein Kovarianzschätzer für* Σ, *basierend auf der Stichprobe* $\tilde{X}_n$. *Dann heißt:*

$$\epsilon_C(\tilde{X}_n, C) = \min_{1\leq k\leq n} \left\{ \frac{k}{n} : \sup_{\tilde{Y}_{n,k}} \tilde{d}\Big(C(\tilde{X}_n), C(\tilde{Y}_{n,k})\Big) = \infty \right\},$$

wobei

$$\tilde{d}\Big(C(\tilde{X}_n), C(\tilde{Y}_{n,k})\Big) : \;=\; \max\left\{ \Big|\lambda_d(C(\tilde{X}_n)) - \lambda_d(C(\tilde{Y}_{n,k}))\Big|, \left|\frac{1}{\lambda_1(C(\tilde{X}_n))} - \frac{1}{\lambda_1(C(\tilde{Y}_{n,k}))}\right| \right\},$$

der finite-sample Bruchpunkt des Kovarianzschätzers C_n.

Bemerkung 2.2 1. Vom Zusammenbruch eines Schätzers $C_n := C(\tilde{X}_n)$ der Kovarianzmatrix Σ wird demnach gesprochen, wenn der größte Eigenwert der Kovarianzmatrix $C_{n,k} = C(\tilde{Y}_{n,k})$ unendlich (Explosion) oder der kleinste Null (Implosion) ist.
2. Der Zusammenbruch in Definition 2.4 wird durch die Spektralnorm erklärt (siehe z.B. Schott (1997), *S.*157). Da die betrachteten Kovarianzmatrizen C_n symmetrisch sind, entspricht die Spektralnorm dem Spektralradius einer Matrix, also dem betragsmäßig größten Eigenwert. Dieser ist definiert als: $||C_n||_2 = |\lambda_d(C_n)|$. Die Spektralnorm erfüllt die Eigenschaften einer Matrixnorm. Durch $||C_n - C_{n,k}||_2$ ist eine Semimetrik aber keine Metrik auf der Menge aller p.d. Matrizen gegeben. Wird dieses Abstandsmaß gewählt, so wird dann von einem Zusammenbruch des Kovarianzschätzers gesprochen werden, falls der „Abstand“ zwischen C_n und $C_{n,k}$ oder der „Abstand“ zwischen C_n^{-1} und $C_{n,k}^{-1}$ beliebig groß wird.

In den in der Literatur zu findenden Definitionen wird der Zusammenbruch eines Kovarianzschätzers bisher nicht über ein geeignetes Abstandsmaß definiert. Es gibt mehrere „sinnvolle“ Möglichkeiten, den Bruchpunkt zu definieren (siehe z.B. Zuo (2001)).
Dies gilt ebenso bei der Definition des finite-sample Bruchpunktes für die simultane Schätzung des Erwartungswerts μ und der Kovarianz Σ. Im Folgenden werden zwei in der Literatur zu findende Bruchpunktdefinitionen vorgestellt und diskutiert.

Definition 2.5 *Sei* $\tilde{X}_n = \{x_1, ..., x_n\}$, $x_i \in \mathbb{R}^d$ *eine Stichprobe von* n *Beobachtungen aus einer Verteilung* $F_\theta \in \mathcal{F} = \{F_\theta : \theta \in \Theta\}$ *mit* $\Theta = \mathbb{R}^d \times \{\Sigma\}_{p.d.}$. *Sei* $\tilde{Y}_{n,k}$ *wie in Definition 2.3 und* $\left\{T(\tilde{X}_n), C(\tilde{X}_n)\right\}$ *ein Schätzer für das Paar* $\theta = \{\mu, \Sigma\}$ *basierend auf der Stichprobe* $\tilde{X}_n$.
a) Dann definiert Davies (1987) den finite-sample Bruchpunkt für $\left\{T(\tilde{X}_n), C(\tilde{X}_n)\right\}$ *wie folgt:*

$$\epsilon_D(\tilde{X}_n, (T, C)) = \min_{1 \le k \le n} \left\{ \frac{k}{n} : \sup_{\tilde{Y}_{n,k}} B_D\Big(\left\{T(\tilde{Y}_{n,k}), C(\tilde{Y}_{n,k})\right\} \Big) = \infty \right\},$$

wobei

$$B_D\Big(\left\{T(\tilde{Y}_{n,k}), C(\tilde{Y}_{n,k})\right\} \Big) = ||T(\tilde{Y}_{n,k})|| + \sum_{i=1}^{d} \Big(\lambda_i(C(\tilde{Y}_{n,k})) + \lambda_i(C(\tilde{Y}_{n,k}))^{-1} \Big).$$

b) Tyler (1994) definiert den Bruchpunkt für $\left\{T(\tilde{X}_n), C(\tilde{X}_n)\right\}$ *wie folgt:*

$$\epsilon_T(\tilde{X}_n, (T, C)) = \min_{1 \le k \le n} \left\{ \frac{k}{n} : \sup_{\tilde{Y}_{n,k}} B_T\Big(\left\{T(\tilde{Y}_{n,k}), C(\tilde{Y}_{n,k})\right\} \Big) = \infty \right\},$$

wobei

$$\begin{aligned} B_T\Big(\left\{T(\tilde{Y}_{n,k}), C(\tilde{Y}_{n,k})\right\} \Big) \quad = \quad & \max \Big\{ ||C^{-1/2}(\tilde{X}_n) \left(T(\tilde{X}_n) - T(\tilde{Y}_{n,k}) \right) ||, \\ & tr \left(C(\tilde{X}_n) C(\tilde{Y}_{n,k})^{-1} + C(\tilde{X}_n)^{-1} C(\tilde{Y}_{n,k}) \right) \Big\}, \end{aligned}$$

wobei $tr(A)$ *die Spur der Matrix* A *bezeichnet.*

Bemerkung 2.3 Anstatt die Verzerrung zwischen der Schätzung mittels der ursprünglichen Stichprobe und der kontaminierten Stichprobe zu betrachten, wird in

der von Davies (1987) vorgeschlagenen Definition der Zusammenbruch nur durch die kontaminierte Stichprobe $\tilde{Y}_{n,k}$ erklärt. Der Schätzer bricht demnach zusammen, wenn der durch die kontaminierte Stichprobe erhaltene Lokationsschätzer bzw. der größte Eigenwert der geschätzten Kovarianz den Wert unendlich annehmen oder wenn der durch die kontaminierte Stichprobe berechnete kleinste Eigenwert der geschätzten Kovarianz Null ist.
Auch nach Tylers Definition (1994) bricht der Schätzer $\left\{T(\tilde{X}_n), C(\tilde{X}_n)\right\}$ zum einem zusammen, wenn $||T(\tilde{X}_n) - T(\tilde{Y}_{n,k})||$ beliebig groß wird zum anderen, wenn der auf der kontaminierten Stichprobe beruhende größte Eigenwert der geschätzten Kovarianzmatrix den Wert unendlich annimmt oder kleinste Eigenwert der geschätzten Kovarianzmatrix Null ist (Becker (2001)).
Beide vorgestellten Bruchpunkt Definitionen führen zum selben Wert des Bruchpunkts für $\left\{T(\tilde{X}_n), C(\tilde{X}_n)\right\}$.
Bei beiden Definitionen basiert der Zusammenbruch auf einer Semimetrik.

In den vorgestellten Definitionen wird der Zusammenbruch der Schätzung durch das Ersetzen von Beobachtungen herbeigeführt. Daher wird dieser Bruchpunkt in der Literatur finite-sample „Ersetzungs-Bruchpunkt“ genannt. Ein anderer Ansatz, bei dem zu einer gegebenen Stichprobe Beobachtungen hinzugefügt werden, so dass ein Zusammenbruch des Schätzers erreicht wird, ist der so genannten „Hinzufügungs-Bruchpunkt“. Zuo (2001) untersucht den Zusammenhang zwischen diesen beiden Definitionen für Lokations- und Kovarianzschätzer. Für den Fall, dass beide Bruchpunkte wohl definiert sind, ist es unter bestimmten Voraussetzungen möglich, den Wert des Hinzufügungs-Bruchpunkts zu ermitteln, wenn der des Ersetzungs-Bruchpunktes bekannt ist und umgekehrt. Es ist daher in vielen Fällen nicht entscheidend, auf welchem der beiden Ansätze die Definition des finite-sample Bruchpunktes beruht.
Eigenschaften der Bruchpunktdefinitionen von Lokations- und Kovarianzschätzern werden in den Arbeiten von Davies (1987) und Lopuhäa und Rousseeuw (1991) untersucht. In diesen Arbeiten werden unter anderem die kleinsten oberen Schranken für den finite-

sample Bruchpunkt einiger Schätzerklassen hergeleitet.

2.2.1 Finite-sample Explosions-Bruchpunktdefinition

Im Folgenden soll die Definition des finite-sample Bruchpunktes auf Schätzer konvexer Körper erweitert werden.
Dazu sind zunächst Überlegungen notwendig, was unter dem Zusammenbruch eines Schätzers eines konvexen Körpers verstanden werden soll.
In den Definitionen 2.3 und 2.4 brechen die Schätzer zusammen, wenn diese an den Rand des zugehörigen Parameterraumes geschoben werden. Werden nun Schätzer konvexer Körper betrachtet, so sollte in Analogie von einem Zusammenbruch gesprochen werden, wenn der Schätzer gegen den Rand des zugehörigen Parameterraumes rückt. Dies geschieht, wenn der Schätzer eines konvexen Körpers zu einer konvexen nicht beschränkten Teilmenge des $\mathbb{R}^d$ degeneriert. Somit kann von einem Zusammenbruch gesprochen werden, wenn der Hausdorff Abstand zwischen dem auf der kontaminierten Stichprobe basierenden Schätzwert und dem auf der regulären Stichprobe basierenden Schätzwert beliebig groß wird. Dies führt zur folgenden Definition:

Definition 2.6 *Sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine Stichprobe vom Umfang n mit $x_i \in \mathbb{R}^d$ von Beobachtungen aus einer Verteilung $F_K \in \mathcal{F} = \{F_K : K \in \Theta\}$, wobei $\Theta \subseteq \mathcal{K}^d$. Sei $\tilde{Y}_{n,k}$ wie in Definition 2.3 und $T(\tilde{X}_n)$ ein Schätzer für den konvexen Körper K. Dann heißt:*

$$\epsilon_K(\tilde{X}_n, T) = \min_{1 \le k \le n} \left\{ \frac{k}{n} : \sup_{\tilde{Y}_{n,k}} d_H\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big) = \infty \right\},$$

der finite-sample (Explosions-)Bruchpunkt für Schätzungen konvexer Körper, wobei $d_H\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big)$ den Hausdorff Abstand aus (2.2) bezeichnet.

Die Stützfunktion eines konvexen Körpers

Zur Bruchpunktberechnung von Schätzern konvexer Körper bietet es sich an, den Hausdorff Abstand mit Hilfe der so genannten Stützfunktion darzustellen.

Die Stützfunktion erweist sich als ein nützliches Hilfsmittel zur Untersuchung von konvexen und kompakten Mengen. Durch sie kann ein konvexer Körper eindeutig charakterisiert werden.

Definition 2.7 *Zu einem gegebenen konvexen Körper $K \in \mathcal{K}^d$ ist die Stützfunktion $h(K, \cdot) : \mathbb{R}^d \to \mathbb{R}$ definiert als:*

$$h(K,p) = \max\Big\{\langle x, p\rangle : x \in K\Big\}, \quad p \in \mathbb{R}^d,$$

dabei bezeichne $\langle \cdot, \cdot \rangle$ das Skalarprodukt.

Somit ist die Stützfunktion eine Abbildung, die jedem $p \in \mathbf{R}^d$ in Abhängigkeit von K ein Element aus den reellen Zahlen zuordnet.

Bemerkung 2.4 1. Die Funktion $h(K, \cdot)$ ist stetig, konvex und homogen vom Grad Eins, d.h. $h(K, t \cdot p) = t \cdot h(K, p)$, für $t \in \mathbf{R}_+$. Demnach ist die Stützfunktion bereits durch ihre Werte für alle p mit $||p|| = 1$ eindeutig festgelegt.
2. Für $p_i \in \mathbf{R}^d$, $i = 1, 2$ gilt: $h(K, p_1 + p_2) \leq h(K, p_1) + h(K, p_2)$. Enthält der konvexe Körper den Ursprung in seinem Inneren, so gilt $h(K, p) > 0$, für $p \neq \mathbf{0}$.
3. Für zwei konvexe und kompakte Mengen K_1 und K_2 gilt:

1. $h(K_1 \oplus K_2, \cdot) = h(K_1, \cdot) + h(K_2, \cdot)$,
2. $h(K_1, \cdot) < h(K_2, \cdot)$ genau dann, wenn $K_1 \subset K_2$,
3. $h(K_1, \cdot) = h(K_2, \cdot)$ genau dann, wenn $K_1 = K_2$.

Die Stützfunktion h einer kompakten konvexen Menge K des $\mathbf{R}^d$ lässt sich folgendermaßen interpretieren:
Für alle Punkte x des konvexen Körpers K und für ein beliebiges aber festes p gilt nach Definition von h:

$$\langle x, p\rangle \leq h(K, p). \tag{2.4}$$

Für festes p ist $h(K,p)$ demnach das Maximum von $\langle x,p\rangle$, wenn x alle Punkte von K durchläuft.

Nun gilt:

Lemma 2.1 Jede stetige, konvexe und homogene Funktion vom Grad Eins $h : \mathbb{R}^d \to \mathbb{R}$ entspricht der Stützfunktion eines konvexen Körpers $K \in \mathcal{K}^d$. Der konvexe Körper K wird durch diese Funktion eindeutig bestimmt und ist gegeben durch die Menge:

$$K = \left\{x \in \mathbb{R}^d : \langle x,p\rangle \leq h(p)\right\}, \text{ für alle } p \in \mathbb{R}^d \text{ mit } ||p|| = 1.$$

Beweis: (Valentine (1968))

q.e.d.

Weiterhin gelten die folgenden nützlichen Zusammenhänge.

Lemma 2.2 Es seien $K_i \in \mathcal{K}^d$ mit zugehörigen Stützfunktionen $h(K_i,\cdot)$, $i = 1,...,n$. Dann gilt:

a) Die Stützfunktion von $K = \oplus_{i=1}^n K_i$ berechnet sich als

$$h(K,\cdot) \quad = \quad \sum_{i=1}^{n} h(K_i,\cdot). \tag{2.5}$$

b) Die Stützfunktion von $K = conv\,(\bigcup_{i=1}^n K_i)$ berechnet sich als

$$h(K,\cdot) \quad = \quad \max_{1\leq i\leq n} h(K_i,\cdot), \tag{2.6}$$

dabei ist $conv(K)$ die konvexe Hülle der Menge K.

Beweis: (Leichtweiß (1979))

q.e.d.

Bemerkung 2.5 Es kann gezeigt werden (siehe Leichtweiß (1979), S. 148), dass für zwei konvexe Körper $K_1, K_2 \in \mathcal{K}^d$ gilt:

$$d_H(K_1,K_2) \quad = \quad \max_{||p||=1} |h(K_1,p) - h(K_2,p)|. \tag{2.7}$$

Somit lässt sich der Hausdorff Abstand zwischen zwei konvexen Körpern als maximale absolute Differenz zwischen den Stützfunktionen beider Körper über alle $||p|| = 1$ bestimmen.

Der Zusammenbruch eines Schätzers eines konvexen Körpers wurde über den Hausdorff Abstand definiert. Ein Schätzer bricht demnach zusammen, falls der Hausdorff Abstand der Schätzungen des konvexen Körpers auf der Grundlage der regulären Stichprobe und der Schätzung auf der Grundlage der kontaminierten Stichprobe beliebig groß wird. Die vorgestellten Bruchpunktdefinition erlaubt die Untersuchung des Bruchpunktverhaltens von Schätzern konvexer Körper. Das folgende Beispiel verdeutlicht zunächst den Zusammenhang des Bruchpunktverhaltens zwischen der Schätzung einer Ellipse und dem Paar $\left\{T(\tilde{X}_n), C(\tilde{X}_n)\right\}$.

Beispiel 2.4 Sei Θ wie in Beispiel 2.2. Es bezeichne $\mathcal{N}_2 = \left\{N_K : K \in \Theta\right\}$ die Klasse der multivariaten Normalverteilungen. Gegeben sei eine Stichprobe $\tilde{X}_n$ vom Umfang n, wobei $x_i \in \mathbb{R}^d$ Realisationen der Verteilung $F \in \mathcal{N}_2$. Es sei der unbekannte konvexe Körper $K = \{x \in \mathbb{R}^d : (x-\mu)^T\Sigma^{-1}(x-\mu) \leq r^2\}$, $r > 0$ bekannt, zu schätzen. Es wird die folgende Schätzung von K betrachtet:

$$T(\tilde{X}_n) = \left\{x \in \mathbb{R}^d : (x-\mu_n)^T C_n^{-1}(x-\mu_n) \leq r^2\right\},$$

wobei μ_n ein affin äquivarianter Lokationsschätzer für μ ist und $C_n \in \mathbb{R}^{d\times d}$ ein nichtsingulärer affin äquivarianter Kovarianzschätzer für Σ ist.
Es gilt:

$$\begin{aligned} T(\tilde{X}_n) &= \left\{x \in \mathbb{R}^d : (x-\mu_n)^T C_n^{-1}(x-\mu_n) \leq r^2\right\} \\ &= C_n^{1/2}\left\{x \in \mathbb{R}^d : \langle x, x\rangle \leq r^2\right\} \oplus \mu_n \\ &= C_n^{1/2} B(\mathbf{0}, r) \oplus \mu_n, \end{aligned}$$

wobei $A^{1/2} \in \mathbb{R}^{d\times d}$ die Quadratwurzel der nichtsingulären Matrix A und $B(\mathbf{0}, r)$ eine Kugel um den Ursprung $\mathbf{0}$ mit Radius r bezeichnet. Diese Kugel hängt nicht von der Stichprobe $\tilde{X}_n$ ab.
Zur Bestimmung des Bruchpunktes wird die Stützfunktion von $T(\tilde{X}_n)$ benötigt. Diese ist gegeben durch:

$$h(T(\tilde{X}_n), p) \quad = \quad h(C_n^{1/2} B(\mathbf{0}, r) \oplus \mu_n, p)$$

$$
\begin{aligned}
&\overset{(2.5)}{=} && h(C_n^{1/2}B(\mathbf{0},r),p) + h(\mu_n,p) \\
&= && ||C_n^{1/2}p||h\left(B(\mathbf{0},r),\frac{(C_n^{1/2})^T p}{||C_n^{1/2}p||}\right) + \langle \mu_n, p\rangle \\
&= && r||C_n^{1/2}p|| + \langle \mu_n, p\rangle,
\end{aligned}
$$

da $C_n^{1/2} = (C_n^{1/2})^T$ und $h(B(\mathbf{0},r),p) = r||p||$. Damit berechnet sich der finite-sample Bruchpunkt von $T(\tilde{X}_n)$ als:

$$
\epsilon_K(\tilde{X}_n, T) = \min_{1\le k\le n}\left\{\frac{k}{n} : \sup_{\tilde{Y}_{n,k}} d_H\left(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\right) = \infty\right\}, \text{ wobei}
$$

$$
d_H\left(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\right) = \max_{||p||=1}\left| r\left(||C_n^{1/2}p|| - ||C_{n,k}^{1/2}p||\right) + \langle \mu_n, p\rangle - \langle \mu_{n,k}, p\rangle\right|.
$$

Dabei bezeichnen $\mu_{n,k}$ und $C_{n,k}$ die Schätzungen für μ und Σ, die sich aus der kontaminierten Stichprobe ergeben.

Im Folgenden wird der Zusammenhang zwischen den Bruchpunktdefinitionen für Lokations- und Kovarianzschätzer und der in dieser Arbeit vorgeschlagenen Bruchpunktdefinition 2.6 für Schätzer konvexer Körper verdeutlicht. Dazu wird gezeigt, dass gilt:

$$
d_H\left(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\right) \to \infty \text{ genau dann, wenn } ||\mu_{n,k}|| \to \infty \text{ oder } \lambda_d(C(\tilde{Y}_{n,k})) \to \infty.
$$

Zunächst gilt:

$$
\begin{aligned}
d_H\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big) &\le \max_{||p||=1}\left(\left| r\left(||C_n^{1/2}p|| - ||C_{n,k}^{1/2}p||\right)\right| + |\langle \mu_n, p\rangle - \langle \mu_{n,k}, p\rangle|\right) \\
&\le r\max_{||p||=1}\Big| ||C_n^{1/2}p|| - ||C_{n,k}^{1/2}p||\Big| + \max_{||p||=1}|\langle \mu_n - \mu_{n,k}, p\rangle| \\
&= r\max_{||p||=1}\Big| ||C_n^{1/2}p|| - ||C_{n,k}^{1/2}p||\Big| + ||\mu_n - \mu_{n,k}|| \\
&\le r\max_{||p||=1}\left(\max\{||C_n^{1/2}p||, ||C_{n,k}^{1/2}p||\}\right) + ||\mu_n - \mu_{n,k}|| \\
&\le r\max\{\lambda_d^{1/2}(C_n), \lambda_d^{1/2}(C_{n,k})\} + ||\mu_n - \mu_{n,k}||.
\end{aligned}
$$

Die letzte Ungleichheit folgt aus der so genannten Verträglichkeit zwischen der Spektralnorm und der euklidischen Vektornorm. Verträglichkeit bedeutet, dass $||Ap|| \le \lambda_d^{1/2}(A^TA)||p||$, mit $A \in \mathbb{R}^{d\times d}$ und $p \in \mathbb{R}^d$ (siehe Zurmühl, Falk (1987), S. 36).

Weiterhin gilt für $p := \frac{\mu_{n,k}}{||\mu_{n,k}||}$:

$$d_H\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big) \geq \Big|r\Big(||C_n^{1/2}\frac{\mu_{n,k}}{||\mu_{n,k}||}|| - ||C_{n,k}^{1/2}\frac{\mu_{n,k}}{||\mu_{n,k}||}||\Big) + \langle \mu_n, \frac{\mu_{n,k}}{||\mu_{n,k}||}\rangle - ||\mu_{n,k}||\Big|.$$

Für $\tilde{p}$, den normierten Eigenvektor zum größten Eigenwert der Matrix $C_{n,k}^{1/2}$, gilt die folgende Abschätzung:

$$\begin{aligned} d_H\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big) &\geq \Big|r\Big(||C_n^{1/2}\tilde{p}|| - ||C_{n,k}^{1/2}\tilde{p}||\Big) + \langle \mu_n, \tilde{p}\rangle - \langle \mu_{n,k}, \tilde{p}\rangle\Big| \\ &= \Big|r\Big(||C_n^{1/2}\tilde{p}|| - \lambda_d(C_{n,k}^{1/2})\Big) + \langle \mu_n, \tilde{p}\rangle - \langle \mu_{n,k}, \tilde{p}\rangle\Big|. \end{aligned}$$

Es folgt also, dass der Schätzer $T(\tilde{X}_n)$ genau dann zusammenbricht, wenn der Lokationsschätzer der kontaminierten Stichprobe oder der größte Eigenwert der geschätzten Kovarianzmatrix der kontaminierten Stichprobe gegen unendlich streben. Somit hängt das Bruchpunktverhalten von $T(\tilde{X}_n)$ nur von den verwendeten Lokations- und Kovarianzschätzern ab.

Bemerkung 2.6 Der Zusammenbruch in Definition 2.6 ergibt sich dadurch, dass einer der Randpunkte der Schätzung beliebig groß wird. In diesem Fall werden die Schätzungen an den Rand des Parameterraumes geschoben und die Schätzung degeneriert zu einer konvexen nicht beschränkten Teilmenge des $\mathbf{R}^d$. Über einen Zusammenbruch sollte jedoch ebenfalls gesprochen werden, falls durch eine verunreinigte Stichprobe der Schätzer zu einem $(d-1)$-dimensionalen Gebilde degeneriert. Auch dann wird die Schätzung an den Rand des Parameterraumes gedrängt. Im Folgenden wird eine Erweiterung von Definition 2.6 vorgestellt, die auch diesen praxisrelevanten Sachverhalt mit berücksichtigt.

2.2.2 Finite-sample Implosions-Bruchpunktdefinition

Definition 2.6 soll nun soweit für Schätzungen konvexer Körper erweitert werden, so dass der Zusammenbruch auch dann definiert ist, wenn die Schätzung auf der Grundlage einer kontaminierten Stichprobe zu einem $(d-1)$-dimensionalen Gebilde degeneriert. Zur Erweiterung der Definition wird die so genannte Polarmenge von $K \in \mathcal{K}^d$ benötigt.

Durch diese Menge wird der konvexe Körper K in eineindeutiger Weise festgelegt. Zudem wird eine weitere Hilfsfunktion, die so genannte Distanzfunktion, eingeführt.

Distanzfunktion und Polarmenge eines konvexen Körpers

Im Abschnitt 2.2 wurde die Stützfunktion als Werkzeug zur Untersuchung von konvexen Körpern eingeführt. Eine weitere Funktion, die einen konvexen Körper eindeutig charakterisiert, ist die Distanzfunktion.

Definition 2.8 *Sei $K \in \mathcal{K}^d$ ein konvexer Körper. Der Koordinatenursprung $\mathbf{0}$ liege im Inneren von K. Es sei $p \neq \mathbf{0}$, und $x \in \mathbb{R}^d$ der einzige Schnittpunkt der von $\mathbf{0}$ ausgehenden Halbgeraden $\mathbf{0}p$ mit dem Rand von K. Unter der Distanzfunktion $g(K,p)$ von K wird dann der Quotient der Abstände vom Ursprung $\mathbf{0}$ zum Punkt p und vom Ursprung $\mathbf{0}$ zum Punkt x verstanden:*

$$g(K,p) = \frac{||p||}{||x||}.$$

Bemerkung 2.7 1. Die Distanzfunktion nimmt den Wert 1 genau dann an, wenn p ein Randpunkt des konvexen Körpers ist, und $g(K,p) < 1$ für alle Punkte p im Inneren von K. Ist $g(K,p) > 1$ so liegt p außerhalb von K.
2. Die Distanzfunktion $g(K,\cdot) : \mathbf{R}^d \to \mathbf{R}$ von K hat die folgenden Eigenschaften:

- $g(K,p) > 0$, für $p \neq \mathbf{0}$ und $g(K,\mathbf{0}) = 0$,
- $g(K,\lambda p) = \lambda g(K,p)$ für $\lambda \geq 0$ (positive Homogenität vom Grad Eins),
- und $g(K,p_1 + p_2) \leq g(K,p_1) + g(K,p_2)$ (Subadditivität).

Durch die Distanzfunktion ist ein konvexer Körper ebenso wie durch die Stützfunktion eindeutig beschrieben.

Satz 2.1 Ist g eine nichtnegative, vom Grad Eins positiv homogene und subadditive Funktion, so existiert genau ein konvexer Körper K, der den Ursprung $\mathbf{0}$ im Innern enthält und dessen Distanzfunktion g ist. Dieser konvexe Körper kann dargestellt werden durch:

$$K = \{p \in \mathbb{R}^d : g(p) \leq 1\}.$$

Beweis: (Leichtweiß (1970))

q.e.d.

Ist K ein konvexer Körper und der Nullpunkt in seinem Inneren gelegen, so ist die Stützfunktion $h(K, p)$ von K positiv für $p \neq \mathbf{0}$. Die Stützfunktion hat somit ebenfalls die Eigenschaften aus Bemerkung 2.7 Punkt 2, die für eine Distanzfunktion charakteristisch sind. Deshalb kann auch jede Stützfunktion eines konvexen Körpers als Distanzfunktion eines anderen konvexen Körpers K^* aufgefasst werden. Dies führt zur folgenden Definition.

Definition 2.9 *Es sei $K \in \mathcal{K}^d$ und der Ursprung sei im Inneren von K enthalten. Dann heißt*

$$K^* = \{p \in \mathbb{R}^d : h(K, p) \leq 1\}$$

Polarmenge von K bzgl. $\mathbf{0}$.

Lemma 2.3 Im Fall eines kompakten konvexen Körpers K des $\mathbb{R}^d$, der in seinem Inneren den Ursprung $\mathbf{0}$ enthält, ist die Stützfunktion h von K mit der Distanzfunktion g^* von K^* und ebenso die Stützfunktion h^* von K^* mit der Distanzfunktion g von K identisch, d.h.: $h(K^*, p) = g(K, p)$ und entsprechend $h(K, p) = g(K^*, p)$.

Beweis: (Leichtweiß (1979))

q.e.d.

Falls K den Ursprung in seinem Inneren enthält, so sind in diesem Fall die Stützfunktion und die Distanzfunktion zueinander „dual“. Es gilt dann $(K^*)^* = K$. Der konvexe Körper K ist somit durch K^* eindeutig festgelegt und ebenso K^* durch K.

Falls der konvexe Körper einer Ellipse entspricht, so gilt der folgende Zusammenhang:

Lemma 2.4 Falls $K = \{x \in \mathbf{R}^d : x^T \Sigma^{-1} x \leq r^2\} \in \mathcal{K}^d$ eine Ellipse mit $r > 0$ und Σ p.d. ist, dann gilt:

1. $\max\limits_{||p||=1} |h(K,p)| \longrightarrow \infty,$ genau dann wenn $\lambda_d(\Sigma) \to \infty$ und

2. $\max\limits_{||p||=1} |h(K^*,p)| \longrightarrow \infty,$ genau dann wenn $\lambda_1(\Sigma) \to 0$.

Beweis:

Die erste Behauptung folgt direkt aus Beispiel 2.4.

Aus Definition 2.9 ergibt sich, dass die Polarmenge K^* von K gegeben ist durch:

$$
\begin{aligned}
K^* &= \Big\{x \in \mathbf{R}^d : h(K,x) \leq 1\Big\} = \Big\{x \in \mathbf{R}^d : r||\Sigma^{1/2}x|| \leq 1\Big\} \\
&= \Big\{x \in \mathbf{R}^d : r^2(\Sigma^{1/2}x)^T(\Sigma^{1/2}x) \leq 1\Big\} \\
&= \left\{x \in \mathbf{R}^d : x^T \Sigma x \leq \frac{1}{r^2}\right\},
\end{aligned}
$$

mit zugehöriger Stützfunktion: $h(K^*,p) = \frac{1}{r}||\Sigma^{-1/2}p||$.

Sei $\tilde{p}$ der auf Länge Eins normierte Eigenvektor zum größten Eigenwert der Matrix $\Sigma^{-1/2}$, dann folgt analog zu Beispiel 2.4, dass $\max_{||p||=1} |h(K^*,p)| = \frac{1}{r}\lambda_d(\Sigma^{-1/2})$, da

$$
\frac{1}{r}\lambda_d(\Sigma^{-1/2}) \leq \frac{1}{r}||\Sigma^{-1/2}\tilde{p}|| \leq \max_{||p||=1} |h(K^*,p)| \leq \frac{1}{r}\lambda_d(\Sigma^{-1/2}).
$$

Nun strebt $\max_{||p||=1} |h(K^*,p)|$ genau dann gegen unendlich, wenn $\lambda_1(\Sigma) \to 0$, da gilt $\lambda_d(\Sigma^{-1/2}) = \lambda_1^{-1/2}(\Sigma)$.

q.e.d.

Beispiel 2.4: (Fortsetzung)

Von einem Zusammenbruch des Schätzers $T(\tilde{X}_n)$ zu einem $(d-1)$-dimensionalen Körper kann wegen Lemma 2.4 gesprochen werden, wenn die Stützfunktion der Schätzung der Polarmenge von $T(\tilde{X}_n)$ auf der Grundlage der kontaminierten Stichprobe beliebig groß wird. Um sicherzugehen, dass die Schätzung $T(\tilde{X}_n)$ den Ursprung in ihrem Inneren enthält, wird die Polarmenge der um Null zentrierten Schätzung betrachtet, d.h. die Polarmenge von $T(\tilde{X}_n - \mu_n) = \{x \in \mathbf{R}^d : x^T C_n^{-1} x \leq r^2\}$ wird untersucht. Die Schätzung bricht dann zusammen, wenn der Hausdorff Abstand zwischen der um Null zentrierten Polarmenge der ursprünglichen Stichprobe und der Polarmenge der kontaminierten Stichprobe beliebig groß wird. Aus Lemma 2.4 folgt, dass dies genau dann geschieht, wenn $\lambda_d^{1/2}(C_{n,k}^{-1}) = \frac{1}{\lambda_1^{1/2}(C_{n,k})}$ gegen unendlich strebt. Wird dieser Fall bei der Bestimmung des Bruchpunktes zusätzlich betrachtet, so wird man von einem Zusammenbruch der Schätzung $T(\tilde{X}_n) = \left\{x \in \mathbb{R}^d : (x-\mu_n)^T C_n^{-1}(x-\mu_n) \leq r^2\right\}$ sprechen, wenn $\mu_{n,k}$ oder $\lambda_d^{1/2}(C_{n,k})$ gegen unendlich streben oder wenn $\lambda_1^{1/2}(C_{n,k})$ gegen Null strebt.

Aus den vorherigen Überlegungen ergibt sich die folgende Bruchpunktdefinition, die sowohl die Explosion als auch die Implosion der Schätzung eines konvexen Körpers berücksichtigt.

Definition 2.10 *Sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine Stichprobe vom Umfang n mit $x_i \in \mathbb{R}^d$ von Beobachtungen aus einer Verteilung $F_K \in \mathcal{F} = \{F_K : K \in \Theta\}$, wobei $\Theta \subseteq \mathcal{K}^d$. Sei $\tilde{Y}_{n,k}$ wie in Definition 2.3. Sei $T(\tilde{X}_n)$ ein Schätzer für den konvexen Körper K und entsprechend $T^*(\tilde{X}_n)$ ein Schätzer für die Polarmenge des um Null zentrierten konvexen Körpers K. Dann heißt:*

$$\epsilon_{K,K^*}(\tilde{X}_n, T) = \min\left\{\epsilon_K(\tilde{X}_n, T), \epsilon_K(\tilde{X}_n, T^*)\right\}$$

der finite-sample Bruchpunkt für $T(\tilde{X}_n)$.
Dabei bezeichne

$$\epsilon_K(\tilde{X}_n, T) = \min_{1\leq k\leq n}\left\{\frac{k}{n} : \sup_{\tilde{Y}_{n,k}} \max_{||p||=1} |h(T(\tilde{X}_n), p) - h(T(\tilde{Y}_{n,k}), p)| = \infty\right\},$$

den finite-sample Explosions-Bruchpunkt und

$$\epsilon_K(\tilde{X}_n, T^*) = \min_{1\leq k\leq n}\left\{\frac{k}{n} : \sup_{\tilde{Y}_{n,k}} \max_{||p||=1} |h(T^*(\tilde{X}_n), p) - h(T^*(\tilde{Y}_{n,k}), p)| = \infty\right\},$$

den so genannten finite-sample Implosions-Bruchpunkt.

Bemerkung 2.8 In Definition 2.6 und 2.10 wurde der Zusammenbruch jeweils durch das Ersetzen von Beobachtungen herbeigeführt. In beiden Situationen kann der Zusammenbruch auch durch das Hinzufügen von Beobachtungen erreicht werden. Inwieweit es Zusammenhänge zwischen diesen beiden Situationen gibt, muss noch untersucht werden.

Zusammenhänge

Wird die Parametrisierung aus Beispiel 2.2 zu Grunde gelegt, so erlaubt die vorgestellte Definition 2.6 die Untersuchung des Bruchpunktverhaltens eines Lokationsschätzers, falls davon ausgegangen wird, dass Σ bekannt ist.
Zudem erlaubt Definition 2.10 die simultane Bruchpunkt-Untersuchung von Lokations- und Kovarianzschätzern, da es eine eineindeutige Zuordnung zwischen $K = \{x \in \mathbf{R}^d : (x-\mu)^T\Sigma^{-1}(x-\mu) \leq r^2\}$, $r > 0$ und dem Tupel $\{\mu, \Sigma\}$ gibt. Um diese zu sehen, wird die Schätzung $T_K(\tilde{X}_n) := T(\tilde{X}_n) \oplus T^*(\tilde{X}_n)$ betrachtet, wobei $T(\tilde{X}_n)$ ein Schätzer für den konvexen Körper K und entsprechend $T^*(\tilde{X}_n)$ ein Schätzer für die Polarmenge des um Null zentrierten konvexen Körpers K ist. Wird dieser Schätzer bzgl. seines Bruchpunktes untersucht, reicht es aus, den Wert des Bruchpunktes nach Definition 2.6 zu bestimmen, da gilt:

$$h(T_K(\tilde{X}_n), p) = h(T(\tilde{X}_n), p) + h(T^*(\tilde{X}_n), p).$$

Von einem Zusammenbruch der Schätzung wird gesprochen, falls der Hausdorff Abstand zwischen $T_K(\tilde{X}_n)$ und $T_K(\tilde{Y}_{n,k})$ beliebig groß wird. Es wird somit der folgende Abstand betrachtet:

$$\begin{aligned} d_H\Big(T_K(\tilde{X}_n), T_K(\tilde{Y}_{n,k})\Big) &= \max_{||p||=1} \Big| r\Big(||C_n^{1/2}p|| - ||C_{n,k}^{1/2}p||\Big) + \frac{1}{r}\Big(||C_n^{-1/2}p|| - ||C_{n,k}^{-1/2}p||\Big) \\ &+ \langle \mu_n, p\rangle - \langle \mu_{n,k}, p\rangle \Big|. \end{aligned}$$

Diese vorgestellte Bruchpunkt Definition für $\Big\{T(\tilde{X}_n), C(\tilde{X}_n)\Big\}$ führt zum selben Wert des Bruchpunkts eines Schätzers wie die Vorschläge von Davies (1987) und Tyler (1994). Der Zusammenbruch der Schätzung von K und somit auch von $\{\mu, \Sigma\}$ wird mittels des Hausdorff Abstandes erklärt, also einer geeigneten Metrik, im Gegensatz zu den von Davies (1987) und Tyler (1994) vorgeschlagenen Bruchpunkt Definitionen (siehe Definition 2.5).

Für den Fall $d = 1$ und $r = 1$ ist der Abstand vom Ursprung $\mathbf{0}$ zu der Schätzung $T_K(\tilde{Y}_{n,k})$ gegeben durch:

$$d_H\Big(\mathbf{0}, T_K(\tilde{Y}_{n,k})\Big) = \left| |T_K(\tilde{Y}_{n,k})| + \sigma_{n,k} + \frac{1}{\sigma_{n,k}} \right|$$

und entspricht somit der von Davies (1987) definierten Größe B_D aus Definition 2.5, wobei $\sigma_{n,k}$ die mittels der kontaminierten Stichprobe geschätzte Standardabweichung ist.

Ebenso wird durch Definition 2.10 der Zusammenbruch bei bekanntem μ und unbekannter Kovarianzmatrix Σ erfasst. Im Gegensatz zur Bruchpunktdefiniton 2.4 wird der Zusammenbruch über eine Metrik und nicht über einer Semimetrik erklärt.

Basiert die Schätzung des interessierenden konvexen Körpers auf der konvexen Hülle einer endlichen Punktmenge, also auf einem Polytop, so ist es ebenfalls möglich, mit Definition 2.10 den Wert des Bruchpunktes zu bestimmen (siehe zum Beispiel Lemma 3.10).

2.3 Eigenschaften der finite-sample Bruchpunkte für Schätzer konvexer Körper

In diesem Abschnitt werden die Eigenschaften der in dieser Arbeit vorgeschlagenen finite-sample Bruchpunkte für Schätzer konvexer Körper untersucht. Die Beweise werden jeweils für den finite-sample Explosions-Bruchpunkt $\epsilon_K(\tilde{X}_n, T)$ (siehe Definition 2.6) durchgeführt. Alle gezeigten Aussagen gelten aufgrund der Eigenschaften der Stützfunktion auch für den Implosions-Bruchpunkt $\epsilon_K(\tilde{X}_n, T^*)$ aus Definition 2.10.

Im folgenden Lemma wird gezeigt, dass der finite-sample Bruchpunkt eines affin äquivarianten Schätzers invariant ist unter affinen Transformationen.

Lemma 2.5 Gegeben sei eine Stichprobe $\tilde{X}_n = \{x_1, ..., x_n\}$, mit $x_i \in \mathbb{R}^d$, $i = 1, ..., n$. Der Schätzer $T(\tilde{X}_n) \in \mathcal{K}^d$ sei affin äquivariant, dann gilt:

$$\epsilon_K(\tilde{X}_n, T) = \epsilon_K(A\tilde{X}_n + b, T), \tag{2.8}$$

für jede nichtsinguläre Matrix $A \in \mathbb{R}^{d\times d}$ und jedes $b \in \mathbb{R}^d$.

Beweis:
Sei $A \in \mathbb{R}^{d\times d}$ eine nichtsinguläre Matrix und $b \in \mathbb{R}^d$. Sei $\tilde{X}_n$ eine Stichprobe regulärer Beobachtungen vom Umfang n. Mit $\tilde{Y}_{n,k}$ werde eine Stichprobe bezeichnet, die durch Austausch von k Beobachtungen von $\tilde{X}_n$ entsteht. Falls der Schätzer $T(\tilde{X}_n)$ affin äquivariant ist, ergibt sich:

$$\begin{aligned} d_H\Big(T(A\tilde{X}_n + b), T(A\tilde{Y}_{n,k} + b)\Big) &= \max_{||p||=1} |h(T(A\tilde{X}_n + b), p) - h(T(A\tilde{Y}_{n,k} + b), p)| \\ &= \max_{||p||=1} |h(T(A\tilde{X}_n) \oplus b, p) - h(T(A\tilde{Y}_{n,k}) \oplus b, p)| \\ &= \max_{||p||=1} |h(T(A\tilde{X}_n), p) - h(T(A\tilde{Y}_{n,k}), p)| \\ &= d_H\Big(T(A\tilde{X}_n), T(A\tilde{Y}_{n,k})\Big). \end{aligned}$$

Weiterhin ist über die Stützfunktion von $T(A\tilde{X}_n)$ bekannt, dass:

$$h(T(A\tilde{X}_n), p) \;=\; \max\Big\{\langle x, p\rangle, \quad x \in T(A\tilde{X}_n)\Big\}$$

$$
\begin{aligned}
&= \max\Big\{\langle x, p\rangle, \quad x \in AT(\tilde{X}_n)\Big\} \\
&= \max\Big\{\langle Ay, p\rangle, \quad y \in T(\tilde{X}_n)\Big\}, \text{ wobei } y := A^{-1}x \\
&= \max\Big\{\langle y, A^T p\rangle, \quad y \in T(\tilde{X}_n)\Big\} \\
&= h(T(\tilde{X}_n), A^T p).
\end{aligned}
$$

Da die Stützfunktion homogen vom Grad Eins ist, folgt:

$$
d_H\Big(T(A\tilde{X}_n + b), T(A\tilde{Y}_{n,k} + b)\Big) = \max_{||p||=1} ||A^T p|| \left| h\left(T(\tilde{X}_n), \frac{A^T p}{||A^T p||}\right) - h\left(T(\tilde{Y}_{n,k}), \frac{A^T p}{||A^T p||}\right) \right|.
$$

Für eine Abschätzung werden die folgenden Beziehungen gebraucht:

1. Seien $f : K \to \mathbf{R}$ und $g : K \to \mathbf{R}$ Funktionen mit $K \subseteq \mathbf{R}^d$, so dass $|f(p)| \leq M_1$ und $|g(p)| \leq M_2$ für alle $p \in K$, dann gilt:
   $\min_p |g(p)| \max_p |f(p)| \leq \max_p |g(p)||f(p)| \leq \max_p |g(p)| \max_p |f(p)|$.

2. Für symmetrische $d \times d$-Matrizen $B = A^T A$ und $||p|| = 1$ gilt:
$$
\lambda_1(B) = \min_p \frac{p^T B p}{p^T p} \text{ und } \lambda_d(B) = \max_p \frac{p^T B p}{p^T p}.
$$

Sei $g(p) = ||A^T p||^2 = p^T \left(A^T A\right) p$ und $f(p) = \left| h\left(T(\tilde{X}_n), \frac{A^T p}{||A^T p||}\right) - h\left(T(\tilde{Y}_{n,k}), \frac{A^T p}{||A^T p||}\right) \right|^2$, dann lässt sich der quadrierte Hausdorff Abstand wie folgt abschätzen:

$$
\begin{aligned}
1.\ d_H\Big(T(A\tilde{X}_n + b), T(A\tilde{Y}_{n,k} + b)\Big)^2 &\leq \lambda_d(A^T A) d_H\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big)^2, \\
2.\ d_H\Big(T(A\tilde{X}_n + b), T(A\tilde{Y}_{n,k} + b)\Big)^2 &\geq \lambda_1(A^T A) d_H\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big)^2,
\end{aligned}
$$

so dass insgesamt folgt:

$$
\lambda_1(A^T A) \leq \frac{d_H\Big(T(A\tilde{X}_n + b), T(A\tilde{Y}_{n,k} + b)\Big)^2}{d_H\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big)^2} \leq \lambda_d(A^T A).
$$

Somit sind $\sup_{A\tilde{Y}_{n,k}+b} d\Big(T(A\tilde{X}_n + b), T(A\tilde{Y}_{n,k} + b)\Big)$ über alle möglichen Stichproben $A\tilde{Y}_{n,k} + b$ und $\sup_{\tilde{Y}_{n,k}} d\Big(T(\tilde{X}_n), T(\tilde{Y}_{n,k})\Big)$ über alle möglichen Stichproben $\tilde{Y}_{n,k}$, die sich in höchstens k Punkten von der Stichprobe $A\tilde{X}_n + b$ unterscheiden, gleichzeitig endlich oder unendlich. Das bedeutet, dass der Bruchpunkt jeweils übereinstimmen muss.

q.e.d.

Im Abschnitt 2.1 wurden die Eigenschaften der Translationsäquivarianz und der affinen Äquivarianz eines Schätzer eingeführt. Zur Bewertung des Bruchpunktverhaltens eines translationsäquivarianten bzw. affin äquivarianten Schätzers ist es notwendig, die kleinste obere Schranke des Bruchpunktes zu kennen. Im Folgenden wird die kleinste obere Schranke des Explosions-Bruchpunktes gemäß Definition 2.6 für translationsäquivariante und affin äquivariante Schätzer konvexer Körper bestimmt.

Lemma 2.6 Gegeben sei eine Stichprobe $\tilde{X}_n$ vom Umfang n. Falls $T(\tilde{X}_n) \in \mathcal{K}^d$ ein translationsäquivarianter Schätzer eines konvexen Körpers ist, gilt:

$$\epsilon_K(\tilde{X}_n, T) \leq \lfloor (n+1)/2 \rfloor / n,$$

wobei $\lfloor x \rfloor$ die größte ganze Zahl kleiner oder gleich x bezeichnet.

Beweis:

Der Beweis wird über einen Widerspruch geführt. Dazu wird angenommen, dass $\epsilon_K(\tilde{X}_n, T) > \lfloor (n+1)/2 \rfloor / n$ ist. Dann muss ein $c \in \mathbf{R}$ existieren, so dass

$$\max_{||p||=1} |h(T(\tilde{X}_n), p) - h(T(\tilde{Y}_{n,k}), p)| \leq c < \infty, \tag{2.9}$$

für alle kontaminierten Stichproben $\tilde{Y}_{n,k}$, die durch Austausch von $k = \lfloor (n+1)/2 \rfloor$ Beobachtungen der Stichprobe $\tilde{X}_n$ entstehen. Sei nun $q = n - k$ die Anzahl der Beobachtungen von $\tilde{X}_n$, welche nicht ersetzt wurden. Da $2q \leq n$, kann für jedes $b \in \mathbf{R}^d$ eine kontaminierte Stichprobe $\tilde{Y}_{n,k,b}$ konstruiert werden, welche $x_1, ..., x_q, x_1 + b, ..., x_q + b$ enthält und eine entsprechende Stichprobe $\tilde{Z}_{n,k,b} = \tilde{Y}_{n,k,b} - b$. Die kontaminierten Stichproben enthalten mindestens q Punkte von $\tilde{X}_n$ und nach Annahme (2.9) gilt zum einem, dass

$$\max_{||p||=1} |h(T(\tilde{X}_n), p) - h(T(\tilde{Y}_{n,k,b}), p)| \leq c$$

und zum anderen:

$$\begin{aligned} c &\geq \max_{||p||=1} |h(T(\tilde{X}_n), p) - h(T(\tilde{Z}_{n,k,b}), p)| \\ &= \max_{||p||=1} |h(T(\tilde{X}_n), p) - h(T(\tilde{Y}_{n,k,b}), p) + \langle b, p \rangle| \\ &\geq \left| h\left(T(\tilde{X}_n), \frac{b}{||b||}\right) - h\left(T(\tilde{Y}_{n,k,b}), \frac{b}{||b||}\right) + ||b|| \right|, \end{aligned}$$

aufgrund der Translationsäquivarianz des betrachteten Schätzers.
Falls $||b||$ gegen unendlich strebt, können nicht beide Ungleichungen erfüllt sein, woraus die Behauptung folgt.

q.e.d.

Im Folgenden Beispiel wird ein Schätzer vorgestellt, der diese obere Schranke annimmt.

Beispiel 2.5 Sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine Stichprobe vom Umfang n mit $x_i \in \mathbb{R}^d$, $i = 1, ..., n$. Gegeben sei der folgende Schätzer $T(\tilde{X}_n) = \{x \in \mathbb{R}^d : (x-\mu_{n,1})^T(x-\mu_{n,1}) \leq 1\}$, wobei $\mu_{n,1}$ den L_1-Median bezeichne. Der L_1-Median ist definiert als derjenige Vektor $\mu_{n,1} \in \mathbb{R}^d$, für den gilt:

$$\sum_{i=1}^{n} ||x_i - \mu_{n,1}|| \leq \sum_{i=1}^{n} ||x_i - t|| \text{ für alle } t \in \mathbb{R}^d.$$

Der L_1-Median ist translationsäquivariant, aber nicht affin äquivariant. Lopuhäa und Rousseeuw (1991) zeigen, dass der finite-sample Bruchpunkt nach Definition 2.3 dieses Lokationsschätzers gleich $\lfloor (n+1)/2 \rfloor / n$ ist. Somit folgt aus Beispiel 2.4, dass der Bruchpunkt des Schätzers $T(\tilde{X}_n)$ der in Lemma 2.6 ermittelten oberen Schranke des Bruchpunktes für translationsäquivariante Schätzer konvexer Körper entspricht.

Der Wert der kleinsten oberen Schranke des Bruchpunktes eines translationsäquivarianten Schätzers hängt also nicht von der Dimension d ab.
Um die kleinste obere Schranke des Bruchpunktes eines affin äquivarianten Schätzers zu ermitteln, wird eine zusätzliche Annahme getroffen. Die Stichprobe $\tilde{X}_n$ muss in so genannter allgemeiner Lage sein.

Definition 2.11 *Sei $\tilde{X}_n$ eine Stichprobe vom Umfang n, mit $x_i \in \mathbb{R}^d$, $i = 1, ..., n$. Die Stichprobe $\tilde{X}_n$ heißt in allgemeiner Lage, wenn in jedem $(d-1)$-dimensionalen affinen Unterraum des $\mathbb{R}^d$ höchstens d Punkte der Stichprobe $\tilde{X}_n$ liegen.*

Unter dieser Annahme gilt:

Lemma 2.7 Gegeben sei eine Stichprobe $\tilde{X}_n$ in allgemeiner Lage, mit $n \geq d+1$. Falls $T(\tilde{X}_n)$ ein affin äquivarianter Schätzer eines konvexen Körpers ist, dann gilt:

$$\epsilon_{K,K^*}(\tilde{X}_n, T) \leq \frac{\lfloor (n-d+1)/2 \rfloor}{n}.$$

Beweis:

Zunächst werden zwei ineinander überführbare Stichproben konstruiert (siehe Davies (1987)).

Sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine d-dimensionale Stichprobe vom Umfang n mit $x_i = (x_{i1}, ..., x_{id})^T \in \mathbb{R}^d$ und $\tilde{Y}_{n,k}$ eine durch $k = \lfloor (n-d+1)/2 \rfloor$ Beobachtungen kontaminierte Stichprobe. In $\tilde{Y}_{n,k}$ sind somit noch $n - k \geq d$ Beobachtungen der Stichprobe $\tilde{X}_n$ enthalten. Aus diesen werden d Beobachtungen gewählt und E bezeichne die durch diese Punkte aufgespannte Hyperebene. Da der betrachtete Schätzer affin äquivariant ist, kann $E = \{y = (y_1, ..., y_d)^T \in \mathbb{R}^d : y_d = 0\}$ gewählt werden. Nun bleiben in $\tilde{Y}_{n,k}$ noch $n' = n - k - d$ unverfälschte Beobachtungen, bezeichnet mit $\tilde{x}_1, ..., \tilde{x}_{n'}$. Es werden n' dieser kontaminierten Beobachtungen ersetzt durch: $x_{ij} = \tilde{x}_{ij}$ für $1 \leq i \leq n', 1 \leq j \leq (d-1)$, und $x_{id} = t\tilde{x}_{id}$, wobei $t > 0$ und $x_i = (x_{i1}, ..., x_{id})^T$. Diese so konstruierte Stichprobe wird mit $\tilde{Y}_{n,k,t}$ bezeichnet.

Sei $\Delta(t) = diag(1, ..., 1, t^{-1}) \in \mathbb{R}^{d \times d}$ und $\tilde{Y}_{n,k,1/t} = \Delta(t)\tilde{Y}_{n,k,t}$. Somit entsteht $\tilde{Y}_{n,k,1/t}$, ebenso wie $\tilde{Y}_{n,k,t}$, aus $\tilde{X}_n$ durch den Austausch von höchstens k Beobachtungen.

Nun gilt:

$$\begin{aligned}
\frac{1}{t}\left| h\left(T(\tilde{Y}_{n,k,t}), \frac{\Delta(t)^T \tilde{p}}{||\Delta(t)^T \tilde{p}||}\right)\right| &\leq \max_{||p||=1} |h(T(\tilde{Y}_{n,k,1/t}), p)| \\
&= \max_{||p||=1} \left| ||\Delta(t)^T p|| h\left(T(\tilde{Y}_{n,k,t}), \frac{\Delta(t)^T p}{||\Delta(t)^T p||}\right)\right| \\
&\leq \frac{1}{t} \max_{||p||=1} \left| h\left(T(\tilde{Y}_{n,k,t}), \frac{\Delta(t)^T p}{||\Delta(t)^T p||}\right)\right|,
\end{aligned}$$

für $\tilde{p} = (0, ..., 0, 1)^T \in \mathbb{R}^d$ und $t < 1$.

Falls t gegen Null strebt so ergibt sich, dass entweder $\max_{||p||=1} |h(T(\tilde{Y}_{n,k,t}), p)|$ oder $\max_{||p||=1} \left| h\left(T(\tilde{Y}_{n,k,t}), \frac{\Delta(t)^T p}{||\Delta(t)^T p||}\right)\right|$ gegen Null oder gegen ∞ konvergieren muss.

Strebt nun $\max_{||p||=1} |h(T(\tilde{Y}_{n,k,t}), p)| = \max_{||p||=1} |g(T^*(\tilde{Y}_{n,k,t}), p)|$ gegen Null, so ist dies äquivalent zu $\max_{||p||=1} |h(T^*(\tilde{Y}_{n,k,t}), p)| \to \infty$ und somit folgt die Behauptung.

q.e.d.

Bemerkung 2.9 Im fünften Kapitel werden affin äquivariante Schätzer konvexer Körper vorgestellt, die einen Bruchpunkt von $\lfloor \frac{n-d+1}{2} \rfloor / n$ aufweisen und somit die kleinste obere Schranke für affin äquivariante Schätzer konvexer Körper annehmen.

Die vorgestellten Bruchpunktdefinitionen erlauben die Untersuchung des Bruchpunktverhaltens von Schätzern konvexer Körper. In den nächsten beiden Kapiteln werden Beispiele konvexer Körper vorgestellt und deren Schätzungen auf ihr Bruchpunktverhalten hin untersucht.

Kapitel 3

Zonoide von Verteilungen

In diesem Kapitel wird eine spezielle Klasse von konvexen Körpern, die so genannten Zonoide von Verteilungen, vorgestellt. Diese weisen die Besonderheit auf, dass sie dem Erwartungswert eines zufälligen konvexen Körpers entsprechen und zwar dem Erwartungswert der Strecke vom Ursprung $\mathbf{0}$ zum Zufallsvektor X.
Um diesen Zusammenhang verdeutlichen zu können, werden in Abschnitt 3.1 zufällige konvexe Körper und deren Erwartungswerte in allgemeiner Form definiert. In den sich anschließenden Abschnitten werden die von Koshevoy und Mosler (1998) vorgestellten Zonoide (Abschnitt 3.2) und Liftzonoide (Abschnitt 3.3) von Verteilungen definiert und deren Eigenschaften untersucht. Es zeigt sich, dass diese konvexen Körper dem Erwartungswert einer zufälligen Strecke entsprechen. Desweiteren werden erste Schätzmethoden für Zonoide und Liftzonoide beschrieben und die zugehörigen Bruchpunkte ermittelt.

3.1 Der Erwartungswert eines zufälligen konvexen Körpers

In diesem Abschnitt liegt der Schwerpunkt auf dem Erwartungswert zufälliger Strecken. Um diese Körper adäquat beschreiben zu können, werden Begriffe aus der stochastischen Geometrie benötigt. In der stochastischen Geometrie werden statistische Modelle zur Beschreibung zufälliger geometrischer Strukturen entwickelt und untersucht. Die statistische Behandlung zufälliger geometrischer Strukturen beginnt im 18. Jahrhundert mit dem so genannten Buffonschen Nadelproblem (siehe Benz (1980)). Dieses Problem fragt nach der Wahrscheinlichkeit, dass eine zufällig auf ein paralleles ebenes Geradenmuster mit Abstand D geworfene Nadel mit Länge $L < D$ eine der Geraden schneidet. Anwendung findet die stochastische Geometrie heute beispielsweise in der Bilderkennung und der Stereologie. Das Anliegen der Stereologie besteht darin, Aussagen über geometrische Eigenschaften dreidimensionaler Strukturen zu machen, für die nur Informationen aus ebenen oder linienförmigen Schnitten verfügbar sind, wie beispielsweise in der Computertomografie.

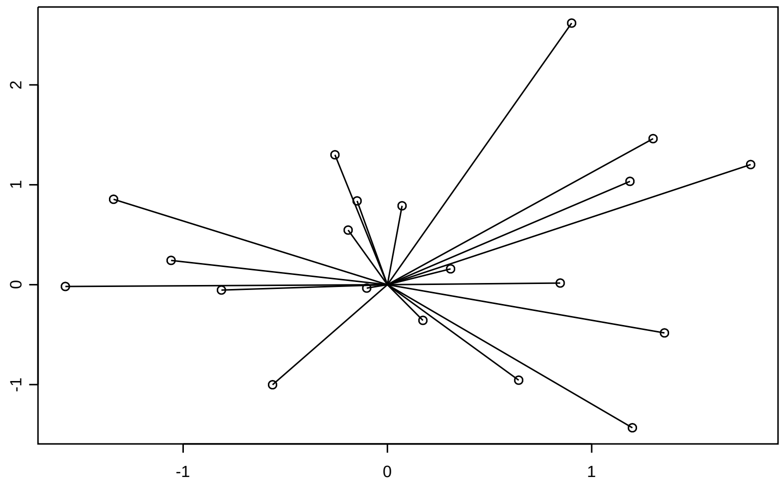
Zunächst werden in diesem Abschnitt benötigte Grundlagen vorgestellt, um später den Erwartungswert einer zufälligen Strecke einführen zu können. Eine ausführliche und weitergehende Darstellung ist beispielsweise bei Schneider und Weil (2000) oder Stoyan und Mecke (1983) zu finden.

Zufällige Mengen werden entsprechend dem üblichen Vorgehen in der Statistik als mengenwertige Zufallsvariablen eingeführt, also als eine messbare Abbildung von einem Wahrscheinlichkeitsraum in ein Teilsystem des $\mathbf{R}^d$, versehen mit einer geeigneten σ-Algebra. Als Teilsystem kann $\mathcal{S}^d$, die Menge aller abgeschlossenen Teilmengen im $\mathbf{R}^d$, betrachtet werden mit den Borelschen Teilmengen $\mathcal{B}(\mathcal{S}^d)$ als geeigneter σ-Algebra (vgl. Schneider und Weil (2000)). Diese sind folgendermaßen definiert:

Definition 3.1 *Sei $(\Omega, \mathcal{A}, P)$ ein Wahrscheinlichkeitsraum. Eine Abbildung $Z : \Omega \to \mathcal{S}^d$ heißt zufällige abgeschlossene Menge (ZAM), wenn $Z^{-1}(B) \in \mathcal{A}$ für alle $B \in \mathcal{B}(\mathcal{S}^d)$.*

Damit ist Z eine $(\mathcal{A}, \mathcal{B}(\mathcal{S}^d))$-messbare Abbildung eines Wahrscheinlichkeitsraumes $(\Omega, \mathcal{A}, P)$ in $(\mathcal{S}^d, \mathcal{B}(\mathcal{S}^d))$. Das Bildmaß $P_Z := Z(P)$ von P unter Z heißt dann Verteilung von Z. Für $B \in \mathcal{B}(\mathcal{S}^d)$ wird $P_Z(B) = P(Z \in B)$ als Abkürzung für $P(\{\omega \in \Omega : Z(\omega) \in B\})$ benutzt. Ist $P(Z \in B) = 1$, so gilt die Aussage $Z \in B$ fast sicher.

Abbildung 3.1: 20 Realisationen zufälliger Strecken $[\mathbf{0}, X]$, wobei $X \sim N(\mathbf{0}, I_2)$.



Beispiele für ZAM sind zufällige Punkte oder Kugeln mit zufälligen Mittelpunkten und/oder Radien. In Abbildung 3.1 werden 20 Realisationen zufälliger Strecken der Form $[\mathbf{0}, X]$, wobei $X \sim N(\mathbf{0}, I_2)$ und I_2 die Einheitsmatrix ist, betrachtet. Zu erkennen ist, dass sowohl die Richtung als auch die Länge der Strecken, die alle im Ursprung starten, zufällig sind.

Werden im Folgenden mehrere ZAM gleichzeitig betrachtet, so wird vorausgesetzt, dass der zugrunde liegende Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, P)$ derselbe ist. Schneider und Weil (2000) zeigen, dass für eine ZAM Z auch αZ, $\alpha \in \mathbf{R}$ und die konvexe Hülle von Z, diese werde mit $conv(Z)$ bezeichnet, ZAM sind. Weiterhin gilt: falls $Z_1, ..., Z_n$ ZAM sind, dann ist $\oplus_{i=1}^n Z_i$ eine ZAM. Gilt für eine ZAM Z fast sicher, dass $Z \in \mathcal{K}^d$ ist, so wird Z als zufälliger konvexer Körper (ZKK) bezeichnet. Im Weiteren seien alle betrachteten zufälligen Mengen ZKK. Diese werden mit K bezeichnet.

Im Folgenden wird der Erwartungswert eines ZKK K definiert. Diese Definition stammt von Artstein und Vitale (1975). Zur Definition des Erwartungswerts eines ZKK K wird der Begriff der Auswahl benötigt.

Definition 3.2 *Gegeben sei ein Wahrscheinlichkeitsraum* $(\Omega, \mathcal{A}, P)$*. Es stammen der ZKK* $K \in \mathcal{K}^d$ *und der Zufallsvektor* $X \in \mathbb{R}^d$ *aus diesem Wahrscheinlichkeitsraum. Dann heißt* X *Auswahl von* K*, falls gilt:*

$$P_K(X) = P(X \in K) = P(\{\omega \in \Omega : X(\omega) \in K(\omega)\}) = 1.$$

Ein Zufallsvektor X, der mit Wahrscheinlichkeit Eins in einem ZKK K liegt, ist demnach eine Auswahl von K. Werden zufällige Strecken der Form $K = [\mathbf{0}, X]$ mit $X \sim P$ betrachtet, so ist beispielsweise aX, mit $a \in [0,1]$ eine mögliche Auswahl von K. Mit Hilfe dieser Definition kann nun der Erwartungswert eines ZKK definiert werden.

Definition 3.3 *Gegeben sei ein ZKK* K*, so dass für jede Auswahl* X *von* K *gilt:* $E(X) < \infty$*. Der Erwartungswert von* K*,* $E(K)$*, ist definiert als die Menge* $\{E(X) : X$ *ist eine Auswahl von* $K\}$*.*

Somit ist der Erwartungswert eines ZKK die Menge, die durch die Erwartungswertvektoren aller Auswahlen gebildet wird. Eine hinreichende Bedingung für die Existenz des Erwartungswertes ist die Forderung, dass $E(||K||) := E(\max\{||X|| : X \text{ ist Auswahl von } K\}) < \infty$ ist. Diese Bedingung sichert zudem die Kompaktheit von $E(K)$. Die Konvexität von $E(K)$ ist sichergestellt, falls das Wahrscheinlichkeitsmaß atomfrei ist, (d.h. $P(\{K = k\}) = 0$ für alle $k \in \mathcal{B}(\mathcal{S}^d)$) (siehe Weil und Wieacker (1993)). Im Folgenden seien alle betrachteten Wahrscheinlichkeitsmaße atomfrei. Falls $K = \{X\}$ mit $X \sim P$, so ist der Erwartungswert nach Definition 3.3 äquivalent zum Erwartungswert eines Zufallsvektors, da dann die Menge aller möglichen Auswahlen nur X selbst enthält.

Da jeder konvexe Körper eindeutig durch die zugehörige Stützfunktion charakterisiert wird (siehe Lemma 2.5), basiert eine äquivalente Definition des Erwartungswerts eines zufälligen konvexen Körpers auf der Stützfunktion (siehe Vitale (1991)).

Satz 3.1 Gegeben sei ein ZKK K, mit $E(||K||) = E(\max\{||X|| : X$ ist Auswahl von $K\}) < \infty$, dann existiert für jedes $p \in \mathbf{R}^d$ mit $||p|| = 1$ der Erwartungswert $E(h(K,p))$. Die stetige und konvexe Funktion $h(p) = E(h(K,p))$ ist Stützfunktion eines konvexen Körpers. Dieser konvexe Körper ist gegeben durch $\{E(X) : X$ ist Auswahl von $K\} = E(K)$.

Beweis: (Artstein (1974))

q.e.d.

Bemerkung 3.1

1. Die Stützfunktion des Erwartungswertes eines ZKK K entspricht dem Erwartungswert der Stützfunktion von K, d.h. $h(E(K),p) = E(h(K,p))$ für alle $p \in \mathbf{R}^d$.

2. Die Bedingung $E(||K||) < \infty$ sichert die Existenz des Integrals $E(h(K,p))$. Sei $||p|| = 1$, dann folgt aus der Ungleichung von Cauchy-Schwarz:

$$\begin{aligned} \infty &> E(||K||) = E(\max\{||X|| : X \text{ ist Auswahl von } K\}) \\ &\geq E(\max\{|\langle X,p\rangle| : X \text{ ist Auswahl von } K\}) = E(h(K,p)). \end{aligned}$$

3.2 Zonoide

Bolker (1969) untersucht eine Untermenge der konvexen Körper im $\mathbf{R}^d$, die so genannten Zonoide. Zonoide weisen die Besonderheit auf, dass sie durch die Minkowski-Summe von Strecken, die auch Zonotope genannt werden, bzgl. des Hausdorff Abstandes approximiert werden können. Diese Eigenschaft wurde von Blascke (1956) erkannt. Eine erste Verbindung zwischen dem Gebiet der stochastischen Geometrie und Zonoiden von Wahrscheinlichkeitsmaßen stellt Vitale (1991) her. Für Verteilungen F, deren erstes absolutes Moment existiert, zeigen Koshevoy und Mosler (1998), dass das zugehörige Zonoid dem Erwartungswert der zufälligen Strecke $[\mathbf{0}, X]$ entspricht, wobei X eine $\mathbf{R}^d$-ZV mit $X \sim F$. In diesem Unterabschnitt werden Zonoide, die durch eine Verteilung bestimmt werden, definiert und deren Eigenschaften untersucht. Da die

Zonoide zweier verschiedener Verteilungen nicht notwendigerweise unterschiedlich sind, also vom Zonoid nicht eindeutig auf die zugehörige Verteilung geschlosssen werden kann, definieren Koshevoy und Mosler (1998) das so genannte Liftzonoid. Im Anschluss werden Möglichkeiten zur Schätzung von Zonoiden und Liftzonoiden vorgestellt.

Die Borelsche σ-Algebra im $\mathbb{R}^d$ werde mit $\mathcal{B}^d$ notiert und $\mathcal{F}$ bezeichne die Klasse aller Verteilungen auf $(\mathbb{R}^d, \mathcal{B}^d)$. In dieser Arbeit werden nur Verteilungen betrachtet, die durch das Lebesgue Maß dominiert werden. Falls $F \in \mathcal{F}$, dann wird im Folgenden, sofern nicht explizit erwähnt, sowohl die zugrunde liegende Verteilung als auch die Verteilungsfunktion durch F notiert. Ist X ein $\mathbb{R}^d$-Zufallsvektor mit Verteilung F bzw. F^X, dann bezeichne F^{AX+b} die Verteilung des ZV $AX + b$, wobei A eine $d \times d$ Matrix und $b \in \mathbb{R}^d$ ist. Die Untermenge der Verteilungen, für die das erste absolute Moment existiert, also $\int_{\mathbb{R}^d} ||x||dF(x) < \infty$ gilt, wird mit $\mathcal{F}_0$ bezeichnet.

Angelehnt an die Definition von Bolker (1969) definieren Koshevoy und Mosler (1998) das Zonoid einer Verteilung wie folgt:

Definition 3.4 *Sei $F \in \mathcal{F}_0$. Der Punkt $\xi(F, g) \in \mathbb{R}^d$ sei für eine beliebige meßbare Abbildung $g : \mathbb{R}^d \to [0, 1]$ definiert als:*

$$\xi(F, g) = \int_{\mathbb{R}^d} g(x)xdF(x) \in \mathbb{R}^d.$$

Dann heißt die Menge

$$Z(F) := \left\{\xi(F, g) :\ g : \mathbb{R}^d \to [0, 1] \ \textit{meßbar}\right\} \subset \mathbb{R}^d,$$

Zonoid von F.

Das Zonoid $Z(\cdot)$ ist somit eine Abbildung der Verteilungsklasse $\mathcal{F}_0$ in eine Teilmenge des $\mathbb{R}^d$.

Im Folgenden werden einige Charakteristika von Zonoiden zusammengestellt.

Lemma 3.1 1. Das Zonoid von F ist symmetrisch um $\frac{1}{2}E(X)$, wobei $E(X)$ der Erwartungswert von $X \sim F$ ist.

2. Das Zonoid von F enthält $\mathbf{0}$ und ist Teilmenge des $\mathbb{R}^d$.

Beweis: (Bolker (1969))

q.e.d.

Da nach Lemma 3.1 das Zonoid ein konvexer Körper ist, kann die Stützfunktion von $Z(F)$ ermittelt werden. Diese ist gegeben durch:

Lemma 3.2 Sei $F \in \mathcal{F}_0$ und $p \in \mathbb{R}^d$. Die Stützfunktion von $Z(F)$ ist dann:

$$h(Z(F), p) = \int_{\{x:\langle x,p\rangle \geq 0\}} \langle x, p\rangle dF(x) = \int_{\mathbb{R}^d} \max\{0, \langle x, p\rangle\} dF(x).$$

Beweis: (Koshevoy und Mosler (1998))
Nach Definition ist:

$$h(Z(F), p) = \max\{\langle \xi(F, g), p\rangle : \xi(F, g) \in Z(F)\},$$

mit $\xi(F, g) = \int_{\mathbb{R}^d} g(x) x dF(x)$.
Für jede messbare Abbildung $g : \mathbb{R}^d \to [0, 1]$ und für jedes $p \in \mathbb{R}^d$ gilt:

$$g(x)\langle x, p\rangle \leq I_{H_\mathbf{0}}(x)\langle x, p\rangle,$$

wobei $I_{H_\mathbf{0}}(x)$ die Indikatorfunktion über dem geschlossenen Halbraum $H_\mathbf{0} = \{x \in \mathbb{R}^d : \langle x, p\rangle \geq 0\}$, der durch den Ursprung $\mathbf{0}$ führt, bezeichnet.
Es ergibt sich also:

$$\begin{aligned} \langle \xi(F, g), p\rangle &= \int_{\mathbb{R}^d} g(x)\langle x, p\rangle dF(x) \leq \int_{\mathbb{R}^d} I_{H_\mathbf{0}}(x)\langle x, p\rangle dF(x) \\ &= \langle \xi(F, I_{H_\mathbf{0}}), p\rangle \\ &= \int_{H_\mathbf{0}} \langle x, p\rangle dF(x) = \int_{\mathbb{R}^d} \max\{0, \langle x, p\rangle\} dF(x). \end{aligned}$$

Da $\xi(F, I_{H_\mathbf{0}}) \in Z(F)$, folgt die Behauptung.

q.e.d.

Bemerkung 3.2 Für ein festes $p \in \mathbb{R}^d$ ist der Punkt $\xi(F, I_{H_\mathbf{0}}) = \int_{\{x:<x,p>\geq 0\}} x dF(x)$ der Punkt aus $Z(F)$, der den Wert der Stützfunktion $h(Z(F), p)$ maximiert. Er ist somit auch Randpunkt des konvexen Körpers $Z(F)$. Werden alle geschlossenen Halbräume der Form $\mathcal{H}_{\mathbf{0}}^d = \Big\{ \{x \in \mathbb{R}^d : \langle x, p \rangle \geq 0\} : ||p|| = 1 \Big\}$ betrachtet, so entspricht die konvexe Hülle der Randpunkte $\xi(F, I_{H_{\mathbf{0}}})$ dem konvexen Körper $Z(F)$. Somit ist:

$$Z(F) = conv \left\{ \left(\int_{H_{\mathbf{0}}} x dF(x) \right) : H_{\mathbf{0}} \in \mathcal{H}_{\mathbf{0}}^d \right\}. \tag{3.1}$$

Beispiel 3.1 Sei $F = N(\mu, \Sigma)$ die multivariate Normalverteilung mit Erwartungswert $\mu \in \mathbb{R}^d$ und der Kovarianzmatix $\Sigma \in \mathbb{R}^{d \times d}$.
Dann ist das Zonoid gegeben durch:

$$Z(N(\mu, \Sigma)) = \left\{ x \in \mathbb{R}^d : (x - \frac{1}{2}\mu)^T \Sigma^{-1} (x - \frac{1}{2}\mu) \leq \frac{1}{2\pi} \right\}.$$

Zwei konvexe Körper sind genau dann gleich, wenn auch die zugehörigen Stützfunktionen gleich sind. Es genügt daher zu zeigen, dass die Stützfunktion von $Z(N(\mu, \Sigma))$ der Stützfunktion dieser Ellipse entspricht. Sei $p \in \mathbb{R}^d$ beliebig, aber fest.
Es ergibt sich:

$$\begin{aligned}
h(Z(F), p) &= \langle \xi(N(\mu, \Sigma)), I_{H_{\mathbf{0}}}), p \rangle = \int_{\{x : \langle x, p \rangle \geq 0\}} \langle x, p \rangle dF^X(x) \\
&= \int_{\{y : y > 0\}} y dF^Y(y), \qquad \text{wobei } Y \sim N(p^T \mu, p^T \Sigma\, p) \\
&= \int_{\{y : y > 0\}} y (2\pi)^{-1/2} (p^T \Sigma\, p)^{-1/2} \exp \left\{ \frac{-(y - p^T \mu)^2}{2 p^T \Sigma\, p} \right\} dy \\
&= \frac{1}{2} p^T \mu - \sqrt{\frac{p^T \Sigma\, p}{(2\pi)}} \exp \left\{ \frac{-y^2}{2 p^T \Sigma\, p} \right\} \Bigg|_0^\infty \\
&= \sqrt{\frac{p^T \Sigma\, p}{2\pi}} + p^T \left(\frac{1}{2} \mu \right) = \sqrt{\frac{1}{2\pi}} ||\Sigma^{1/2} p|| + \langle \frac{1}{2} \mu, p \rangle.
\end{aligned}$$

Dies entspricht der Stützfunktion der Ellipse $E := \{x \in \mathbb{R}^d : (x - \frac{1}{2}\mu)^T \Sigma^{-1} (x - \frac{1}{2}\mu) \leq \frac{1}{2\pi}\}$ (vgl. Beispiel 2.4).

Im Folgenden wird gezeigt, dass es möglich ist, das Zonoid von F als Erwartungswert einer zufälligen Strecke darzustellen.

Satz 3.2 Sei $X \sim F$. Dann entspricht der Erwartungswert der zufälligen Strecke $[\mathbf{0}, X]$ dem Zonoid von F, d.h.:

$$Z(F) = E([\mathbf{0}, X]).$$

Beweis: (Koshevoy und Mosler (1998)):
Die beiden kompakten Mengen $Z(F)$ und $E([\mathbf{0}, X])$ sind genau dann gleich, wenn ihre Stützfunktionen gleich sind. Nach Lemma 3.2 gilt:

$$h(Z(F), p) = \int_{\mathbf{R}^d} \max\{0, \langle x, p\rangle\} dF(x).$$

Andererseits ist $h([\mathbf{0}, X], p) = \max\{0, \langle X, p\rangle\}$.
Daraus folgt:

$$h(E([\mathbf{0}, X], p) \overset{Satz 3.1}{=} E(h([\mathbf{0}, X], p)) = \int_{\mathbf{R}^d} \max\{0, \langle x, p\rangle\} dF(x) = h(Z(F), p).$$

Somit folgt die Behauptung.

q.e.d.

Bemerkung 3.3 1. Jeder ZV $Xg(X)$, wobei die Abbildung g eine beliebige aber feste messbare Abbildung $g : \mathbf{R}^d \to [0, 1]$, ist eine Auswahl des ZKK $[\mathbf{0}, X]$.
2. Die Einschränkung auf die Verteilungsklasse $\mathcal{F}_0$ sichert die Existenz von $K = E([\mathbf{0}, X])$, da gilt:

$$E(||[\mathbf{0}, X]||) = \int_{\mathbf{R}^d} ||x|| dF(x) < \infty.$$

3. Für den Fall, dass X multivariat standardnormalverteilt ist, entspricht das Zonoid bzw. der Erwartungswert der zufälligen Strecke $[\mathbf{0}, X]$ einer Kugel um den Ursprung mit Radius $\frac{1}{\sqrt{2\pi}}$.

Nach Lemma 3.1 ist $Z(F)$ symmetrisch um den Punkt $\frac{1}{2}E(X)$ und daher kann die Stützfunktion des zentrierten Zonoids $Z^+(F) = Z(F) \oplus \frac{-1}{2}E(X)$ in folgender Form geschrieben werden.

Lemma 3.3 Für das zentrierte Zonoid $Z^+(F)$ gilt:

$$h(Z^+(F), p) = \frac{1}{2}(h(Z(F), p) + h(Z(F), -p)) = \frac{1}{2} \int_{\mathbf{R}^d} |\langle x, p\rangle| dF(x).$$

Beweis:

Da $Z^+(F)$ symmetrisch um den Ursprung **0** ist, gilt für die Stützfunktion: $h(Z^+(F),p) = h(Z^+(F),-p)$.

Somit ist:

$$\begin{aligned}
2h(Z^+(F),p) &= h(Z^+(F),p) + h(Z^+(F),-p) \\
&= h(Z(F),p) - \langle \frac{-1}{2}E(X), -p\rangle + h(Z(F),-p) - \langle \frac{-1}{2}E(X), -p\rangle \\
&= h(Z(F),p) + h(Z(F),-p) \\
&= \int_{\{x:\langle x,p\rangle \geq 0\}} \langle x,p\rangle dF(x) + \int_{\{x:\langle x,-p\rangle \geq 0\}} \langle x,-p\rangle dF(x) \\
&= \int_{\{x:\langle x,p\rangle \geq 0\}} |\langle x,p\rangle| dF(x) + \int_{\{x:\langle x,-p\rangle \geq 0\}} |\langle x,-p\rangle| dF(x) \\
&= \int_{\mathbb{R}^d} |\langle x,p\rangle| dF(x).
\end{aligned}$$

Hieraus folgt die Behauptung.

q.e.d.

Bemerkung 3.4 Aus Bemerkung 3.2 folgt, dass zwei Zonoide genau dann gleich sind, wenn beide Körper dieselben Randpunkte besitzen. Es gilt $Z(F) = Z(G)$ für $F, G \in \mathcal{F}_0$ genau dann, wenn

$$\int_{\{x:\langle x,p\rangle \geq 0\}} x dF(x) = \int_{\{x:\langle x,p\rangle \geq 0\}} x dG(x) \text{ für jedes } p \in \mathbb{R}^d \text{ mit } ||p|| = 1.$$

Für den Fall $d = 1$ bedeutet dies, dass die Zonoide zweier Verteilungen genau dann übereinstimmen, wenn gilt:

$$\int_0^\infty x dG(x) = \int_0^\infty x dF(x) \quad \text{und} \quad \int_{-\infty}^0 x dG(x) = \int_{-\infty}^0 x dF(x).$$

Werden die Zufallsvariablen X und Y betrachtet, wobei $X \sim N(0,1) = F$ und $Y \sim U[-\frac{4}{\sqrt{2\pi}}, \frac{4}{\sqrt{2\pi}}] = G$.

Dann gilt: $\int_0^\infty x dF(x) = \int_0^\infty y dG(y)$ und $\int_{-\infty}^0 y dG(y) = \int_{-\infty}^0 x dF(x)$.

Somit ist $Z(F) = Z(G) = \{x : x \in [-\frac{1}{\sqrt{2\pi}}, \frac{1}{\sqrt{2\pi}}]\}$, d.h. die Zonoide beider betrachteter Verteilungen stimmen überein.

Ebenfalls ergibt sich die Gleichheit der Stützfunktionen: $h(Z(F),p) = h(Z(G),p) =$

$\frac{1}{\sqrt{2\pi}}|p|$, $p \in \mathbb{R}$.
Zwar wird jeder Verteilung in eindeutiger Weise genau ein Zonoid zugeordnet, die Umkehrung trifft aber im Allgemeinen nicht zu.

3.3 Liftzonoide

Es gibt mehrere Möglichkeiten, eine Verteilung eindeutig zu charakterisieren. Zu erwähnen sind beispielsweise die Verteilungsfunktion, die charakteristische Funktion oder eine Dichte. Im Folgenden wird ein konvexer Körper vorgestellt, der eine Verteilung eindeutig beschreibt.
In Bemerkung 3.4 wurde gezeigt, dass Zonoide die Verteilung nicht eindeutig charakterisieren. Koshevoy und Mosler (1998) definieren deshalb das Liftzonoid einer Verteilung, einen konvexen Körper, durch den eine Verteilung, deren erstes absolutes Moment existiert, eindeutig festgelegt wird.

Definition 3.5 *Sei $F \in \mathcal{F}_0$ und X ein $\mathbb{R}^d$ ZV, mit $X \sim F$. Dann heißt die Menge:*

$$LZ(F) := E([\mathbf{0}, (1, X^T)^T]) \subset \mathbb{R}^{d+1},$$

das Liftzonoid von F.

Bemerkung 3.5 Sei $X \in \mathbb{R}^d$ ein gemäß F verteilter ZV. Wird nun der ZV $(1, X^T)^T$ betrachtet, so ist der Stichprobenraum dieses ZV als Teilmenge der Hyperebene $E_1 = \{x = (x_0, x_1, ..., x_d)^T \in \mathbb{R}^{d+1} : x_0 = 1\}$ im $\mathbb{R}^{d+1}$ eingebettet. Der ZV wird also in E_1 realisiert. Wird die Verteilung des ZV $(1, X^T)^T \in \mathbb{R}^{d+1}$ mit $F^{(1,X^T)^T}$ bezeichnet, dann gilt:

$$Z(F^{(1,X^T)^T}) = LZ(F), \quad Z(F) = proj_d(LZ(F)), \tag{3.2}$$

wobei $proj_d$ die Projektion in die letzten d Koordinaten bezeichne. Ein Liftzonoid ist somit ein Zonoid bzgl. der Verteilung $F^{(1,X^T)^T}$. Daraus folgt unmittelbar:

Lemma 3.4 1. Jedes Liftzonoid von $F \in \mathcal{F}_0$ ist ein konvexer Körper, enthält den Ursprung $\mathbf{0}$ und ist Teilmenge des $\mathbb{R}^{d+1}$.

2. Falls $F \in \mathcal{F}_0$, dann ist das zugehörige Liftzonoid symmetrisch um $\frac{1}{2}(1, E(X)^T)^T$, wobei $E(X)$ der Erwartungswert von $X \sim F$ ist.

Beweis: (Bolker (1969))

q.e.d.

Da nach Lemma 3.4 $LZ(F)$ ein konvexer Körper ist, ist das Liftzonoid durch die zugehörige Stützfunktion eindeutig charakterisiert.

Lemma 3.5 Sei $F \in \mathcal{F}_0$ und $(p_0, p^T)^T \in \mathbb{R}^{d+1}$, wobei $p_0 \in \mathbb{R}$. Die Stützfunktion von $LZ(F)$ ist dann:

$$\begin{aligned} h(LZ(F), (p_0, p^T)^T) &= \int_{\{x:p_0+\langle x,p\rangle \geq 0\}} p_0 + \langle x, p\rangle dF(x) \\ &= \int_{\mathbb{R}^d} \max\{0, p_0 + \langle x, p\rangle\} dF(x). \end{aligned}$$

Beweis:

Analog zum Beweis von Lemma 3.2.

q.e.d.

Bemerkung 3.6 Zu beachten ist, dass die Stützfunktion nicht von der Verteilung $F^{(1,X^T)^T}$ abhängt, sondern nur von F.

Lemma 3.6 Die folgenden Definitionen der Abbildung $LZ : \mathcal{F}_0 \to \mathcal{K}^{d+1}$ sind äquivalent:

$$\begin{aligned} 1.\ LZ(F) &= E\left(\left[\mathbf{0}, (1, X^T)^T\right]\right), \\ 2.\ LZ(F) &= \left\{\left(\int g(x)dF(x), \int xg(x)dF(x)^T\right)^T : g\colon \mathbb{R}^d \to [0,1] \text{ messbar}\right\}, \\ 3.\ LZ(F) &= conv\left\{\left(\int_{H_{p_0}} dF(x), \int_{H_{p_0}} xdF(x)^T\right)^T : H_{p_0} \in \mathcal{H}^d_{p_0}\right\}, \end{aligned}$$

wobei $\mathcal{H}^d_{p_0}$ die Menge aller geschlossenen Halbräume der Form $H_{p_0} = \{x \in \mathbb{R}^d : p_0 + \langle x, p\rangle \geq 0;\ p \in \mathbb{R}^d \text{ und } p_0 \in \mathbb{R}\}$ ist.

Beweis:

In Analogie zum Beweis von Satz 3.2.

q.e.d.

Satz 3.3 Jede Verteilung $F \in \mathcal{F}_0$ ist eindeutig durch das zugehörige Liftzonoid $LZ(F)$ festgelegt.

Bevor dieser Satz bewiesen werden kann, wird folgende Aussage benötigt.

Lemma 3.7 Bezeichne F_p die Verteilung der Zufallsvariable $\langle X, p\rangle$, wobei $p \in \mathbb{R}^d$ und X ein ZV mit zugehöriger Verteilung F sei. Entsprechendes gilt für G_p.
Dann gilt:

$$LZ(F) = LZ(G) \text{ genau dann, wenn } LZ(F_p) = LZ(G_p) \text{ für alle } p \in \mathbb{R}^d.$$

Beweis: (Koshevoy und Mosler (1998))

q.e.d.

Bemerkung 3.7 1. Für die zugehörigen Stützfunktionen von F, G, F_p und G_p gilt die folgende Beziehung aus Lemma 3.7:

$$h(LZ(F),\cdot) = h(LZ(G),\cdot) \Longleftrightarrow h(LZ(F_p),\cdot) = h(LZ(G_p),\cdot) \text{ für alle } p \in \mathbb{R}^d.$$

Somit ist ein Liftzonoid durch seine zweidimensionalen Projektionen eindeutig festgelegt.
2. Für den Fall $d = 1$ gilt wegen Punkt 3 in Lemma 3.6, dass sich $LZ(F)$ bestimmen lässt aus der konvexen Hülle von:

$$\left\{ \left(\int_{(-\infty,y]} dF(x), \int_{(-\infty,y]} x dF(x) \right)^T, \quad \left(\int_{[y,+\infty)} dF(x), \int_{[y,+\infty)} x dF(x) \right)^T \right\} \quad \forall\, y \in \mathbb{R}.$$

Diese Punkte lassen sich auch schreiben als:

$$\left(t, \int_0^t F^{-1}(s)ds\right)^T, \quad \text{und } \left(t, \int_{1-t}^1 F^{-1}(s)ds\right)^T \quad t \in [0,1],$$

wobei $F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$ das α-Quantil von F ist.

Beweis von Satz 3.3:
Seien $F, G \in \mathcal{F}_0$ und sei der ZV X gemäß F und Y gemäß G verteilt.
Zu zeigen ist: Falls $LZ(F) = LZ(G)$, dann gilt $F = G$.

Nach Bemerkung (3.7.1) folgt, dass $LZ(F) = LZ(G)$ genau dann gilt, wenn für alle $p \in \mathbf{R}^d$ und $q \in \mathbf{R}^2$: $h(LZ(F_p), q) = h(LZ(G_p), q)$.

Somit gilt nach Punkt 2 von Bemerkung 3.7 für festes p, dass genau ein $t \in [0,1]$ existiert, so dass:

$$g(t) := \int_0^t F_p^{-1}(s)d(s) = \int_0^t G_p^{-1}(s)d(s) =: f(t) \text{ für alle } t \in [0,1].$$

Nun folgt hieraus, dass $\frac{\partial f(t)}{\partial t} = F_p^{-1}(t) = G_p^{-1}(t) = \frac{\partial g(t)}{\partial t}$ für alle $t \in [0,1]$, d.h $F_p = G_p$. Somit gilt für alle $p \in \mathbf{R}^d$, $\langle X, p\rangle \stackrel{d}{=} \langle Y, p\rangle$, wobei $\stackrel{d}{=}$ die Gleichheit der Verteilungen der ZV beschreibt. Nun folgt aus dem Cramer-Wold Theorem (siehe z.B. Mardia (1979)) die Gleichheit der Verteilungen der ZV X und Y und somit $F = G$.

q.e.d.

Durch die eindeutige Zuordnung zwischen Liftzonoiden und Verteilungen und der Tatsache, dass Liftzonoide konvexe Körper sind, kann ein Abstand zwischen Verteilungen wie folgt definiert werden:

Definition 3.6 *Sei $d : \mathcal{F}_0 \times \mathcal{F}_0 \to \mathbb{R}_+$ eine Funktion der folgenden Form:*

$$d(F_1, F_2) := d_H(LZ(F_1), LZ(F_2)), \quad F_1, F_2 \in \mathcal{F}_0,$$

wobei d_H der Hausdorff Abstand ist. Dann ist $d(\cdot,\cdot)$ eine Metrik auf dem Raum aller Verteilungen, deren erstes Moment existiert.

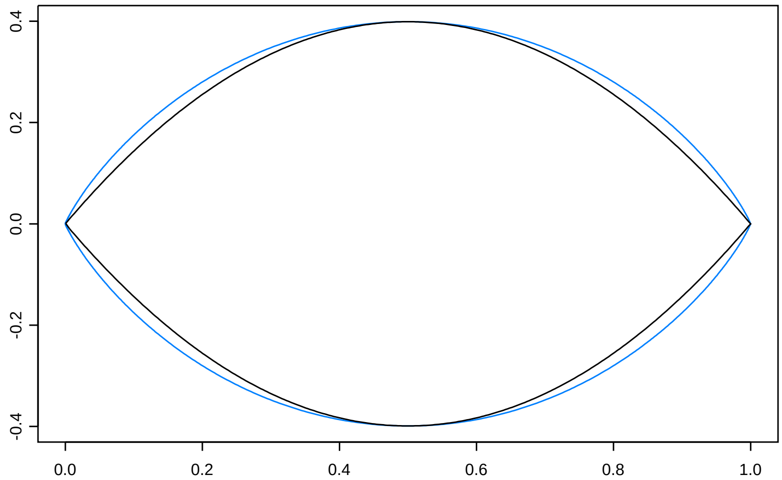
Bemerkung 3.8 Eine äquivalente Definition dieser Metrik ist wegen Formel (2.7) gegeben durch:

$$d_H(F_1, F_2) = \max_{||(p_0, p^T)^T||=1} |h(LZ(F_1), (p_0, p^T)^T) - h(LZ(F_2), (p_0, p^T)^T)|, \quad F_1, F_2 \in \mathcal{F}_0,$$

wobei h jeweils die zugehörige Stützfunktion bezeichnet.

In Abbildung 3.2 sind die Liftzonoide einer Standardnormalverteilung $N(0,1)$ (blau) und einer Gleichverteilung auf dem Intervall $[-\frac{4}{\sqrt{2\pi}}, \frac{4}{\sqrt{2\pi}}]$ (schwarz) dargestellt. In der Abbildung sind die Unterschiede zwischen diesen beiden konvexen Körpern deutlich zu erkennen (vgl. Beispiel 3.4).

Abbildung 3.2: Liftzonoide einer Normalverteilung $N(0,1)$ (blau) und einer Gleichverteilung $U[-\frac{4}{\sqrt{2\pi}}, \frac{4}{\sqrt{2\pi}}]$



3.3.1 Die Schätzung von Liftzonoiden und Zonoiden

In diesem Abschnitt wird ein möglicher Ansatz zur Schätzung von Liftzonoiden und Zonoiden von Verteilungen vorgestellt. Ein naheliegender Ansatz besteht darin, die Verteilung durch die empirische Verteilung zu ersetzen.

Lemma 3.8 Seien $X_1, ..., X_n$ u.i.v. $\mathbf{R}^d$-ZV mit $X_i \sim F \in \mathcal{F}_0$, $i = 1, ..., n$. Dann gilt:

$$LZ(F_n) = \frac{1}{n} \bigoplus_{i=1}^{n} [\mathbf{0}, (1, X_i^T)^T].$$

Beweis:

Die empirische Verteilung F_n von $X_1, ...X_n$ ist definiert als:

$F_n = \frac{1}{n} \sum_{i=1}^{n} I_{X_i}$, dabei bezeichne I_X die Einpunktverteilung in X.

Um die Gleichheit der konvexen Körper zu zeigen, reicht es aus, die Gleichheit beider zugehöriger Stützfunktionen zu zeigen. Es gilt:

$$h(LZ(F_n), p) \quad = \quad \int_{\mathbf{R}^d} \max\{0, p_0 + \langle x, p \rangle\} dF_n(x)$$

$$
\begin{aligned}
&= \int_{\mathbf{R}^d} \max\{0, p_0 + \langle x, p\rangle\} d\frac{1}{n}\sum_{i=1}^n I_{X_i}(x) \\
&= \frac{1}{n}\sum_{i=1}^n \int_{\mathbf{R}^d} \max\{0, p_0 + \langle x, p\rangle\} dI_{X_i}(x) \\
&= \frac{1}{n}\sum_{i=1}^n \max\{0, p_0 + \langle X_i, p\rangle\} \\
&= h(\frac{1}{n}\bigoplus_{i=1}^n [\mathbf{0}, (1, X_i^T)^T], p).
\end{aligned}
$$

Das letzte Gleichheitszeichen folgt aus den Eigenschaften der Stützfunktion (siehe Lemma 2.2 und Lemma 3.5).

q.e.d.

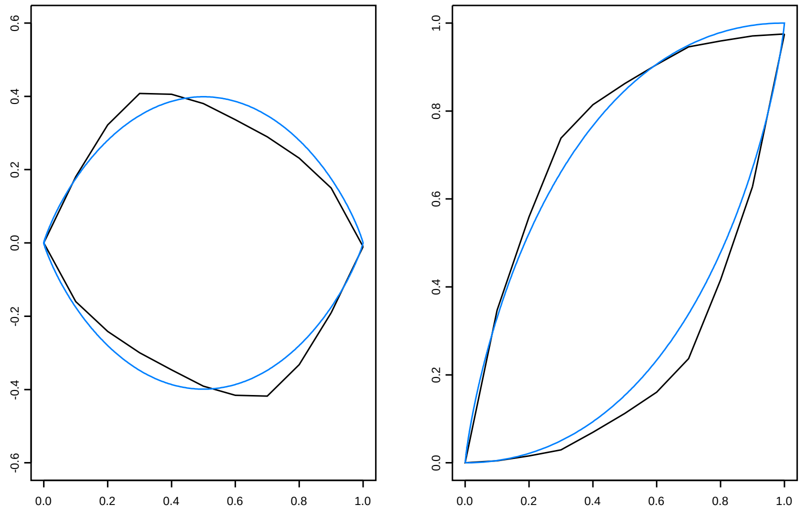
Bemerkung 3.9 1. Somit entspricht das Liftzonoid der empirischen Verteilung F_n der Minkowski-Summe der zufälligen Strecken $\frac{1}{n}[\mathbf{0}, (1, X_i^T)^T]$, $i = 1, ..., n$.
2. In der Literatur werden konvexe Körper, die durch eine endliche Minkowski-Summe von Strecken entstehen, auch Zonotope genannt.
3. Die in Lemma 3.8 vorgestellte Schätzung des Liftzonoids, die auf der empirischen Verteilung beruht, heißt zufälliges Liftzonotop $LZ(F_n) =: LZ(X_1, ..., X_n)$.
4. Da zufällige Liftzonotope spezielle zufällige Polytope sind (darunter wird die zufällige konvexe Hülle einer endlichen zufälligen Punktmenge verstanden), ergibt sich unmittelbar:

$$
\begin{aligned}
LZ(F_n) &= conv\Big\{\sum_{i=1}^n \tau_i\Big(\frac{1}{n}, \frac{X_i^T}{n}\Big)^T : \tau_i \in \{0,1\}, i \in N\Big\} && (3.3) \\
&= conv\Big\{\bigcup_{k=0}^n \{\sum_{j=1}^k \Big(\frac{1}{n}, \frac{X_{i_j}^T}{n}\Big)^T\} : \{i_1, ..., i_k\} \subset N\Big\}, && (3.4)
\end{aligned}
$$

wobei $N = \{1, ..., n\}$. Für die Äquivalenz dieser Darstellungen sei beispielsweise auf Grünbaum (1967) verwiesen. Die Schätzung eines Liftzonoids basiert demnach auf der Bestimmung der konvexen Hülle einer endlichen Punktemenge.

Satz 3.4 Sei $F \in \mathcal{F}_0$ und F_n die empirische Verteilung der u.i.v. Zufallsvektoren $X_1, ..., X_n$, wobei $X_i \sim F, i = 1, ..., n$. Dann konvergiert $LZ(F_n)$ bzgl. des Hausdorff Abstandes fast sicher gegen $LZ(F)$.

Abbildung 3.3: Liftzonoid (blau) und Schätzung eines Liftzonoid (schwarz) einer $N(0,1)$-Verteilung und einer $Exp(1)$-Verteilung mit jeweils 10 Beobachtungen



Beweis: (Koshevoy und Mosler (1998))

q.e.d.

Abbildung 3.3 illustriert die Schätzung von Liftzonoiden für den Fall, dass die Beobachtungen Realisierungen einer univariaten standardormal- bzw. exponentialverteilten Zufallsvariablen sind (blau). Es wird eine Stichprobe vom Umfang $n = 10$ betrachtet. In schwarzer Farbe sind die Liftzonoide der entsprechenden Verteilungen dargestellt. Deutlich zu erkennen ist, dass beide Schätzungen jeweils die Punkte $(0,0)^T$ und $(1,\bar{x})^T$ beinhalten, wobei $\bar{x}$ das arithmetische Mittel der Stichprobe ist.
Für die Schätzung von Zonoiden von Verteilungen ist ein analoges Vorgehen möglich:

Lemma 3.9 Seien $X_1, ..., X_n$ u.i.v. $\mathbf{R}^d$-ZV mit $X_i \sim F \in \mathcal{F}_0$, $i = 1, ..., n$. Das Zonoid $Z(F_n)$ der empirischen Verteilung F_n von $X_1, ..., X_n$ ist gegeben durch:

$$Z(F_n) = proj_d(LZ(F_n)) = \frac{1}{n}\bigoplus_{i=1}^{n}[\mathbf{0}, X_i] = conv\Big\{\bigcup_{k=0}^{n}\{\sum_{j=1}^{k}\frac{X_{i_j}}{n} : \{i_1, ..., i_k\} \subset N\}\Big\},$$

wobei $N = \{1, ..., n\}$. Der konvexe Körper $Z(F_n)$ wird zufälliges Zonotop genannt.

Beweis:

Da das Zonoid der empirischen Verteilung der Projektion eines zufälligen Liftzonotops auf die letzten d Koordinaten entspricht (siehe Formel (3.2)), folgt unmittelbar Korollar 3.9.

q.e.d.

Bemerkung 3.10 1. Für Zonoide von Verteilungen gilt genau wie für Liftzonoide, dass $Z(F_n)$ bzgl. des Hausdorff Abstandes fast sicher gegen $Z(F)$ konvergiert.
2. Für den Fall $d = 1$ wird $Z(F)$ geschätzt durch:

$$Z(F_n) = \left[\frac{1}{n} \sum_{i=1}^{n} X_i(1 - I_{[0,\infty)}(X_i)), \frac{1}{n} \sum_{i=1}^{n} X_i I_{[0,\infty)}(X_i) \right] \tag{3.5}$$

3. Da die Schätzung von Liftzonoiden und Zonoiden im gleichen Maße von der empirischen Verteilung wie auch von der Stichprobe abhängt, wird folgende Schreibweise vereinbart: $LZ(\tilde{X}_n) := LZ(F_n)$ bzw. $Z(\tilde{X}_n) := Z(F_n)$.

3.3.2 Robustheitsuntersuchungen von Zonoiden

Die Schätzung der in diesem Kapitel vorgestellten konvexen Körper basiert auf der konvexen Hülle aus Mittelwerten geeignet gewählter Teilstichproben. Dieses Vorgehen legt die Vermutung nahe, dass schon eine einzelne Beobachtung ausreicht, die Schätzung zusammenbrechen zu lassen. Mit Hilfe des Explosions-Bruchpunktes aus Definition 2.6 wird diese Vermutung bestätigt.

Lemma 3.10 Gegeben sei eine Stichprobe $\tilde{X}_n = \{x_1, ..., x_n\}$ vom Umfang n mit $x_i \in \mathbf{R}^d$ Realisationen auf $F \in \mathcal{F}_0$. Dann gilt für den Schätzer $LZ(\tilde{X}_n) = \frac{1}{n} \bigoplus_{i=1}^{n} [\mathbf{0}, (1, x_i^T)^T]$:

$$\epsilon_K(\tilde{X}_n, LZ) = \frac{1}{n}.$$

Beweis:

Zu zeigen ist, dass die durch den Austausch einer einzigen Beobachtung entstandene kontaminierte Stichprobe ausreicht, die Schätzung zusammenbrechen zu lassen.

Sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine Stichprobe vom Umfang n, mit $x_i \in \mathbf{R}^d$ und $\tilde{Y}_{n,1} := \{x_{i_1}, ..., x_{i_{n-1}}, y\}$, eine durch Austausch der Beobachtung x_{i_n} aus $\tilde{X}_n$ entstehende Stichprobe. Zu zeigen ist:

$$\sup_{\tilde{Y}_{n,1}} \left(\max_{||(p_0,p^T)^T||=1} \left| h(LZ(\tilde{X}_n), (p_0, p^T)^T) - h(LZ(\tilde{Y}_{n,1}), (p_0, p^T)^T) \right| \right) = \infty,$$

wobei $h(LZ(\tilde{X}_n, p)) = \frac{1}{n} \sum_{i=1}^n \max\{0, p_0 + \langle x_i, p \rangle\}$ (vgl. Beweis von Lemma 3.8).
Sei oBdA $y \in \mathbf{R}^d_+$. Dabei bezeichne $\mathbf{R}^d_+$ die Untermenge aller Punkte im $\mathbf{R}^d$, die nur positive Komponenten enthalten. Somit existieren ein $p_0 \in \mathbf{R}$ und $p \in \mathbf{R}^d$ mit $||(\tilde{p_0}, \tilde{p}^T)^T|| = 1$, so dass

$$\tilde{p_0} + \langle y, \tilde{p} \rangle > 0. \tag{3.6}$$

Mit $P = \{(p_0, p^T)^T \in \mathbf{R}^{d+1} : ||(p_0, p^T)^T|| = 1, p_0 + \langle y, p \rangle > 0\}$ werde die Menge aller Punkte bezeichnet, welche die Ungleichung (3.6) erfüllen. Dann ist $\tilde{p}_1 = (0, \frac{y}{||y||}^T)^T \in P$.
Nun gilt:

$$\begin{aligned}
d_H\Big(LZ(\tilde{X}_n), LZ(\tilde{Y}_{n,1})\Big) &\geq \max_{(p_0,p^T)^T \in P} \Big| h(LZ(\tilde{X}_n), p) - h(LZ(\tilde{Y}_{n,1}), p) \Big| \\
&= \max_{(p_0,p^T)^T \in P} \Bigg| \frac{1}{n} \sum_{\tilde{X}_n} \max\{0, p_0 + \langle x_i, p \rangle\} \\
&- \frac{1}{n} \Big(\sum_{\tilde{Y}_{n,1} \setminus \{y\}} \max\{0, p_0 + \langle x_i, p \rangle\} + \langle y, p \rangle + p_0 \Big) \Bigg| \\
&\geq \Bigg| \frac{1}{n} \sum_{\tilde{X}_n} \max \left\{ 0, \left\langle x_i, \frac{y}{||y||} \right\rangle \right\} \\
&- \frac{1}{n} \left(\sum_{\tilde{Y}_{n,1} \setminus \{y\}} \max \left\{ 0, \left\langle x_i, \frac{y}{||y||} \right\rangle \right\} + \left\langle y, \frac{y}{||y||} \right\rangle \right) \Bigg| \\
&= \begin{cases} \left| \frac{1}{n} \left(\left\langle x_{i_n}, \frac{y}{||y||} \right\rangle - ||y|| \right) \right| & : \text{ falls } \left\langle x_{i_n}, \frac{y}{||y||} \right\rangle > 0 \\ \frac{1}{n} \, ||y|| & : \text{ sonst} \end{cases} \\
&\longrightarrow \infty, \text{ falls } ||y|| \to \infty.
\end{aligned}$$

Somit ist der Bruchpunkt $\epsilon_K(\tilde{X}_n, LZ) = \frac{1}{n}$. Eine Beobachtung reicht aus, den Schätzer zusammenbrechen zu lassen.

q.e.d.

Bemerkung 3.11 1. Da $LZ(\tilde{X}_n)$ nicht den Ursprung in seinem Inneren enthält, ist $\epsilon_{K,K^*}(\tilde{X}_n, LZ)$ (siehe Definition 2.10) nicht definiert.
2. Für den Schätzer $Z(\tilde{X}_n)$ kann gezeigt werden, dass gilt $\epsilon_{K,K^*}(\tilde{X}_n, Z) = 1/n$. Somit reicht eine Beobachtung aus, den Schätzer zusammenbrechen zu lassen. Der Beweis verläuft analog zum Beweis von Lemma 3.10.

Kapitel 4

Kontur-Toleranzintervalle und Datentiefe-Funktionen

Kontur-Toleranzintervalle und Datentiefe-Funktionen sind zu einem wichtigen Werkzeug in der modernen Statistik geworden. Anwendung finden sie in der explorativen Datenanalyse hochdimensionaler Daten und in der nichtparametrischen Statistik.
Unter Kontur-Toleranzintervallen werden Bereiche des $\mathbf{R}^d$ verstanden, in denen Punkte mindestens in einer festgelegten „Entfernung“ vom Zentrum der Verteilung entfernt liegen. Diese Bereiche erlauben somit Aussagen über die Form und Struktur eines hochdimensionalen Datensatzes und können als Analogon zu univariaten Quantilsintervallen angesehen werden.
Kontur-Toleranzbereiche sind eng verbunden mit der so genannten Datentiefe-Funktion. Diese Funktion gibt an, wie weit ein Punkt $x \in \mathbf{R}^d$ vom Zentrum der Verteilung F entfernt ist. Auf diese Weise entsteht eine „natürliche“Anordnung von Punkten im $\mathbf{R}^d$ bzgl. eines Zentrums und zwar vom Zentrum der Verteilung nach außen führend. Die erste Datentiefe-Funktion stammt von Tukey (1975). Weitere Datentiefe-Funktionen wurden beispielsweise von Liu (1990), Donoho und Gasko (1992), Liu und Singh (1992) und Koshevoy und Mosler (1997) vorgeschlagen.
Im ersten Abschnitt wird in allgemeiner Form der Begriff des Kontur-Toleranzbereiches

für unimodale Verteilungen definiert. Basierend auf dieser Definition ergibt sich die Definition der Datentiefe-Funktion. Im Abschnitt 4.2 werden zwei Beispiele vorgestellt. Zum einen werden die von Liu (1992) vorgeschlagenen Mahalanobis-Konturen und die (α-getrimmten) zonoiden Zonen (Koshevoy und Mosler (1997)) beschrieben und deren Schätzungen auf ihr Bruchpunktverhalten hin untersucht.

4.1 Einführung

Kontur-Toleranzbereiche werden um ein Zentrum $\theta \in \mathbb{R}^d$ einer Verteilung $F \in \mathcal{F}$ konstruiert. Dieses Zentrum kann zum Beispiel dem Erwartungswert oder einem multivariaten Median einer Verteilung entsprechen. Zunächst wird eine allgemeine Definition für das Zentrum einer Verteilung gegeben.

Definition 4.1 *Sei der d-dimensionale ZV X gegeben, $X \sim F \in \mathcal{F}$ und $\chi \subset \mathbb{R}^d$ bezeichne den zugehörigen Stichprobenraum des ZV X. Dann heißt $\theta \in \mathbb{R}^d$ Zentrum von F bzgl. der Abbildung $Z_F : \chi \longrightarrow \mathbb{R}_+$ genau dann, wenn für alle $x \in \chi$ gilt:*

$$Z_F(x) \leq Z_F(\theta) < \infty.$$

Die Abbildung Z_F wird dann Kriterium genannt.

Bemerkung 4.1 Sei θ das Zentrum bzgl. Z_F. Falls $Z_F(x) < Z_F(\theta)$ für alle $x \neq \theta$, $x \in \chi$, dann besteht das Zentrum aus nur einem Punkt, d.h. es ist eindeutig.

Die folgenden Beispiele sollen Definition 4.1 verdeutlichen:

Beispiel 4.1 ($K1$) Die Untermenge der Verteilungen, für die das zweite Moment existiert wird mit $\mathcal{F}_1$ bezeichnet. Sei $Z_F(x) := \Sigma^{1/2} f(x)$, wobei f die Dichtefunktion und Σ die Kovarianzmatrix von $X \sim F \in \mathcal{F}_1$ bezeichne. Dann entspricht θ den Modalwerten von F. Falls die Verteilung F unimodal ist, besteht das Zentrum θ aus einem Punkt.
($K2$) Es werde $Z_F(x) = \max\{F(B) : B \in \mathcal{B}^d$ und $\int_B y dF(y) = x\}$ betrachtet, mit $F \in \mathcal{F}_0$. Das Maximum dieser Funktion wird angenommen, falls $B = \chi$ und x der

Erwartungswert von F ist. In diesem Fall ist θ eindeutig.

$(K3)$ Seien $x, p \in \mathbb{R}^d$ und $Z_F(x) = \inf_{||p||=1} F(H_{\langle x,p \rangle})$.
Dabei bezeichne $H_{\langle x,p\rangle} = \{y \in \mathbb{R}^d : \langle y,p\rangle \geq \langle x,p\rangle\}$. Für $d = 1$ ergibt sich gerade $Z_F(x) = \min\{F(x), 1 - F(x)\}$. Der Index $Z_F(x)$ wird maximiert, falls x dem Median von F entspricht. Für $d \geq 1$ kann der Punkt, der $Z_F(x)$ maximiert, als eine mögliche multivariate Median-Definition (Tukey Median, vgl. Tukey (1975)) angesehen werden.

Im Folgenden wird angenommen, dass das Zentrum aus einem einzelnen Punkt $\theta \in \mathbb{R}^d$ besteht und es bezeichne $\mathcal{D}^d$ die Menge aller konvexen Teilmengen des $\mathbb{R}^d$. Dann ist die Definition des Kontur-Toleranzbereiches für unimodale Verteilungen gegeben durch:

Definition 4.2 *Eine Abbildung $D_F : [0, \varepsilon] \to \mathcal{D}^d$ heißt Kontur-Toleranzbereich von F bezüglich des Kriteriums Z_F, mit $\varepsilon := \mathrm{argsup}_{\mathrm{x} \in \mathbb{R}^{\mathrm{d}}} \mathrm{Z_F(x)}$, wenn sie die folgenden Eigenschaften erfüllt:*

1. *Für ein festes $\alpha \in (0, \varepsilon]$ gilt: $D_F(\alpha) \in \mathcal{K}^d$.*

2. *Der Bereich $D_F(\alpha)$ ist affin äquivariant, d.h. $D_{F^{AX+b}}(\alpha) = AD_{F^X}(\alpha) \oplus\ b$, für jeden ZV $X \in \mathbb{R}^d$, jede nichtsinguläre $d \times d$ Matrix A und jeden Vektor $b \in \mathbb{R}^d$.*

3. *Es gilt: $\max_\alpha D_F(\alpha) = \theta \in \mathbb{R}^d$.*

4. *Die Bereiche $D_F(\alpha)$ sind verschachtelt, d.h. für $\alpha_1 \geq \alpha_2$ gilt: $D_F(\alpha_1) \subseteq D_F(\alpha_2)$.*

5. *Für $\alpha \longrightarrow 0$ gilt:*

$$D_F(\alpha) \longrightarrow \bigcap \{H_{p_0} \in \mathcal{H}^d_{p_0} : F(H_{p_0}) = 1\} = \chi,$$

wobei $H := H_{p_0} = \{x \in \mathbb{R}^d : \langle x,p\rangle \geq p_0\}$ den geschlossenen Halbraum bezeichne und $\mathcal{H}^d := \mathcal{H}^d_{p_0} = \{H_{p_0} : p \in \mathbb{R}^d,\ p_0 \in \mathbb{R}\}$ die Menge aller geschlossenen Halbräume.

Bemerkung 4.2 1. Punkt 4 in Definition (4.2) sagt aus, dass das Kontur-Toleranzintervall für $\alpha = 0$ der konvexen Hülle des Stichprobenraumes der Verteilung

entspricht.
2. Definition 4.2 ist nur für unimodale Verteilungen sinnvoll, da sonst die Struktur der Verteilung nicht widergespiegelt wird.

Beispiel 4.2 Gegeben sei das Kriterium (K1) aus Beispiel 4.1:
Für einen ZV $X \in \mathbb{R}^d$ mit Dichtefunktion f und Kovarianzmatrix Σ erfüllen die Bereiche:

$$L_F(\alpha) := \{x \in \mathbb{R}^d : det(\Sigma)^{1/2} f(x) \geq \alpha\},$$

die in Definition 4.2 geforderten Bedingungen. Diese Bereiche werden Konturen gleicher Dichte genannt. In diesem Fall sind die Bereiche um den Modalwert der Verteilung zentriert. Die Determinante $det(\Sigma)^{1/2}$ sichert die affine Äquivarianz dieses Kontur-Toleranzbereichs.

Bemerkung 4.3 Fraimann, Liu und Meloche (1997) betrachten Dichteschätzer, mit denen es möglich ist, auch Konturen bimodaler Verteilungen zu schätzen.

Datentiefe-Funktionen messen, wie weit ein Punkt vom Zentrum der Verteilung entfernt ist. Basierend auf Kontur-Toleranzintervallen ergibt sich die Datentiefe-Funktion wie folgt:

Definition 4.3 *Sei $D_F(\alpha)$ ein Kontur-Toleranzbereich. Dann heißt die Abbildung: $DT_F : \mathbb{R}^d \to [0, \infty)$ mit*

$$DT_F(x) = \sup \{\alpha : x \in D_F(\alpha)\}$$

Datentiefe-Funktion.

Aus Definition 4.3 lassen sich die folgenden Eigenschaften einer Datentiefe-Funktion ableiten.

Satz 4.1 Sei $DT_F(x)$ eine Datentiefe-Funktion. Dann gilt:

1. Die Werte der Funktion $DT_F(x)$ sind nichtnegativ und beschränkt.

2. Es gilt: $DT_{F^{AX+b}}(Ax+b) = DT_{F^X}(x)$ für jeden Zufallsvektor $X \sim F$ im $\mathbb{R}^d$, jede nichtsinguläre $d \times d$ Matrix A und jeden Vektor $b \in \mathbb{R}^d$.

3. Für jede Verteilung $F \in \mathcal{F}$ mit Zentrum θ gilt:
   $DT_F(\theta) = \sup_{x \in \mathbb{R}^d} DT_F(x) = \varepsilon.$

4. Für $F \in \mathcal{F}$ gilt: $DT_F(x) \to 0$, falls $||x|| \to \infty$, für $F \in \mathcal{F}$.

5. Sei $F \in \mathcal{F}$ mit Zentrum θ. Dann gilt: $DT_F(x) \leq DT_F(\beta x + (1-\beta)\theta)$ für alle $x \in \mathbb{R}^d$ und $\beta \in [0,1]$.

Beweis:

1. Folgt direkt, da $0 \leq \alpha \leq \varepsilon < \infty$.

2. Da für jede nichtsinguläre $d \times d$ Matrix A und jeden Vektor $b \in \mathbb{R}^d$ gilt:

$$\begin{aligned} DT_{F^X}(x) &= \sup\{\alpha : x \in D_{F^X}(\alpha)\} = \sup\{\alpha : Ax + b \in A\, D_{F^X}(\alpha) \oplus b\} \\ &= \sup\{\alpha : Ax + b \in D_{F^{AX+b}}(\alpha)\} = DT_{F^{AX+b}}(Ax+b), \end{aligned}$$

folgt die Behauptung.

3. Folgt aus der Tatsache, dass Kontur-Toleranzbereiche verschachtelt sind und dass $\max_\alpha D_F(\alpha) = \theta$ gilt.

4. Sei $(x_n)_{n \in \mathbb{N}}$, mit $x_n \in \mathbb{R}^d$ eine Folge, so dass $||x_n|| \to \infty$ falls $n \to \infty$. Angenommen $DT_F(x_n)$ strebt nicht gegen Null, dann existiert ein $0 < \alpha \leq \varepsilon$, so dass gilt: $DT_F(x_n) \geq \alpha > 0$ für alle n. Daraus folgt, dass $x_n \in D_F(\alpha)$ für alle x_n. Dies ist ein Widerspruch zur Kompaktheit von $D_F(\alpha)$ für $0 < \alpha \leq \varepsilon$, da die Folge $(x_n)_{n \in \mathbb{N}}$ keinen Grenzwert in $D_F(\alpha)$ hat.

5. Sei $x \in D_F(\alpha')$, wobei $\alpha' = \sup\{\alpha : x \in D_F(\alpha)\}$. Dann folgt, dass $\theta \in D_F(\alpha')$. Da $D_F(\alpha')$ eine konvexe Menge ist, folgt $(\beta x + (1-\beta)\theta) \in D_F(\alpha')$ und damit $DT_F(\beta x + (1-\beta)\theta) \geq DT_F(x)$ für $\beta \in [0,1]$.

q.e.d.

Bemerkung 4.4 Eine Datentiefe-Funktion hat somit die folgenden Eigenschaften:

1. Die Datentiefe eines Punktes hängt nicht von dem zugrunde gelegten Koordinatensystem ab, d.h., sie ist invariant gegenüber affinen Transformationen.

2. Falls die Verteilung, wie angenommen, ein eindeutig definiertes Zentrum θ besitzt, wird der Wert der Datentiefe-Funktion (auch Tiefe genannt) an der Stelle θ maximal.
3. Für Punkte $x \in \mathbf{R}^d$, die sich vom Zentrum der Verteilung F entfernen, nimmt der Wert der Datentiefe-Funktion ab und strebt gegen Null.
Diese Eigenschaften der Datentiefe-Funktion werden schon von Liu (1990) vorgeschlagen. Sie resultieren somit aus einer geeigneten Definition von Kontur-Toleranzbereichen.

Beispiel 4.2: (Fortsetzung)
Die zugehörige Datentiefe-Funktion zu Beispiel 4.2 ergibt sich wie folgt:

$$LT_F(x) = det(\Sigma)^{1/2} f(x).$$

Diese Funktion erfüllt die Eigenschaften von Satz 4.1. Die Dichtefunktion selbst kann somit als eine Datentiefe-Funktion interpretiert werden.

Jede Datentiefe-Funktion ist durch die zugehörigen Kontur-Toleranzbereiche festgelegt. Andererseits können bei gegebener Datentiefe-Funktion die entsprechenden Kontur-Toleranzbereiche durch Betrachtung der Mengen:

$$D_F(\alpha) = \{x \in \mathbf{R}^d : DT_F(x) \geq \alpha\},\ 0 \leq \alpha < \infty,$$

erhalten werden. Dieses Vorgehen wird von Zuo und Serfling (2000) verfolgt. Da jedoch in dieser Arbeit konvexe Körper im Mittelpunkt der Betrachtung stehen, erscheint es zweckmäßiger aus Kontur-Toleranzbereichen auf die Datentiefe-Funktion zu schließen. Schätzungen von Kontur-Toleranzbereichen und Datentiefe-Funktionen können erhalten werden, indem F durch die empirische Verteilung F_n ersetzt wird.

Mit $EC_d(\mu, \Sigma)$ werde die Klasse der elliptisch symmetrischen Verteilungen im $\mathbf{R}^d$ bezeichnet.
Die multivariate Normalverteilung gehört zur Klasse der elliptisch symmetrischen Verteilungen.

Falls $F \in EC_d(\mu, \Sigma)$, so gilt für alle Kontur-Toleranzbereiche und die zugehörige Datentiefe-Funktion der folgende Zusammenhang.

Lemma 4.1 Sei der $\mathbf{R}^d$-Zufallsvektor $X \sim F$, wobei $F \in EC_d(\mu, \Sigma)$. Dann sind die Kontur-Toleranzbereiche (Hyper-)Ellipsen, d.h. sie sind in folgender Form darstellbar:

$$D_F(\alpha) = \{x \in \mathbf{R}^d : (x - \mu)^T \Sigma^{-1} (x - \mu) \leq r_\alpha^2\},$$

wobei $r_\alpha > 0$.

Beweis: Zuo und Serfling (2000) q.e.d.

Bemerkung 4.5 Kontur-Toleranzbereiche elliptisch symmetrischer Verteilungen entsprechen Ellipsen, die um den Erwartungswert der zugrunde liegenden Verteilungen zentriert sind.

4.2 Beispiele für Kontur-Toleranzbereiche

Bei Vorgabe unterschiedlicher Kriterien (nach Definition 4.1) resultieren wie zuvor gesehen unterschiedliche Zentren. Unter Zuhilfenahme der Datentiefe-Funktion können um diese Zentren herum Kontur-Toleranzbereiche konstruiert werden. Im Folgenden wird anhand von zwei Beispielen das Prinzip erläutert und die entsprechenden Schätzungen für Kontur-Toleranzbereiche angegeben und auf ihr Bruchpunktverhalten untersucht.

4.2.1 Mahalanobis-Konturen

Mit $\mathcal{F}_1$ werde die Menge aller Verteilungen bezeichnet, deren zweites Moment existiert. Zur Konstruktion von so genannten Mahalanobis-Konturen, wird das Kriterium (K2) zugrunde gelegt. Das Zentrum einer Verteilung ist der Erwartungswert μ. Um dieses konstruiert Liu (1992) Kontur-Toleranzbereiche, die auf dem Mahalanobis-Abstand zwischen $x \in \mathbf{R}^d$ und $\mu \in \mathbf{R}^d$ basieren
Basierend auf diesem Abstand bestimmt Liu (1992) den folgenden Bereich:

Definition 4.4 *Sei $F \in \mathcal{F}_1$. Sei $\mu \in \mathbb{R}^d$ der Erwartungswert und $\Sigma \in \mathbb{R}^{d \times d}$ die Kovarianz der Verteilung F und $\alpha \in [0,1]$. Dann heißen die Bereiche:*

$$MT_F(\alpha) := \left\{ x \in \mathbb{R}^d : (x-\mu)^T \Sigma^{-1} (x-\mu) \leq \frac{1-\alpha}{\alpha} \right\},$$

Mahalanobis-Konturen.

Es gilt:

Satz 4.2 Die Mahalanobis-Konturen $MT_F(\alpha)$ erfüllen die in Definition 4.2 geforderten Eigenschaften und sind somit Kontur-Toleranzbereiche bzgl. des Kriteriums (K2).

Beweis:
Um zu zeigen, dass $MT_F(\alpha)$ ein Kontur-Toleranzbereich ist, sind fünf Voraussetzungen zu überprüfen:

1. Da die Bereiche für $\alpha \in (0,1]$ Ellipsen entsprechen, gilt: $MT_F(\alpha) \in \mathcal{K}^d$.

2. Affine Äquivarianz ist gegeben, da für jede nichtsinguläre $d \times d$ Matrix A und jedes $b \in \mathbb{R}^d$ gilt:

$$\begin{aligned}
AMT_{F^X}(\alpha) \oplus b &= A\left\{ x \in \mathbb{R}^d : (x-\mu)^T \Sigma^{-1} (x-\mu) \leq \frac{1-\alpha}{\alpha} \right\} \oplus b \\
&= \left\{ Ax + b \in \mathbb{R}^d : (x-\mu)^T \Sigma^{-1} (x-\mu) \leq \frac{1-\alpha}{\alpha} \right\} \\
&= \left\{ x \in \mathbb{R}^d : (A^{-1}x - A^{-1}b - \mu)^T \Sigma^{-1} (A^{-1}x - A^{-1}b - \mu)) \leq \frac{1-\alpha}{\alpha} \right\} \\
&= \left\{ x \in \mathbb{R}^d : (A^{-1}(x - (A\mu + b)))^T \Sigma^{-1} (A^{-1}(x - (A\mu + b)) \leq \frac{1-\alpha}{\alpha} \right\} \\
&= \left\{ x \in \mathbb{R}^d : (x - (A\mu + b))^T \left(A \Sigma A^T \right)^{-1} (x - (A\mu + b)) \leq \frac{1-\alpha}{\alpha} \right\} \\
&= MT_{F^{AX+b}}(\alpha).
\end{aligned}$$

3. Die Bereiche sind verschachtelt, da für $\alpha_1 \geq \alpha_2$ gilt:
Ist $x \in MT_F(\alpha_1)$ folgt dass:

$$x \in MT_F(\alpha_1) \Rightarrow (x-\mu)^T \Sigma^{-1} (x-\mu) \geq \frac{1-\alpha_2}{\alpha_2} \geq \frac{1-\alpha_1}{\alpha_1} \Rightarrow x \in MT_F(\alpha_2).$$

Hieraus folgt: $MT_F(\alpha_1) \subseteq MT_F(\alpha_2)$.

4. Für $\alpha = 1$ folgt sofort $MT_F(1) = \mu$. Und da μ dem Zentrum von F entspricht, folgt

die Behauptung.

5. Für $\alpha \to 0$ erfüllt jeder Punkt $x \in \mathbf{R}^d$ die geforderte Bedingung. Die Menge all dieser Punkte entspricht der konvexen Hülle des Stichprobenraums der Verteilung F.

q.e.d.

Bemerkung 4.6 1. Die zugehörige Datentiefe Funktion der Mahalanobis-Konturen ist gegeben durch:

$$MDT_F(x) = \frac{1}{1 + (x-\mu)^T \Sigma^{-1} (x-\mu)}.$$

2. Die vorgestellten Mahalanobis-Konturen und die zugehörige Datentiefe-Funktion sind nur definiert, falls das erste und zweite Moment der Verteilung F existiert. Da die Bereiche Ellipsen entsprechen, wird die wahre Struktur und Variabilität der zugrunde liegenden Verteilung F insbesondere erfasst, falls F zur Klasse der elliptisch symmetrischen Verteilungen $EC_d(\mu, \Sigma)$ gehört.

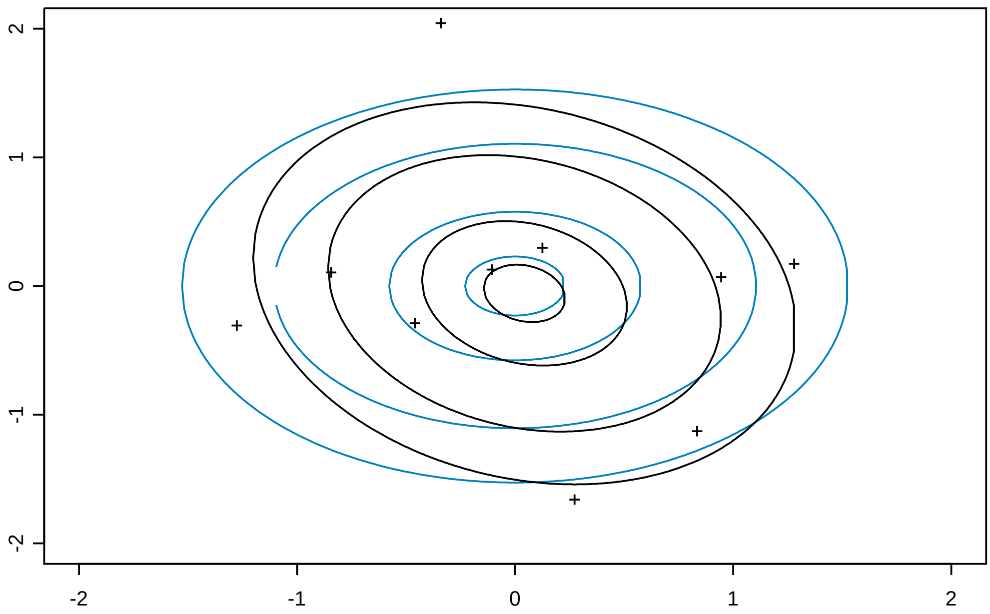
3. Wird F durch die empirische Verteilung F_n ersetzt, so folgt aus dem Beweis von Satz 4.2, dass $MT_{F_n}(\alpha)$ genau dann ein Kontur-Toleranzbereich ist, falls μ_n und C_n affin äquivariante Schätzer für μ und Σ sind. Ein Kovarianzschätzer wird affin äquivariant genannt, wenn gilt: $C(A\tilde{X}_n + b) = AC(\tilde{X}_n)A^T$.

Beispiel 4.3 In Abbildung 4.1 sind Mahalanobis-Konturen (gestrichelt) dargestellt für den Fall, dass $F = N((0,0)^T, I_2)$ und $\alpha \in \{0.1, 0.2, 0.3, 0.4\}$. Zur Schätzung wurden 10 Realisationen aus $N((0,0)^T, I_2)$ benutzt. In schwarzer Farbe sind die entsprechenden Schätzungen abgetragen, wobei μ mit Hilfe des arithmetischen Mittelwertvektors und Σ durch die empirische Kovarianzmatrix geschätzt werden. Da die multivariate Normalverteilung zur Klasse der elliptisch symmetrischen Verteilungen gehört, spiegelt Abbildung 4.1 die Form und Struktur der Verteilung wider.

Robustheitsuntersuchungen für Schätzer von Mahalanobis-Konturen

Aus den Überlegungen im zweiten Kapitel folgt, dass der Wert des Bruchpunktes von $MT_F(\alpha)$ nur von den affin äquivarianten Lokations- und Kovarianzschätzern abhängt.

Abbildung 4.1: Mahalanobis-Konturen einer bivariaten Normalverteilung (gestrichelt) und die zugehörigen Schätzungen (schwarz).



Da der Bruchpunkt der in Beispiel 4.3 verwendeten Lokations- und Kovarianzschätzer (nach Definiton 2.3 und 2.4) gerade $1/n$ ist, ergibt sich derselbe Bruchpunkt nach Definition 2.10 auch für die Schätzung der Mahalanobis-Konturen (vgl. Beispiel 2.4). Um also einen Schätzer der Mahalanobis-Konturen mit hohem Bruchpunkt zu garantieren, sollten robuste Lokations- bzw. Kovarianzschätzer verwendet werden. Hierauf wird im fünften Kapitel ausführlich eingegangen.

4.2.2 Zonoide Zonen

Koshevoy und Mosler (1997) schlagen ein weiteres Verfahren vor, um multivariate Kontur-Toleranzbereiche zu bestimmen, die so genannten (α-getrimmten) zonoiden Zonen. Diese Bereiche sind genau wie die Mahalanobis-Konturen um den Erwartungswert zentriert, d.h., es wird wieder (K2) als Kriterium benutzt. Im Gegensatz dazu können (α-getrimmte) zonoide Zonen auch eingesetzt werden, wenn das zweite Moment der interessierenden Verteilung nicht existiert.

Definition 4.5 *Sei $\alpha \in (0,1]$ und $F \in \mathcal{F}_0$. Dabei sei $\mathcal{F}_0$ die Menge aller Verteilungen, deren erstes absolutes Moment existiert. Dann heißt:*

$$Z_F(\alpha) = \left\{ \int_{\mathbb{R}^d} x\,\tilde{g}(x)dF(x) : \tilde{g} : \mathbb{R}^d \to \left[0, \frac{1}{\alpha}\right] \textit{ messbar und } \int_{\mathbb{R}^d} \tilde{g}(x)dF(x) = 1 \right\}$$

(α-getrimmte) zonoide Zone von F.

Bemerkung 4.7 Zonoide Zonen resultieren aus der Projektion von Liftzonoiden in den $\mathbb{R}^d$. Dazu sei $\alpha \in (0,1]$, $F \in \mathcal{F}_0$ und $L(F,\alpha)$ die Schnittmenge zwischen der Hyperebene $E_\alpha = \{(x_0, x_1, ..., x_d)^T \in \mathbb{R}^{d+1} : x_0 = \alpha\}$ und dem Liftzonoid. Die α-getrimmte zonoide Zone $Z_F(\alpha)$ ist gegeben als Projektion von $L(F,\alpha)$ in die letzten d Koordinaten multipliziert mit $1/\alpha$, d.h:

$$Z_F(\alpha) = \frac{1}{\alpha}\, proj_d(L(F,\alpha)).$$

Beispiel 4.4 Für den Fall $d = 1$ ergeben sich die (α-getrimmten) zonoiden Zonen wie folgt:

Nach Punkt 2 von Bemerkung 3.7 ist $LZ(F)$ gegeben durch die konvexe Hülle der Punkte:

$$\left(t, \int_0^t F^{-1}(s)ds\right)^T, \quad \text{und } \left(t, \int_{1-t}^1 F^{-1}(s)ds\right)^T, \quad t \in [0,1].$$

Sei nun $E_\alpha = \{(x_0, x_1)^T \in \mathbb{R}^2 : x_0 = \alpha\}$.

Dann ist:

$$L(F,\alpha) = \left(E_\alpha \bigcap LZ(F)\right) = \left[\left(\alpha, \int_0^\alpha F^{-1}(s)ds\right)^T, \left(\alpha, \int_{1-\alpha}^1 F^{-1}(s)ds\right)^T\right]$$

eine Strecke im $\mathbf{R}^2$ und somit folgt:

$$Z_F(\alpha) = \frac{1}{\alpha} proj_d\Big(L(F,\alpha)\Big) = \left[\frac{1}{\alpha}\int_0^\alpha F^{-1}(s)ds, \frac{1}{\alpha}\int_{1-\alpha}^1 F^{-1}(s)ds\right].$$

Dabei ist $F^{-1}(s)$ das s-Quantil der Verteilung F.

Lemma 4.2 Die folgende Darstellung der Abbildung $Z_F : (0,1] \to \mathcal{K}^d$ ist äquivalent zu Definition 4.5:

$$Z_F(\alpha) = conv\left\{\frac{1}{\alpha}\int_H xdF(x), H \in \mathcal{H}^d, F(H) = \alpha\right\},$$

wobei $\mathcal{H}^d$ die Menge aller geschlossenen Halbräume bezeichne.
Für $\alpha \to 0$ gilt:

$$Z_F(\alpha) \to \bigcap\Big\{H \in \mathcal{H}^d : F(H) = 1\Big\}.$$

Beweis: (Koshevoy und Mosler (1997))

q.e.d.

Lemma 4.3 Sei F_n die empirische Verteilungsfunktion der u.i.v. $\mathbf{R}^d$-ZV $X_1, ..., X_n$ mit $X_i \sim F \in \mathcal{F}_0$. Dann gilt:

$$\lim_{n\to\infty} Z_{F_n}(\alpha) = Z_F(\alpha) \text{ fast sicher bzgl. des Hausdorff Abstandes für } 0 < \alpha \leq 1.$$

Beweis: (Koshevoy und Mosler (1997))

q.e.d.

Schätzung von zonoiden Zonen

Sei X ein $\mathbf{R}^d$-ZV mit $X \sim F \in \mathcal{F}_0$. Weiter sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine Stichprobe von Realisierungen aus $F \in \mathcal{F}_0$ vom Umfang n. Es sei F_n die empirische Verteilung von $\tilde{X}_n$. Die Schätzung von (α-getrimmten) zonoiden Zonen erfolgt durch Bestimmung von $Z_{\tilde{X}_n}(\alpha) := Z_{F_n}(\alpha) = \frac{1}{\alpha} proj_d\Big(E_\alpha \bigcap LZ(F_n)\Big) = \frac{1}{\alpha} proj_d\Big(L(F_n,\alpha)\Big)$, wobei $E_\alpha = \{x = (x_0, x_1, ..., x_d)^T \in \mathbf{R}^d : x_0 = \alpha\}$ eine Hyperebene im $\mathbf{R}^d$ bezeichnet.

Satz 4.3 Sei X ein $\mathbb{R}^d$-ZV mit $X \sim F \in \mathcal{F}_0$ und $\tilde{X}_n = \{x_1, ..., x_n\}$ eine Stichprobe vom Umfang n aus F. Dann kann $Z_{\tilde{X}_n}(\alpha) = \frac{1}{\alpha} proj_d\Big(L(F_n, \alpha)\Big)$ dargestellt werden durch:

$$Z_{\tilde{X}_n}(\alpha) = conv\left\{\frac{1}{\alpha n}\sum_{j=1}^{k} x_{i_j} + (1 - \frac{k}{\alpha n} x_{i_{k+1}}) : \{i_1, ..., i_{k+1}\} \subset \{1, ..., n\}\right\},$$

für $\alpha \in (\frac{k}{n}, \frac{(k+1)}{n}]$, $k = 1, ..., n-1$, und

$$Z_{\tilde{X}_n}(\alpha) = conv\{x_1, ..., x_n\}, \text{ für } \alpha \in [0, \tfrac{1}{n}].$$

Beweis: (Koshevoy und Mosler (1997))

q.e.d.

Bemerkung 4.8 1. Zur Schätzung von $Z_F(\alpha)$ für $\alpha \in (\frac{k}{n}, \frac{(k+1)}{n}]$, $k = 1, ..., n-1$, werden zunächst aus allen möglichen $(k+1)$-elementigen Teilstichproben der Beobachtungen $\{x_1, ..., x_n\}$ die Mittelwerte berechnet. Eine Schätzung von $Z_F(\alpha)$ ist nun durch die konvexe Hülle aller derartigen Mittelwertvektoren gegeben. Somit müssen zur Schätzung $\binom{n}{(k+1)}$ Teilstichproben bestimmt werden. Für den Fall $\alpha = 1$ erhält man das arithmetische Mittel $\bar{x}$ der Stichprobe $\tilde{X}_n$.

2. Die Schätzung von $Z_F(\alpha)$ ist somit ein Polytop.

3. Bei dieser Art der Schätzung handelt es sich um ein verteilungsfreies Verfahren bzgl. der Verteilungsklasse $\mathcal{F}_0$.

Aus Satz 4.3 folgt die Aussage:

Korollar 4.1 Gegeben sei eine Stichprobe $\tilde{X}_n = \{x_1, ..., x_n\}$ vom Umfang n, wobei $x_1, ..., x_n$ Realisierungen des $\mathbb{R}^d$-ZV $X \sim F \in \mathcal{F}_0$. Dann gilt für $0 \leq \alpha_1 \leq \alpha_2 \leq 1$:

$$Z_{\tilde{X}_n}(\alpha_2) \subseteq Z_{\tilde{X}_n}(\alpha_1).$$

Beweis: Koshevoy und Mosler (1997)

q.e.d.

Korollar 4.2 Die Realisierungen $x_1, ..., x_n$, einer univariaten Zufallsvariable $X \sim F$ seien der Größe nach angeordnet: $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$. Dann ist für ein festes $\alpha \in (0,1]$ die Schätzung der α-getrimmten zonoiden Zonen gegeben durch:

$$Z_{\tilde{X}_n}(\alpha) = \left[\frac{1}{n\alpha} \sum_{i=1}^{\lfloor n\alpha \rfloor} x_{(i)}, \frac{1}{n\alpha} \sum_{i=n-\lfloor n\alpha \rfloor+1}^{n} x_{(i)} \right].$$

Beweis: folgt direkt aus Satz 4.3.

q.e.d.

Im Folgenden wird gezeigt, dass zonoide Zonen Konturtoleranz-Bereiche nach Definition 4.2 sind.

Satz 4.4 Die α-getrimmten zonoiden Zonen $Z_F(\alpha)$ erfüllen die in Definition 4.2 geforderten Eigenschaften und sind somit Kontur-Toleranzbereiche bzgl. des Kriteriums (K2).

Beweis:
Um zu zeigen, dass $Z_\alpha(F)$ ein Kontur-Toleranzbereich ist, sind fünf Punkte zu überprüfen:
Punkt 1 und Punkt 5 folgen aus Lemma 4.3.
2. Seien $x_1, ..., x_n$ Realisationen des ZV X. Dann sind $Ax_i + b, ..., Ax_n + b$ zugehörige Realisationen des ZV $AX + b$. Die empirische Verteilung von $Ax_i + b, ..., Ax_n + b$ werde mit $\tilde{F}_n$ bezeichnet. Nun folgt aus Satz 4.3 direkt: $Z_{\tilde{F}_n}(\alpha) = AZ_{F_n}(\alpha) \oplus b$. Nach Lemma 4.3 folgt für $\alpha \in (0,1]$:

$$Z_{F_{AX+b}}(\alpha) = \lim_{n\to\infty} Z_{\tilde{F}_n}(\alpha) = \lim_{n\to\infty} AZ_{F_n}(\alpha) \oplus b = AZ_F(\alpha) \oplus b.$$

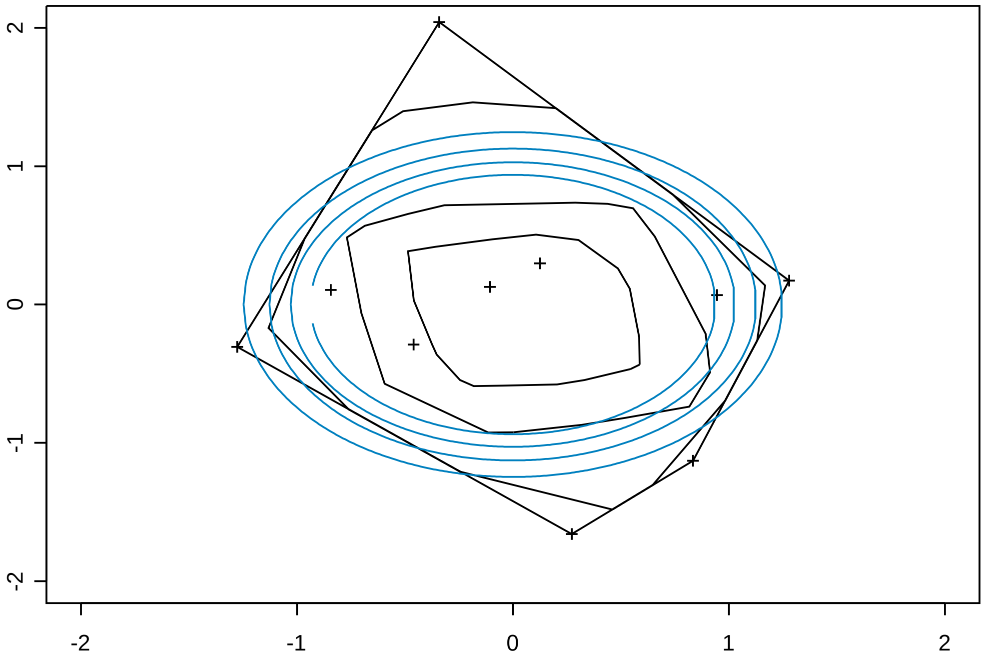
3. Für $\alpha = 1$ gilt $Z_F(1) = \mu$.
4. Sei $\alpha_1 \leq \alpha_2$, dann folgt aus Korollar 4.1 und Lemma 4.3.

$$Z_F(\alpha_2) = \lim_{n\to\infty} Z_{F_n}(\alpha_2) \subseteq \lim_{n\to\infty} Z_{F_n}(\alpha_1) = Z_F(\alpha_1).$$

Somit folgt die Behauptung.

q.e.d.

Abbildung 4.2: Getrimmte zonoide Zonen einer bivariaten Standardnormalverteilung (blau) und die zugehörigen Schätzungen (schwarz).



Beispiel 4.5 Für eine multivariate Normalverteilung mit Erwartungswert $\mu \in \mathbf{R}^d$ und einer $d \times d$-Kovarianzmatrix Σ sind die (α-getrimmten) zonoiden Zonen gegeben durch:

$$Z_\alpha(N(\mu, \Sigma)) = \{x \in \mathbf{R}^d : (x-\mu)^T \Sigma^{-1}(x-\mu) \leq r_\alpha^2\},$$

wobei $r_\alpha = \frac{exp(-q^2_{(1-\alpha)}/2)}{\sqrt{2\pi}\alpha}$ und q_s das s-Quantil der Standardnormalverteilung bezeichne. Sie entsprechen somit Ellipsen (vgl. Lemma 4.1). In Abbildung 4.2 sind α-getrimmte zonoide Zonen und deren Schätzung für den Fall einer $F = N((0,0)^T, I_2)-$Verteilung für $\alpha = \{0.1, 0.2, 0.4, 0.5\}$ abgebildet.

Robustheitsuntersuchungen für Schätzer von zonoiden Zonen

Die in Satz 4.3 vorgestellte Schätzung der α-getrimmten zonoiden Zonen legt die Vermutung nahe, dass für jedes $\alpha \in [0,1]$ schon eine Beobachtung ausreichen wird, um den Schätzer zusammenbrechen zu lassen. Im Folgenden wird diese Vermutung bestätigt.

Lemma 4.4 Gegeben sei eine Stichprobe $\tilde{X}_n$ vom Umfang n mit $x_i \in \mathbf{R}^d$, $i = 1, ..., n$. Dann gilt für den finite-sample Explosions-Bruchpunkt nach Definition 2.6 von $Z_{\tilde{X}_n}(\alpha)$ für alle $\alpha \in [0,1]$:

$$\epsilon_K(\tilde{X}_n, Z_{\cdot}(\alpha)) = \frac{1}{n}.$$

Beweis:
Sei $\alpha \in \left(\frac{k}{n}, \frac{(k+1)}{n}\right]$, $k = 1, ..., n-1$, und $\tilde{X}_n = \{x_1, ..., x_n\}$, $x_i \in \mathbf{R}^d$, $i = 1, ..., n-1$, eine Stichprobe vom Umfang n. Nach Satz 4.3 gilt:

$$Z_{\tilde{X}_n}(\alpha) = conv\left\{\left\{\frac{1}{\alpha n}\sum_{j=1}^{k} x_{i_j} + (1 - \frac{k}{\alpha n} x_{i_{k+1}})\right\} : \mathcal{I} := \{i_1, ..., i_{k+1}\} \subset \{1, ..., n\}\right\},$$

und nach Formel (2.6) folgt:

$$h\left(Z_{\tilde{X}_n}(\alpha), p\right) = \max_{\mathcal{I}} \left\langle \left(\frac{1}{\alpha n}\sum_{j=1}^{k} x_{i_j} + \left(1 - \frac{k}{\alpha n} x_{i_{k+1}}\right)\right), p\right\rangle.$$

Sei $\tilde{Y}_{n,1} = \{x_{i_1}, ..., x_{i_{n-1}}, y\}$, eine durch Austausch der Beobachtung x_{i_n} aus $\tilde{X}_n$ entstandene Stichprobe. Die ausgetauschte Beobachtung $y \neq \mathbf{0} \in \mathbf{R}^d$ und $\tilde{p} = \frac{y}{||y||} \in \mathbf{R}^d$ mit $||\tilde{p}|| = 1$ werden so gewählt, dass für alle $x_i \in \tilde{X}_n$ gilt:

$$\langle x_i, \tilde{p}\rangle \leq \langle y, \tilde{p}\rangle, \ i = 1, ..., n.$$

Der Punkt y liegt außerhalb der konvexen Hülle von $\tilde{X}_n$.
Sei α beliebig aber fest. Dann folgt, dass eine Teilstichprobe $\{\tilde{x}_{i_1}, ..., \tilde{x}_{i_{k+1}}\}$ vom Umfang $k+1$ existiert, so dass gilt:

$$\left\langle \frac{1}{\alpha n}\sum_{j=1}^{k} \tilde{x}_{i_j} + \left(1 - \frac{k}{\alpha n}\right)\tilde{x}_{i_{k+1}}, \tilde{p}\right\rangle = h\left(Z_{\tilde{X}_n}(\alpha), \tilde{p}\right),$$

wobei $\langle \tilde{x}_{i_1}, \tilde{p}\rangle \geq \langle \tilde{x}_{i_2}, \tilde{p}\rangle \geq ... \geq \langle \tilde{x}_{i_k}, \tilde{p}\rangle \geq \langle \tilde{x}_{i_{k+1}}, \tilde{p}\rangle$. Aus diesen Überlegungen folgt:

$$h\left(Z_{\tilde{X}_n}(\alpha), \tilde{p})\right) \leq \left\langle \frac{1}{\alpha n} \sum_{j=1}^{k-1} \tilde{x}_{i_j} + \frac{1}{\alpha n} y + \left(1 - \frac{k}{\alpha n}\right) \tilde{x}_{i_k}, \tilde{p} \right\rangle = h\left(Z_{\tilde{Y}_{n,1}}(\alpha), \tilde{p}\right),$$

da $\frac{1}{\alpha n} \geq (1 - \frac{k}{\alpha n})$.

Hieraus folgt für $\tilde{p} = \frac{y}{||y||}$:

$$\begin{aligned} d_H\left(Z_{\tilde{X}_n}(\alpha), Z_{\tilde{Y}_{n,1}}(\alpha)\right) &= \max_{||p||=1} \left| h\left(Z_{\tilde{X}_n)}(\alpha), p\right) - h\left(Z_{\tilde{Y}_{n,1}}(\alpha), p\right) \right| \\ &\geq \left| \left\langle \left(1 - \frac{k}{\alpha n}\right) \tilde{x}_{i_{k+1}} + \left(1 - \frac{(k+1)}{\alpha n} \tilde{x}_{i_k}\right) - \frac{1}{\alpha n} y, \tilde{p} \right\rangle \right| \\ &= \left| \left\langle \left(1 - \frac{k}{\alpha n}\right) \tilde{x}_{i_{k+1}}, \frac{y}{||y||} \right\rangle + \left\langle \left(1 - \frac{(k+1)}{\alpha n} \tilde{x}_{i_k}\right), \frac{y}{||y||} \right\rangle - \frac{1}{\alpha n} ||y|| \right|. \end{aligned}$$

Da für $||y|| \to \infty$ folgt, dass $d_H\left(Z_{\tilde{X}_n}(\alpha), Z_{\tilde{Y}_{n,1}}(\alpha)\right) \to \infty$, folgt die Behauptung.

Für $\alpha \in [0, \frac{1}{n}]$ erfolgt der Beweis analog.

Insgesamt ist somit der Bruchpunkt für α-getrimmte zonoide Zonen $\frac{1}{n}$.

q.e.d.

Im dritten und vierten Kapitel wurden Schätzer konvexer Körper auf ihr Bruchpunktverhalten untersucht. Es hat sich gezeigt, dass schon wenige Beobachtungen ausreichen, um die Schätzung zusammenbrechen zu lassen. Im nächsten Kapitel werden Schätzungen, die einen hohen Bruchpunkt aufweisen, für die hier vorgestellten konvexen Körper konstruiert.

Kapitel 5

Robuste Schätzer konvexer Körper

Im dritten und vierten Kapitel wurden konvexe Körper und erste Schätzungen vorgestellt. Diese Schätzungen basieren auf der konvexen Hülle einer endlichen Punktmenge (Zonoide) oder auch auf Lokations- und Kovarianzschätzern (Mahalanobis-Konturen). Es zeigt sich jedoch, dass diese Schätzungen sensibel auf das Vorhandensein von Ausreißern reagieren.
Ziel dieses Kapitels ist es, Schätzer für konvexe Körper zu finden, die einen hohen Bruchpunkt besitzen und damit weniger empfindlich gegenüber Ausreißern sind.
Dazu wird ein von Rousseeuw (1985) entwickeltes Prinzip auf die hier betrachtete Situation übertragen. Dieses beruht darauf, Schätzer basierend auf derjenigen Teilstichprobe zu bestimmen, die ein vorher festgesetztes Variabilitätsmaß minimiert. Auch zur Schätzung konvexer Körper kann eine solche bereinigte Stichprobe verwendet werden. Mit diesem Konstruktionsprinzip ist es möglich, affin äquivariante Schätzer mit hohem Bruchpunkt herzuleiten. Im Abschnitt 5.1 werden zunächst von Rousseeuw (1985) vorgeschlagene affin äquivariante Lokations- bzw. Kovarianzschätzer, die einen hohen Bruchpunkt aufweisen, beschrieben. Im Anschluss wird ein weiterer Lokations- bzw. Kovarianzschätzer vorgeschlagen, dem ein ähnliches Konstruktionsprinzip zugrunde liegt. All diese Schätzer weisen den höchst möglichen Bruchpunkt für affin äquivariante Lokations- bzw. Kovarianzschätzer auf. Zum Abschluss der Arbeit werden Schätzer konvexer Körper mit hohem Bruchpunkt vorgestellt.

5.1 Lokations- und Kovarianzschätzer mit hohem Bruchpunkt

Eine wichtige Aufgabe in der Statistik ist die Schätzung der Parameter einer Verteilung aufgrund von Stichproben, die gemäß dieser Verteilung erzeugt worden sind. Zu den in der Literatur zu findenden Lokations- bzw. Kovarianzschätzern, die einerseits die Eigenschaft der affinen Äquivarianz besitzen und andererseits einen hohen Bruchpunkt aufweisen, gehören die auf den MVE- bzw. MCD-Kriterien basierenden Lokations- bzw. Kovarianzschätzer.

5.1.1 MVE- und MCD-Schätzer

Den auf den MVE- bzw. MCD-Kriterien basierenden Schätzern liegt die Idee zugrunde, nicht alle Beobachtungen einer Stichprobe zu benutzen, sondern eine Teilstichprobe auszuwählen, die ein vorher festgelegtes Variabilitätsmaß minimiert. Mit der so ermittelten Teilstichprobe können die interessierenden Parameter der Verteilung geschätzt werden. Hierdurch wird sicher gestellt, dass Beobachtungen, die weit entfernt von der Hauptmasse der Daten liegen, nicht mit in die Schätzung einfließen.

Definition 5.1 *Sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine Stichprobe $x_i \in \mathbb{R}^d$, $i = 1, ..., n$, aus einer Normalverteilung mit Parametern $\mu \in \mathbb{R}^d$ und $\Sigma \in \mathbb{R}^{d \times d}$ vom Umfang $n \geq d + 1$ in allgemeiner Lage. Sei $g = \lfloor (n + d + 1)/2 \rfloor$. Es werden alle $\binom{n}{g}$ $g-$elementigen Teilstichproben gebildet. Dann bezeichne:*

1. *$\tilde{X}_{g,E}$ diejenige g-elementige Teilstichprobe, die in derjenigen Ellipse enthalten ist, deren Volumen unter allen Ellipsen minimal ist.*
2. *$\tilde{X}_{g,C}$ diejenige g-elementige Teilstichprobe, deren empirische Kovarianzmatrix minimale Determinante besitzt.*

Bemerkung 5.1 Falls $d = 1$, so entspricht $\tilde{X}_{g,E}$ derjenigen Teilstichprobe unter allen g-elementigen Teilstichproben, deren Spannweite $r_g := |\max\{\tilde{X}_g\} - \min\{\tilde{X}_g\}|$ minimal

ist.
Die Teilstichprobe $\tilde{X}_{g,C}$ ist im univariaten Fall diejenige Teilstichprobe vom Umfang g, deren empirische Varianz $s_g^2 = \frac{1}{g-1}\sum_{x_i \in \tilde{X}_g}(x_i - \bar{x}_g)^2$, mit $\bar{x}_g = \frac{1}{g}\sum_{x_i \in \tilde{X}_g} x_i$, minimal wird.

Für die Teilstichproben $\tilde{X}_{g,E}$ und $\tilde{X}_{g,C}$ aus Definition 5.1 gilt:

Lemma 5.1 Sei $\tilde{X}_n = \{x_1, ..., x_n\}$, $x_i \in \mathbf{R}^d$ eine Stichprobe in allgemeiner Lage vom Umfang $n \geq d+1$, weiter sei A eine nichtsinguläre $d \times d$ Matrix, $b \in \mathbf{R}^d$ und $\tilde{X}_{g,E}$ und $\tilde{X}_{g,C}$ seien wie in Definition 5.1 definiert. Wird die Stichprobe $A\tilde{X}_n + b$ betrachtet, so gilt:

$$\left(A\tilde{X}_n + b\right)_{g,E} = A\tilde{X}_{g,E} + b \qquad \text{und} \qquad \left(A\tilde{X}_n + b\right)_{g,C} = A\tilde{X}_{g,C} + b.$$

Beweis:
Für eine beliebige Ellipse E gilt: $vol(AE) = |det(A)|\, vol(E)$, wobei $vol(K)$ das Volumen des konvexen Körpers K bezeichne. Der Term $|det(A)|$ ist konstant und somit für alle Teilstichproben gleich. Da auch eine Verschiebung des Mittelpunktes um $b \in \mathbf{R}^d$ keine Auswirkung auf das Volumen einer Ellipse hat, folgt die erste Behauptung.
Für die Determinante der empirischen Kovarianzmatrix der transformierten Stichprobe $A\tilde{X}_n + b$ gilt: $det(C(A\tilde{X}_n + b)) = det(AC(\tilde{X}_n)A^T) = (det(A))^2\, det(C(\tilde{X}_n))$. Da $(det(A))^2$ konstant ist, folgt die zweite Behauptung.

q.e.d.

Bemerkung 5.2 Werden affin äquivariante Lokations- oder Kovarianzschätzer auf diese Teilstichproben angewendet, so ist der resultierende Schätzer wiederum affin äquivariant.

Die von Rousseeuw (1985) vorgeschlagenen Schätzer ergeben sich dann wie folgt.

Definition 5.2 *Seien $\tilde{X}_{g,E}$ und $\tilde{X}_{g,C}$ wie in Definition 5.1.*
Dann ist der Minimum Volumen Lokationsschätzer (MVE-Lokationsschätzer) $T_{g,E}(\tilde{X}_n)$

definiert als das Zentrum derjenigen Ellipse, die $\tilde{X}_{g,E}$ enthält und daher minimales Volumen hat.
Der MVE-Kovarianzschätzer $C_{g,E}(\tilde{X}_n)$ entspricht der empirischen Kovarianz der Teilstichprobe $\tilde{X}_{g,E}$.
Der Minimum Covariance Determinant (MCD) Lokationsschätzer $T_{g,C}(\tilde{X}_n)$ ist als das arithmetische Mittel der Teilstichprobe $\tilde{X}_{g,C}$ definiert.
Der MCD-Kovarianzschätzer $C_{g,C}(\tilde{X}_n)$ ergibt sich aus der empirischen Kovarianz der Teilstichprobe $\tilde{X}_{g,C}$.

Bemerkung 5.3 Problematisch an diesem Ansatz ist, dass die Schätzungen der Kovarianzen schlechte Ergebnisse liefern, wenn weniger als g der Beobachtungen Ausreißer sind. Als Korrekturfaktor schlagen Rousseeuw und Leroy (1987) daher vor, den MVE-Kovarianzschätzer mit dem Faktor $\left(\chi^2_{d,0.5}\right)^{-1} m$ zu multiplizieren. Dabei ist $\left(\chi^2_{d,0.5}\right)^{-1}$ der Median der χ^2-Verteilung mit d Freiheitsgraden und m der Median der Mahalanobis-Abstände der Stichprobe bezüglich der geschätzten Parameter. Einen entsprechenden Korrekturfaktor für den MCD-Kovarianzschätzer geben Pison, Van Aelst und Willems (2002) an.
Die affine Äquivarianz des arithmetischen Mittels und der empirischen Kovarianz implizieren die affine Äquivarianz der in Definition 5.2 eingeführten Schätzer.

Lemma 5.2 Sei $\tilde{X}_n$ eine Stichprobe vom Umfang $n \geq d+1$ in allgemeiner Lage mit $x_i \in \mathbf{R}^d$, $i = 1, ..., n$. Gegeben seien die Schätzer aus Definition 5.2. Für $d = 1$ sind die finite-sample Bruchpunkte dieser Schätzer nach Definition 2.3 und 2.4 gegeben durch: $\epsilon_L(\tilde{X}_n, T_{g,C}) = \epsilon_L(\tilde{X}_n, T_{g,E}) = \left\lfloor \frac{n+1}{2} \right\rfloor \Big/ n$ und $\epsilon_C(\tilde{X}_n, C_{g,C}) = \epsilon_C(\tilde{X}_n, C_{g,E}) = \left\lfloor \frac{n}{2} \right\rfloor \Big/ n.$
Falls $d \geq 2$ dann gilt für die entsprechenden finite-sample Bruchpunkte der Schätzer aus Definition 5.2:

$$\epsilon_C(\tilde{X}_n, C_{g,C}) = \epsilon_C(\tilde{X}_n, C_{g,E}) = \epsilon_L(\tilde{X}_n, T_{g,C}) = \epsilon_L(\tilde{X}_n, T_{g,E}) = \left\lfloor \frac{n-d+1}{2} \right\rfloor \Big/ n.$$

Beweis: siehe Rousseeuw und Leroy (1987) und Lopuhaä und Rousseuuw (1991)

q.e.d.

Bemerkung 5.4 Eine exakte Bestimmung der MVE- und MCD-Schätzer ist sehr rechenintensiv. Es wurden jedoch approximative Algorithmen entwickelt. Rousseeuw und van Zomeren (1990) schlagen einen approximativen Algorithmus vor, um die auf dem MVE-Kriterium beruhenden Parameter zu berechnen. Bernholt und Fischer (2001) geben einen Algorithmus zur Berechnung der MCD-Schätzer an. In der Literatur wird die Verwendung der MCD-Parameterschätzer empfohlen, da diese bessere asymptotische Eigenschaften als die MVE-Schätzer aufweisen (siehe Davies (1992)).

5.1.2 MZE-Lokations- und MZE-Kovarianzschätzer

Ein weiteres Variabilitätsmaß für univariate Daten ist die so genannte mittlere absolute Abweichung vom arithmetischen Mittel:

$$d_n := \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}_n|.$$

Es wird im Folgenden gezeigt, dass die multivariate Erweiterung dieser Maßzahl durch das Volumen des geschätzten zentrierten Zonoids gegeben ist. Somit liegt die Idee nahe, multivariate robuste Lokations- und Kovarianzschätzer basierend auf diesem Variabilitätsmaß analog zum Vorgehen von Rousseeuw (1985) herzuleiten. Dazu wird diejenige Teilstichprobe vom Umfang $g = \lfloor (n+d+1)/2 \rfloor$ ausgewählt, für die das Volumen des auf der Teilstichprobe basierenden geschätzten zentrierten Zonoids diese Maßzahl minimiert. Die so ermittelte Teilstichprobe dient als Grundlage zur Schätzung der interessierenden Parameter. Dieses Vorgehen ergibt ebenfalls affin äquivariante Lokations- und Kovarianzschätzer mit dem höchst möglichen Bruchpunkt.
Zur Verdeutlichung des Zusammenhangs zwischen dem Volumen des geschätzten zentrierten Zonoids und der mittleren absoluten Abweichung vom arithmetischen Mittel wird zunächst der univariate Fall betrachtet.

Bemerkung 5.5 Die mittlere absolute Abweichung vom arithmetischen Mittel stimmt mit der Länge des geschätzten zentrierten Zonoids $Z(\tilde{X}_n - \bar{x})$ überein, da gilt (vgl.

Bemerkung 3.10):

$$\begin{aligned} d_n &= \frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}| = \sum_{i=1}^{n}\left|\frac{1}{n}det(x_i-\bar{x})\right| \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})I_{[0,\infty)}(x_i-\bar{x}) - \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(1-I_{[0,\infty)}(x_i-\bar{x})) \\ &= vol\left(\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(1-I_{[0,\infty)}(x_i-\bar{x})), \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})I_{[0,\infty)}(x_i-\bar{x})\right]\right) \\ &= vol(Z(\tilde{X}_n-\bar{x})), \end{aligned}$$

wobei $\tilde{X}_n - \bar{x} := \{x_1-\bar{x}, ..., x_n-\bar{x}\}$.

Im Folgenden wird der Zusammenhang zwischen dem Erwartungswert des Volumens von zufälligen Zonotopen und der Variabilität einer Verteilung verdeutlicht.

Erwartetes Volumen von Liftzonotopen und Zonotopen

Eine wichtige Maßzahl zur Charakterisierung von konvexen Körpern ist das Volumen. Das erwartete Volumen für zufällige Liftzonotope und Zonotope ergibt sich wie folgt:

Satz 5.1 Seien $X_1, ..., X_n$ $\mathbb{R}^d$-ZV u.i.v. mit $X_i \sim F \in \mathcal{F}_0$, $i = 1, ..., n$. Dann ist das erwartete Volumen des zufälligen Zonotops gegeben durch:

$$E(vol(Z(X_1, ..., X_n))) = \frac{n!}{n^d(n-d)!}vol(Z(F)). \tag{5.1}$$

Beweis: (Koshevoy und Mosler (1998))

Um diesen Beweis führen zu können, werden zwei Aussagen benötigt.

1. Es bezeichne M_X eine $d \times d$ Matrix, welche als Spalten u.i.v. Zufallsvektoren enthält, die wie X gemäß F verteilt sind. Unter diesen Voraussetzungen zeigt Vitale (1991), dass gilt:

$$E|det(M_X)| = d!\, vol(Z(F)). \tag{5.2}$$

2. Shepard (1974) zeigt, dass für Punkte $x_1, ..., x_n$ mit $x_i \in \mathbb{R}^d$ das Volumen eines Zonotops $\oplus_{i=1}^{n}[\mathbf{0}, x_i]$ gegeben ist durch:

$$\sum_{1\le i_1<...<i_d\le n} |det(x_{i_1}, ..., x_{i_d})|. \tag{5.3}$$

Hieraus folgt nun:

$$\begin{aligned}
E(vol(Z(X_1,...,X_n))) &\overset{(5.3)}{=} E\left(\sum_{1\leq i_1<...<i_d\leq n}\left|det\left(\frac{X_{i_1}}{n},...,\frac{X_{i_d}}{n}\right)\right|\right)\\
&= \frac{1}{n^d}\sum_{1\leq i_1<...<i_d\leq n} E\left(|det(X_{i_1},...,X_{i_d})|\right)\\
&\overset{(5.2)}{=} \frac{d!}{n^d}\sum_{1\leq i_1<...<i_d\leq n} vol(Z(F))\\
&= \frac{d!}{n^d}\binom{n}{d} vol(Z(F)).
\end{aligned}$$

q.e.d.

Bemerkung 5.6 1. Ein erwartungstreuer Schätzer für das Volumen des Zonoids ist somit gegeben durch:

$$\frac{(n-d)!}{n!}\sum_{1\leq i_1<...<i_d\leq n}|det(X_{i_1},...,X_{i_d})|.$$

2. Für Liftzonoide ist entsprechend ein erwartungstreuer Schätzer für das Volumen gegeben durch (vgl. Koshevoy und Mosler (1997)):

$$\frac{1}{(d+1)!}\binom{n}{d+1}^{-1}\sum_{1\leq i_1<...<i_d\leq n}\left|det\left(\left(1,X_{i_1}^T\right)^T,...,\left(1,X_{i_d}^T\right)^T\right)\right|.$$

3. Aus der Hadamard Determinanten Ungleichung (siehe z.B. Schott (1997)) folgt direkt, dass falls das Volumen von $Z(X_1,...,X_n)$ bzw. $LZ(X_1,...,X_n)$ kleiner unendlich ist, dann gilt $||X_i||<\infty$ für alle $i=1,...,n$.

Korollar 5.1 Gegeben sei eine Stichprobe von $X_1,...,X_n$ u.i.v. $\mathbf{R}$-ZV mit $X_i\sim F\in\mathcal{F}_0$, $i=1,...,n$. Es bezeichne $\tilde{Y}_n=\{X_1-\bar{X},...,X_n-\bar{X}\}$, wobei $\bar{X}$ das arithmetische Mittel sei. Dann ist:

$$vol(Z(\tilde{Y}_n))=\frac{1}{n}\sum_{i=1}^{n}|X_i-\bar{X}|,$$

ein erwartungstreuer Schätzer für das Volumen des zentrierten Zonoids.

Beweis:

Es folgt:

$$E(vol(Z(\tilde{Y}_n))) \;=\; E\left(\frac{1}{n}\sum_{i=1}^{n}\left|det\left(X_i-\bar{X}\right)\right|\right)=E\left(\frac{1}{n}\sum_{i=1}^{n}\left|X_i-\bar{X}\right|\right)\overset{(5.1)}{=} vol(Z(F)).$$

q.e.d.

Bemerkung 5.7 1. Für den Fall $d = 1$ entspricht das Volumen des zentrierten zufälligen Zonotops $Z(\tilde{Y}_n)$ der mittleren absoluten Abweichung vom arithmetischen Mittel.
2. Unter der Verteilungsannahme $X_i \sim N(\mu, \sigma^2)$, $i = 1, ..., n$, konvergiert $\frac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}|$ in Wahrscheinlichkeit gegen $\sqrt{\frac{2}{\pi}}\sigma$. Dies entspricht der Länge des Zonoids von $N(\mu, \sigma^2)$ (vgl. Beispiel 3.1 für $d = 1$).

Das Volumen des Zonoids einer Normalverteilung ist gegeben durch (vgl. Beispiel 3.1):

$$\begin{aligned} vol\left(Z(N(\mu,\Sigma)\right) &= vol\left(\left\{x\in\mathbf{R}^d : (x-\frac{1}{2}\mu)^T\Sigma^{-1}(x-\frac{1}{2}\mu)\le\frac{1}{2\pi}\right\}\right) \\ &= \sqrt{\lambda_1(\Sigma)\cdots\lambda_d(\Sigma)}\;\, vol\left(B\left(\frac{1}{2}\mu,\sqrt{\frac{1}{2\pi}}\right)\right). \end{aligned}$$

Das Volumen des Zonoids $Z(F)$ ist demnach proportional zu σ falls $d = 1$. Für $d > 1$ ist es proportional zum Produkt der Eigenwerte der Kovarianzmatrix Σ. Somit gibt das Volumen des Zonoids Aufschluss über die Variabilität des zugehörigen Zufallsvektors $X \sim N(\mu, \Sigma)$.

Es liegt nahe, analog zu dem Vorgehen von Rousseeuw (1985), die Teilstichprobe vom Umfang $g = \lfloor (n + d + 1)/2 \rfloor$ zu ermitteln, die das Volumen des zentrierten Zonotops minimiert, um dann mit dieser Teilstichprobe interessierende Parameter zu schätzen. Dies führt zu folgender Definition:

Definition 5.3 *Sei $\tilde{X}_n = \{x_1, ..., x_n\}$, $x_i \in \mathbb{R}^d$, $i = 1, ..., n$, eine Stichprobe aus einer Normalverteilung mit Parametern $\mu \in \mathbb{R}^d$ und $\Sigma \in \mathbb{R}^{d \times d}$ in allgemeiner Lage vom Umfang $n \geq d+1$. Sei $g = \lfloor (n+d+1)/2 \rfloor$. Es werden alle $\binom{n}{g}$ $g-$elementigen Teilstichproben gebildet. Dann bezeichne:*
$\tilde{X}_{g,Z}$ diejenige g-elementige Teilstichprobe, deren geschätztes zentriertes Zonoid minimales Volumen besitzt.

Lemma 5.3 Sei $\tilde{X}_n = \{x_1, ..., x_n\}$ eine Stichprobe vom Umfang $n \geq d+1$ mit $x_i \in \mathbb{R}^d$, $i = 1, ..., n$, A eine nichtsinguläre $d \times d$ Matrix und $b \in \mathbb{R}^d$. Weiterhin sei $\tilde{X}_{g,Z}$ wie in Definition 5.3 definiert. Wird die Stichprobe $A\tilde{X}_n + b$ betrachtet, so gilt:

$$\left(A\tilde{X}_n + b\right)_{g,Z} = A\tilde{X}_{g,Z} + b.$$

Beweis:
Zunächst folgt aus Formel 5.3 für jede Stichprobe $\tilde{X}_n$ vom Umfang n:

$$\begin{aligned} vol(Z(\tilde{X}_n - \bar{x})) &= \sum_{1 \leq i_1 < ... < i_d \leq n} \left| det\left(\frac{1}{n}(x_{i_1} - \bar{x}), ..., \frac{1}{n}(x_{i_d} - \bar{x})\right)\right| \\ &= vol\left(Z\left(\tilde{X}_n + b - (\bar{x} + b)\right)\right) \end{aligned}$$

die Translationsinvarianz.
Für jede nicht singuläre $d \times d$ Matrix A gilt:

$$\begin{aligned} vol((Z(A(\tilde{X}_n - \bar{x}))) &= \sum_{1 \leq i_1 < ... < i_d \leq n} \left| det\left(\frac{1}{n}(Ax_{i_1} - A\bar{x}), ..., \frac{1}{n}A(x_{i_d} - A\bar{x})\right)\right| \\ &= \sum_{1 \leq i_1 < ... < i_d \leq n} \left| det(A) det\left(\frac{1}{n}(x_{i_1} - \bar{x}), ..., \frac{1}{n}(x_{i_d} - \bar{x})\right)\right| \\ &= |det(A)| vol(Z(\tilde{X}_n - \bar{x})). \end{aligned}$$

Da $|det(A)|$ konstant ist, folgt die Behauptung.

q.e.d.

Somit ergibt sich die folgende Definition für Lokations- und Kovarianzschätzer.

Definition 5.4 *Sei $\tilde{X}_{g,Z}$ wie in Definition 5.3 gegeben.*
Der Minimum Zonoid (MZE) Lokationsschätzer $T_{g,Z}(\tilde{X}_n)$ ist definiert als das arithmetische Mittel aus der Teilstichprobe $\tilde{X}_{g,Z}$.

Der MZE-Kovarianzschätzer $C_{g,Z}(\tilde{X}_n)$ ergibt sich aus der empirischen Kovarianz der Teilstichprobe $\tilde{X}_{g,Z}$.

Lemma 5.3 sichert die affine Äquivarianz der in Definition 5.4 vorgeschlagenen Lokations- und Kovarianzschätzer.

Lemma 5.4 Sei $\tilde{X}_n$ eine Stichprobe vom Umfang $n \geq d+1$ in allgemeiner Lage mit $x_i \in \mathbb{R}^d$, $i = 1, ..., n$. Dann gilt für den finite-sample Bruchpunkt der Schätzer aus Definition 5.4 für $d = 1$:

$$\epsilon_L(\tilde{X}_n, T_{g,Z}) = \left\lfloor \frac{n+1}{2} \right\rfloor \Big/ n \text{ und } \epsilon_C(\tilde{X}_n, C_{g,Z}) = \left\lfloor \frac{n}{2} \right\rfloor \Big/ n$$

und für $d \geq 2$:

$$\epsilon_L(\tilde{X}_n, C_{g,Z}) = \epsilon_C(\tilde{X}_n, T_{g,Z}) = \left\lfloor \frac{n-d+1}{2} \right\rfloor \Big/ n.$$

Zum Beweis dieses Satzes wird das folgende Korollar benötigt:

Korollar 5.2 Gegeben sei eine Stichprobe $\tilde{X}_g$ vom Umfang g in allgemeiner Lage, wobei $x_i \in \mathbb{R}^d$, $i = 1, ..., g$. Sei $C_g = \frac{1}{g}\sum_{i=1}^g (x_i - \bar{x}_g)(x_i - \bar{x}_g)^T$ und $\bar{x}_g = \frac{1}{g}\sum_{i=1}^g x_i$. Dann gilt:

1. $det(C_g) \to \infty$ genau dann, wenn $vol\left(Z(\tilde{X}_g - \bar{x}_g)\right) = \infty$.

2. $det(C_g) = 0$ genau dann, wenn $vol\left(Z(\tilde{X}_g - \bar{x}_g)\right) = 0$.

Beweis:

Sei o.B.d.A. $\bar{x}_g = 0$.

Zu 1.

$\Rightarrow$ Es strebe $det(C_g) = \infty$, dann existiert mindestens eine Beobachtung $\tilde{x}_i \in \tilde{X}_g$ mit $||\tilde{x}_i|| \to \infty$ und $rang\left(\tilde{X}_g\right) = d$, dabei bezeichne $rang\,(A)$ den Rang der Matrix A. Da dann gilt $0 < |det(\tilde{x}_{i_1}, ..., x_{i_d})| \leq vol\left(Z(\tilde{X}_g)\right) = \infty$ folgt die Behauptung.

$\Leftarrow$ Es gelte $vol\left(Z(\tilde{X}_g)\right) = \infty$, dann folgt aus der Hadamard-Determinanten-Ungleichung, dass ein $x_i \in \tilde{X}_g$ existiert, so dass $||x_i|| \to \infty$ und $rang\left(\tilde{X}_g\right) = d$. Daraus folgt $\lambda_d(C_g) \to \infty$ und somit auch $det(C_g) = \prod_{i=1}^d \lambda_i(C_g) = \infty$.

Zu 2.
Sei $vol\left(Z(\tilde{X}_g)\right) = 0 \Leftrightarrow$ für alle $1 \leq i_1 \leq \cdots \leq i_d \leq g$ gilt $|det(x_{i_1}, ..., x_{i_d})| = 0 \Leftrightarrow rang\left(\tilde{X}_g\right) \leq (d-1) \Leftrightarrow det(C_g) = 0.$

q.e.d.

Beweis von Lemma 5.4
Für den MZE-Lokationsschätzer folgt die Behauptung aus Satz 5.2 auf Seite 82.
Zu zeigen bleibt, dass für $d \geq 2$ gilt: $\epsilon_C(\tilde{X}_n, T_{g,Z}) = \left\lfloor \frac{n-d+1}{2} \right\rfloor \Big/ n$. Für $d = 1$ gilt dieselbe Argumentation.
Angenommen der Bruchpunkt des MZE-Kovarianzschätzers wäre gerade $\left\lfloor \frac{n-d}{2} \right\rfloor \Big/ n$. Dies würde bedeuten, dass durch das Ersetzen von $\left\lfloor \frac{n-d}{2} \right\rfloor$ Beobachtungen der Stichprobe $\tilde{X}_n$ erreicht wird, dass $\lambda_d(C(\tilde{X}_{g,Z})) = \infty$ bzw. $det(C(\tilde{X}_{g,Z})) = \infty$ oder $\lambda_1(C(\tilde{X}_{g,Z})) = 0$ bzw. $det(C(\tilde{X}_{g,Z})) = 0$. Aus Korollar 5.2 folgt dann, dass gelten muss $vol\left(Z(\tilde{X}_{g,Z} - \bar{x}_{g,Z})\right) = \infty$ oder $vol\left(Z(\tilde{X}_{g,Z} - \bar{x}_{g,Z})\right) = 0$. Nun beträgt der Bruchpunkt von $C(\tilde{X}_{g,C})$ gerade $\left\lfloor \frac{n-d+1}{2} \right\rfloor \Big/ n$ (siehe Lemma 5.2). Somit existiert mindestens eine Teilstichprobe vom Umfang $g = \left\lfloor \frac{n+d+1}{2} \right\rfloor$, so dass $0 < \lambda_1(C(\tilde{X}_{g,C})) \leq \cdots \leq \lambda_d(C(\tilde{X}_{g,C})) < \infty$. Dies ist aber ein Widerspruch zur Annahme, da mindestens eine durch das MZE-Kriterium ausgewählte Stichprobe existieren muss, nämlich $\tilde{X}_{g,C} = \tilde{X}_{g,Z}$, so dass $0 < \lambda_1(C(\tilde{X}_{g,Z})) \leq \cdots \leq \lambda_d(C(\tilde{X}_{g,Z})) < \infty$. Somit folgt: $\epsilon_C(\tilde{X}_n, T_{g,Z}) = \left\lfloor \frac{n-d+1}{2} \right\rfloor \Big/ n$, da dies der oberen Schranke für affin äquivariante Kovarianzschätzer entspricht (siehe Davies (1987)).

q.e.d.

Auch die Berechnung der auf dem MZE-Kriterium basierenden Schätzer ist sehr rechenintensiv. Im Folgenden wird anhand eines Beispiels die Schätzung erläutert.

Beispiel 5.1 Gegeben sei eine Stichprobe $\tilde{X}_6 = \{x_1, ..., x_6\}$ mit $x_1 = \binom{-1.33}{0.21}, x_2 = \binom{1.74}{1.24}, x_3 = \binom{-0.09}{0.30}, x_4 = \binom{0.27}{1.71}, x_5 = \binom{5}{5}$ und $x_6 = \binom{-3}{0}$ vom Umfang sechs. Die ersten vier Beobachtungen sind Realisationen einer bivariaten Standardnormalverteilung. Die Beobachtungen x_5 und x_6 wurden künstlich hinzugefügt.
Die auf dem MCD-Kriterium basierende Teilstichprobe vom Umfang 4 ist gegeben

durch $\tilde{X}_{g,C} = \{x_1, x_2, x_3, x_6\}$. Dabei ist $det\Big(C_{g,C}(\tilde{X}_n)\Big) = 0.188$. Dagegen berechnet sich die auf dem MZE-Kriterium bestimmte Teilstichprobe als $\tilde{X}_{g,Z} = \{x_1, x_2, x_3, x_4\}$, mit $vol\Big(Z(\tilde{X}_{g,Z} - \bar{x}_{g,Z})\Big) = 1.67$. In diesem Spezialfall unterscheiden sich die Ergebnisse beider Kriterien voneinander. Man beachte, dass $det\Big(C_{g,Z}(\tilde{X}_n)\Big) = 0.489$ gilt.
Die Lokationsschätzer sind dann gegeben durch:

$$T_{g,Z}(\tilde{X}_n) = \begin{pmatrix} 0.15 \\ 0.87 \end{pmatrix} \quad \text{und} \quad T_{g,C}(\tilde{X}_n) = \begin{pmatrix} -0.67 \\ 0.44 \end{pmatrix}$$

und die entsprechenden Kovarianzschätzer:

$$C_{g,Z}(\tilde{X}_n) = \begin{pmatrix} 1.60 & 0.60 \\ 0.60 & 0.53 \end{pmatrix} \quad \text{und} \quad C_{g,C}(\tilde{X}_n) = \begin{pmatrix} 4.01 & 1.02 \\ 1.02 & 0.30 \end{pmatrix}.$$

5.2 Schätzer konvexer Körper mit hohem Bruchpunkt

Ziel dieses Abschnittes ist die Konstruktion von Schätzern konvexer Körper mit hohem Bruchpunkt. Im Abschnitt 5.1 wurde gezeigt, dass die Verwendung der auf den Teilstichproben $\tilde{X}_{g,E}$, $\tilde{X}_{g,C}$ und $\tilde{X}_{g,Z}$ basierenden Schätzer der Parameter μ und Σ einen hohen Bruchpunkt haben. Basierend auf diesen Stichproben werden nun auch Schätzer konvexer Körper bestimmt, die ebenso einen hohen Bruchpunkt aufweisen. Formal ergeben sich diese Schätzer wie folgt:

Definition 5.5 *Gegeben sei eine Stichprobe $\tilde{X}_n$ vom Umfang n in allgemeiner Lage mit $n \geq (d+1)$. Sei T ein Funktional zur Schätzung eines konvexen Körpers basierend auf der Stichprobe $\tilde{X}_n$. Seien $\tilde{X}_{g,E}, \tilde{X}_{g,C}$ und $\tilde{X}_{g,Z}$ wie in Definition 5.1 und 5.3 definiert, dann heißt:*

1. $T(\tilde{X}_{g,E}) =: T_{g,E}(\tilde{X}_n)$ *MVE-Schätzfunktional eines konvexen Körpers.*

2. $T(\tilde{X}_{g,C}) =: T_{g,C}(\tilde{X}_n)$ *MCD-Schätzfunktional eines konvexen Körpers.*

3. $T(\tilde{X}_{g,Z}) =: T_{g,Z}(\tilde{X}_n)$ *MZE-Schätzfunktional eines konvexen Körpers.*

Bemerkung 5.8 Falls T angewandt auf $\tilde{X}_n$ ein affin äquivarianter Schätzer eines konvexen Körpers ist, so folgt aus den Korollaren 5.1 und 5.3 auch die affine Äquivarianz für $T_{g,E}(\tilde{X}_n)$, $T_{g,C}(\tilde{X}_n)$ und $T_{g,Z}(\tilde{X}_n)$.

Bemerkung 5.9 Stammen die Daten aus einer elliptisch symmetrischen Verteilung, so reduziert sich die Schätzung konvexer Körper häufig auf die Schätzung einer Ellipse (siehe Lemma 4.1 und Beispiel 3.1). Um einen hohen Bruchpunkt des Schätzers der Ellipse nach Definition 2.10 zu garantieren, müssen Lokations- bzw. Kovarianzschätzer mit hohem Bruchpunkt nach Definition 2.3 bzw. 2.4 verwendet werden. Somit folgt direkt, dass bei Verwendung der im vorherigen Abschnitt vorgestellten Lokations- und Kovarianzschätzer auch die zugehörige Schätzung des konvexen Körpers den höchst möglichen Bruchpunkt für affin äquivariante Schätzer konvexer Körper annimmt (siehe Lemma 2.7).

Beispiel 5.2 In Abbildung 5.1 ist die Schätzung der Mahalanobis-Konturen (siehe Abschnitt 4.2.1) einmal mit dem arithmetischen Mittel und der empirischen Kovarianz der Stichprobe aus Beispiel 5.1 und zum anderen mit dem auf dem MZE-Kriterium basierenden Lokations- und Kovarianzschätzer für $\alpha = \{0.9, 0.7, 0.5\}$ angegeben.
Deutlich zu sehen sind die Unterschiede der beiden Schätzungen, die durch die vorhandenen Ausreißer zu Stande kommen.

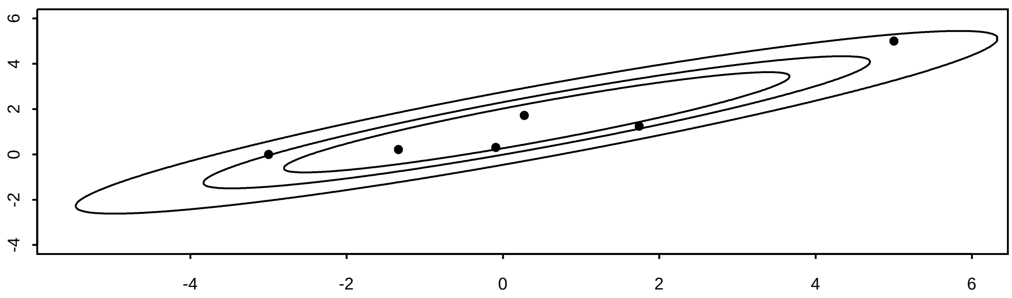
Im Folgenden wird die Schätzerklasse der Polytope betrachtet. Diese konvexen Körper lassen sich wie folgt darstellen:

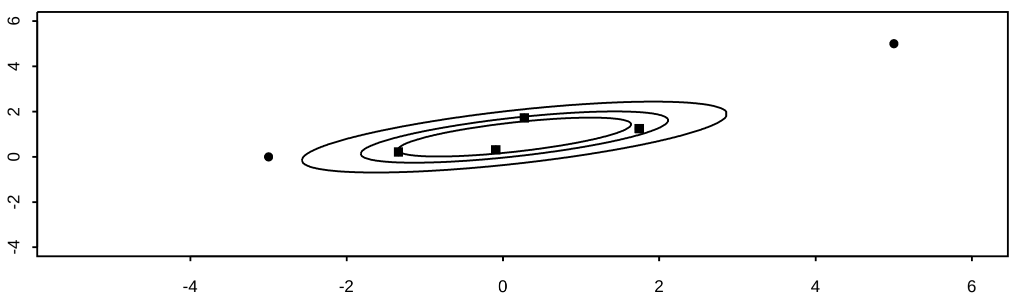
Definition 5.6 *Gegeben sei eine Stichprobe* $\tilde{X}_n = \{x_1, ..., x_n\}$ *vom Umfang* n *mit* $x_i \in \mathbb{R}^d$. *Dann heißt:*

$$T_P(\tilde{X}_n) = conv\{q_1, ..., q_u\},$$

Schätzpolytop. Dabei sind die Punkte q_t *darstellbar als* $q_t = \sum_{i=1}^{n} \lambda_{t,i} x_i$ *mit* $0 \leq \lambda_{t,i} \leq 1$ *und* $t = 1, ..., u$, *wobei* $u < \infty$ *und zu jedem* $||x_i|| \neq \mathbf{0}$ *mindestens ein* $\lambda_{t,i}$ *existiert mit* $||\lambda_{t,i} x_i|| \neq \mathbf{0}$.

Abbildung 5.1: Vergleich der Schätzung mit kontaminierter Stichprobe und mit MZE-Kriterium bereinigter Stichprobe für Mahalanobis-Konturen





Bemerkung 5.10 Somit fließt bei der Bestimmung eines Schätzpolytops jede Beobachtung mindestens einmal mit einem Gewicht $\lambda_{t,i} > 0$ in die Schätzung ein.
Falls $u = 1$ und $\lambda_i = \frac{1}{n}$ ist, so entspricht $T_P(\tilde{X}_n)$ dem arithmetischen Mittel der Stichprobe $\tilde{X}_n$.
Polytope sind invariant gegenüber affin äquivarianten Transformationen. Die Schätzung von Liftzonoiden, Zonoiden und (α-getrimmten) zonoiden Zonen lässt sich auch durch Polytope darstellen.
Die Stützfunktion von $T_P(\tilde{X}_n) = conv\{q_1, ..., q_u\}$ ist gegeben durch (siehe Lemma 2.2):

$$h(T_P(\tilde{X}_n), p) = \max_{1 \leq t \leq u} \langle q_t, p \rangle.$$

Beispiel 5.3 Gegeben sei eine $d-$dimensionale Stichprobe $\tilde{X}_n$ vom Umfang n.

1. Die Schätzung des Zonoids ist gegeben durch (siehe Lemma 3.9):

$$Z(\tilde{X}_n) = conv\Big\{q_1, ..., q_u\Big\},$$

dabei ist $q_t = \sum_{i=1}^{n} \lambda_{t,i} x_i$ mit $\lambda_{t,i} \in \{0, 1/n\}$, $i \in \{1, ..., n\}$ und $u = 2^n$.

2. Die Schätzung von α-getrimmten zonoiden Zonen ist gegeben durch (siehe Satz 4.3):

$$Z_\alpha(\tilde{X}_n) = conv\Big\{q_1, ..., q_u\Big\},$$

wobei $q_t = \frac{1}{\alpha n} \sum_{j=1}^{k+1} \lambda_{t,j} x_{i_j}$, $\{i_1, ..., i_{k+1}\} \subset \{1, ..., n\}$, $\lambda_{t,j} = \frac{1}{\alpha n}$ für $i = 1, ..., k$, $\lambda_{t,(k+1)} = (1 - \frac{k}{\alpha n})$ und $u = \binom{n}{(k+1)}$.

Im Folgenden wird der Explosions-Bruchpunkt eines MZE-Schätzers für einen konvexen Körper beruhend auf einem Polytop ermittelt.

Satz 5.2 Für jede $d-$dimensionale Stichprobe $\tilde{X}_n$ vom Umfang n, mit $n \geq d + 1$ in allgemeiner Lage gilt für den Explosions-Bruchpunkt nach Definition 2.6 eines MZE-Polytopschätzers:

$$\epsilon_K(\tilde{X}_n, T_{P,g,Z}) = \begin{cases} \left\lfloor \frac{(n+1)}{2} \right\rfloor \Big/ n & : \text{ falls } d = 1, \\ \\ \left\lfloor \frac{(n-d+1)}{2} \right\rfloor \Big/ n & : \text{ falls } d > 1. \end{cases}$$

Beweis:

Sei $d > 1$. Im ersten Schritt wird gezeigt, dass der Bruchpunkt größer oder gleich $\lfloor (n - d + 1)/2 \rfloor / n$ sein muss.

Sei $\tilde{X}_n$ eine Stichprobe im $\mathbf{R}^d$ vom Umfang n in allgemeiner Lage.

Bezeichne

$\mathcal{X}_g$ die Menge aller Teilstichproben vom Umfang $g = \lfloor (n + d + 1)/2 \rfloor$ aus $\tilde{X}_n$.

Sei nun $\tilde{Y}_{n,k}$ eine durch Austausch von $k = \lfloor (n - d + 1)/2 \rfloor - 1$ Beobachtungen von $\tilde{X}_n$ durch beliebige Vektoren entstehende Stichprobe.

Bezeichne

$\mathcal{Y}_g$ die Menge aller Teilstichproben vom Umfang $g = \lfloor (n + d + 1)/2 \rfloor$ aus $\tilde{Y}_{n,k}$.

Es existiert mindestens eine Stichprobe aus $\mathcal{Y}_g$, die gerade $\lfloor(n{+}d{+}1)/2\rfloor$ Beobachtungen aus $\tilde{X}_n$ enthält, da gilt: $n - (\lfloor(n-d+1)/2\rfloor - 1) \geq \lfloor(n+d+1)/2\rfloor$. Daher gibt es mindestens eine Teilstichprobe vom Umfang g, die sowohl in $\mathcal{Y}_g$ als auch in $\mathcal{X}_g$ enthalten ist. Bezeichne die Menge all dieser Teilstichproben mit:

$$\tilde{\mathcal{X}}_g = \mathcal{Y}_g \cap \mathcal{X}_g.$$

Sei $\tilde{X}_{g,Z} \in \tilde{\mathcal{X}}_g$ diejenige Teilstichprobe mit zugehörigem arithmetischen Mittel $\bar{x}_{g,Z}$, die das Volumen des geschätzten zentrierten Zonoids

$$\min_{\tilde{X}_g \in \tilde{\mathcal{X}}_g} vol\left(Z(\tilde{X}_g - \bar{x}_g)\right) = vol\left(Z(\tilde{X}_{g,Z} - \bar{x}_{g,Z})\right)$$

minimiert.
Sei nun $\tilde{Y}_{g,Z} \in \mathcal{Y}_g$ diejenige Teilstichprobe, die den folgenden Ausdruck minimiert

$$\min_{\tilde{Y}_g \in \mathcal{Y}_g} vol\left(Z(\tilde{Y}_g - \bar{y}_g)\right) = vol\left(Z(\tilde{Y}_{g,Z} - \bar{y}_{g,Z})\right).$$

Dabei sei $\bar{y}_{g,Z}$ das arithmetische Mittel von $\tilde{Y}_{g,Z}$. Nun enthält $\tilde{Y}_{g,Z}$ mindestens $(d+1)$ Beobachtungen aus $\tilde{X}_n$, da $\lfloor(n+d+1)/2\rfloor - (\lfloor(n-d+1)/2\rfloor - 1) = d+1$. Da $\tilde{X}_n$ eine Stichprobe in allgemeiner Lage ist, ist das Volumen des geschätzten zentrierten Zonoids der Teilstichprobe $\tilde{Y}_{g,Z}$ immer größer Null. Weiterhin gilt $\tilde{\mathcal{X}}_g \subseteq \mathcal{Y}_g$.
Aus diesen Überlegungen folgt:

$$0 < vol\left(Z(\tilde{Y}_{g,Z} - \bar{y}_{g,Z})\right) \leq vol\left(Z(\tilde{X}_{g,Z} - \bar{x}_{g,Z})\right) < \infty. \tag{5.4}$$

Unter der Annahme, dass die Stützfunktion des geschätzten Körpers gegen ∞ strebt, also falls für ein festes $p \in \mathbf{R}^d$ mit $||p|| = 1$ für die Stützfunktion gilt:

$$\left|h\left(T_P(\tilde{Y}_{g,Z}), p\right)\right| = \max_{1 \leq t \leq u} \left\langle \sum_{i=1}^{g} \lambda_{t,i} y_i, p \right\rangle \longrightarrow \infty,$$

muss mindestens eine Beobachtung y_i existieren mit $||y_i|| \to \infty$. Dies lässt sich schreiben als $y_i = ty$ mit $t \to \infty$, $t > 0$. In diesem Fall gilt für das Volumen des geschätzten zentrierten Zonoids:

$$vol\left(Z(\tilde{Y}_{g,Z} - \bar{y}_{g,Z})\right) = \sum_{1 \leq i_1 < \ldots < i_d \leq g} \left|det\left(\frac{1}{g}\left(y_{i_1} - \bar{y}_{g,Z}\right), \ldots, \frac{1}{g}\left(y_{i_d} - \bar{y}_{g,Z}\right)\right)\right|$$

$$
\begin{aligned}
&\geq \left| det\left(\frac{1}{g}\left(y_{i_1} - \bar{y}_{g,Z}\right), ..., \frac{1}{g}\left(y_{i_d} - \bar{y}_{g,Z}\right)\right)\right| \text{ mit } y_{i_j} \neq ty, j = 1, ..., d \\
&= \left| det\left(\left(\frac{1}{g}y_{i_1}, ..., \frac{1}{g}y_{i_d}\right) - \bar{y}_{g,Z}1_d^T\right)\right| \\
&= \left| det\left(\frac{1}{g}y_{i_1}, ..., \frac{1}{g}y_{1_d}\right)\left(1 - 1_d^T\left(\frac{1}{g}y_{i_1}, ..., \frac{1}{g}y_{i_d}\right)^{-1}\bar{y}_{g,Z}\right)\right| \longrightarrow \infty,
\end{aligned}
$$

falls t und somit auch $\bar{y}_{g,Z} = \frac{1}{g}ty + \frac{1}{g}\sum_{y_i \in \tilde{Y}_{g,Z} \setminus \{ty\}} y_i$ gegen unendlich strebt.

Dies ist aber ein Widerspruch zu Formel (5.4), da

$$0 < vol\left(Z(\tilde{Y}_{g,Z} - \bar{y}_{g,Z})\right) \leq vol\left(Z(\tilde{X}_{g,Z} - \bar{x}_{g,Z})\right) < \infty.$$

Demnach muss für alle $p \in \mathbf{R}^d$ mit $||p|| = 1$ gelten: $|h(T_P(\tilde{Y}_{g,Z}), p)| < \infty$.

Hieraus folgt, dass der Bruchpunkt eines MZE-Polytopschätzers mindestens $\lfloor (n-d+1)/2 \rfloor / n$ sein muss.

Um zu zeigen, dass der Bruchpunkt gerade $\lfloor (n-d+1)/2 \rfloor / n$ ist, werden d beliebige Punkte aus $\tilde{X}_n$ gewählt und der $(d-1)$-affine Unterraum H betrachtet, den diese Punkte aufspannen. Nun werden $g = \lfloor (n-d+1)/2 \rfloor$ Punkte von $\tilde{X}_n$ durch Punkte, die in H liegen, ersetzt. Die so entstehende Stichprobe werde mit $\tilde{Y}_{n,g}$ bezeichnet. Dann enthält H gerade $\lfloor (n-d+1)/2 \rfloor$ Punkte der Stichprobe $\tilde{Y}_{n,g}$. Das Volumen des geschätzten zentrierten Zonoids der Stichprobe $\tilde{Y}_{n,g}$, dessen Elemente alle in H liegen, ist Null. Da die Stichprobe $\tilde{X}_n$ nach Voraussetzung in allgemeiner Lage liegt, kann kein mit einer anderen Teilstichprobe geschätztes Zonoid ein Volumen von Null haben. Somit ist $\tilde{Y}_{n,g} = \tilde{Y}_{g,Z}$ und $T_P(\tilde{Y}_{g,Z})$ muss in H liegen. Da H nicht beschränkt ist, kann die Norm der $\lfloor (n-d+1)/2 \rfloor$ kontaminierten Beobachtungen beliebig groß werden. Somit existiert ein $p \in \mathbf{R}^d$ mit $||p|| = 1$, so dass $h(T_P(\tilde{Y}_{g,Z}), p) \to \infty$, falls die Norm einer der Beobachtungen, die in H liegen, beliebig groß wird.

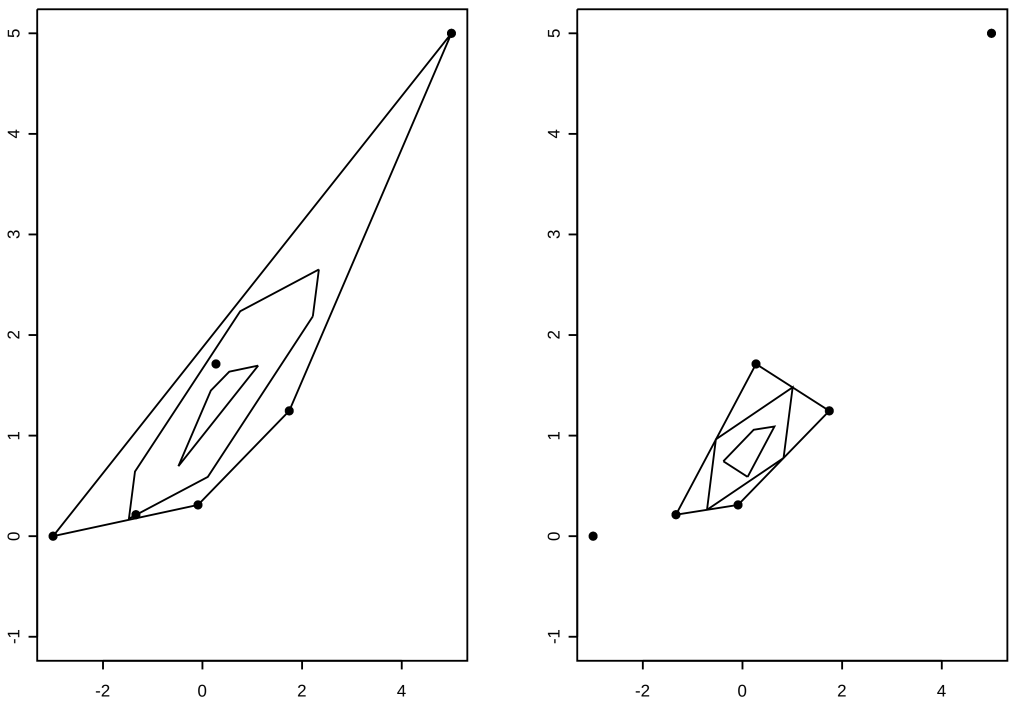
Hieraus folgt die Behauptung.

Für $d = 1$ erfolgt der Beweis mit denselben Überlegungen.

q.e.d.

Bemerkung 5.11 Die Konstruktion eines MZE-Polytopschätzers benötigt im Gegensatz zum MCD-Polytopschätzer nicht die Annahme, dass die Stichprobe einer Verteilung entstammt, deren zweites Moment existiert.

Abbildung 5.2: Vergleich der Schätzung mit kontaminierter Stichprobe und mit MZE-Kriterium bereinigter Stichprobe für zonoide Zonen



Wird der MZD-Polytopschätzer zur Schätzung benutzt, so folgt mit einer ähnlichen Argumentation wie im Beweis von Satz 5.2, dass dieser denselben Explosions-Bruchpunkt aufweist wie der MZE-Polytopschätzer.

Beispiel 5.4 In Abbildung 5.2 wurde der Datensatz aus Beispiel 5.1 (für $\alpha = \{0.16, 0.5, 0.75\}$) und die mittels des MZE-Kriterium bereinigte Teilstichprobe (für $\alpha = \{0.25, 0.5, 0.75\}$) zu Schätzung von zonoiden Zonen (siehe Abbschnitt 4.2.2) verwendet. Auch in diesem Fall wird deutlich wie die Schätzung der einzelnen Bereiche durch Ausreißer verfälscht wird.

Kapitel 6

Ausblick

Eine wichtige Maßzahl zur Beurteilung eines Schätzers ist der finite-sample Bruchpunkt. Diese Größe gibt an, wie groß der Anteil beliebig schlecht platzierter Beobachtungen in einer Stichprobe mindestens sein muss, damit der Schätzer „zusammenbricht“. In dieser Arbeit wird der finite-sample Bruchpunkt für Schätzer konvexer Körper definiert. Der Zusammenbruch wird durch den Hausdorff Abstand erklärt. Dies erfolgt im zweiten Kapitel. Von einem Zusammenbruch eines solchen Schätzers wird einerseits gesprochen, wenn der Hausdorff Abstand zwischen der Schätzung des konvexen Körpers basierend auf der regulären und der kontaminierten Stichprobe beliebig groß wird. Andererseits muss ebenfalls von einem Zusammenbruch gesprochen werden, wenn der Schätzer des konvexen Körpers durch eine verunreinigte Stichprobe zu einem $(d-1)$-dimensionalen Gebilde degeneriert. Dieser Fall tritt ein, wenn der Hausdorff Abstand zwischen der Schätzung der Polarmenge des konvexen Körpers basierend auf der regulären und der kontaminierten Stichprobe beliebig groß wird. Die im zweiten Kapitel vorgestellte Definition betrachtet beide Arten des Zusammenbruchs und erlaubt somit die Untersuchung vorhandener Schätzer konvexer Körper bzgl. ihres Bruchpunktverhaltens.

Im dritten und vierten Kapitel werden Schätzer konvexer Körper vorgestellt. Dabei beruht die Schätzung zum einem auf konvexen Polytopen und zum anderen, da es eine eineindeutige Verbindung zwischen einigen konvexen Körpern und den Momenten der

Verteilung gibt, auf Schätzungen des Erwartungswertes und der Kovarianz. Der Bruchpunkt der betrachteten Schätzer konvexer Körper wird bestimmt. Es zeigt sich, dass viele Schätzer schon bei Austausch einer einzigen Beobachtung zusammenbrechen.

Im fünften Kapitel werden Schätzer konvexer Körper vorgestellt, die einen hohen Bruchpunkt aufweisen. Diese Schätzer gehören zur Klasse der „half-sample" Schätzer. Zur Schätzung wird eine geeignete Teilmenge der Stichprobe verwendet, die höchstens die Hälfte aller Punkte umfasst. Dabei soll diese Teilmenge möglichst wenige Ausreißer enthalten. Zur Bestimmung geeigneter Teilstichproben werden zum einen die von Rousseeuw (1983) vorgeschlagenen MVE- bzw. MCD-Kriterien benutzt und zum anderen das in dieser Arbeit vorgestellte MZE-Kriterium.

Des Weiteren werden affin äquivariante Lokations- bzw. Kovarianzschätzer eingeführt, die auf dem MZE-Kriterium beruhen. Diese Schätzer nehmen die obere Bruchpunkt-Schranke für ihre jeweiligen Schätzerklassen an.

Ein Ansatzpunkt für weitere Forschung wäre die Untersuchung der im fünften Kapitel vorgeschlagenen MZE-Lokations- bzw. MZE-Kovarianzschätzer, bezüglich ihrer Konsistenz sowie ihrer asymptotischen Verteilung. Weiterhin gibt es bislang keine effizienten Algorithmen, welche die praktische Berechnung dieser Schätzer für einen großen Stichprobenumfang ermöglichen.

Die in dieser Arbeit vorgeschlagenen affin äquivarianten Schätzer, die den höchsten Bruchpunkt bzgl. dieser Schätzerklasse aufweisen, benötigen die restriktive Annahme, dass die Stichprobe in allgemeiner Lage ist. In der Arbeit von Davies und Gather (2002) wird ein univariates Varianz-Funktional hergeleitet, das den höchst möglichen Bruchpunkt innerhalb seiner Schätzerklasse annimmt, ohne diese Annahme zu benötigen. Somit stellt sich die Frage, inwieweit die in dieser Arbeit vorgeschlagenen Schätzer modifiziert werden können, um ebenfalls den höchst möglichen Bruchpunkt zu erhalten, ohne die Annahme, dass die Stichprobe in allgemeiner Lage ist.

Des Weiteren ist die Herleitung einer Influenzfunktion für Funktionale konvexer Körper, mit deren Hilfe zum Beispiel die „gross error sensitivity" bestimmt wird, von Interesse für zukünftige Forschung.

Symbolverzeichnis

$X_1, ..., X_n$	$\mathbb{R}^d$ Zufallsvektoren, S. 5
F	Verteilung, S. 5
$\mathcal{F}$	Veteilungsklasse, S. 5
Θ	Parameterraum, S. 6
$N_{\mu,\Sigma}$	Normalverteilung mit Erwartungswert μ und Kovarianzmatrix Σ, S. 6
K	konvexer Körper des $\mathbb{R}^d$, S. 6
$\mathcal{K}^d$	Menge aller konvexen Körper des $\mathbb{R}^d$, S. 6
T	Funktional eines konvexen Körpers, S. 6
$\|\| \cdot \|\|$	euklidische Norm des $\mathbb{R}^d$, S. 7
$d_H(\cdot, \cdot)$	Hausdorff Abstand, S. 7
$\tilde{X}_n = \{x_1, ..., x_n\}$	Stichprobe vom Umfang n, S. 8
$\oplus$	Minkowski Summe, S. 9
ϵ_L	finite-sample Bruchpunkt für Lokations-schätzer, S. 9
$\lambda_d(A) \geq ... \geq \lambda_1(A)$	Eigenwerte der Matrix $A \in \mathbb{R}^{d \times d}$, S. 10
ϵ_C	finite-sample Bruchpunkt für Kovarianz-schätzer, S. 10
$tr(A)$	Spur der Matrix A, S. 11
ϵ_K	finite-sample Explosions-Bruchpunkt für Schätzer Schätzer konvexer Körper, S. 13
$h(K, \cdot)$	Stützfunktion des konvexen Körpers K, S. 14

$g(K,\cdot)$	Distanzfunktion des konvexen Körpers K, S. 19
K^*	Polarmenge von $K \in \mathcal{K}^d$, S. 20
$\epsilon_{K,K*}$	finite-sample Bruchpunkt für Schätzer konvexer Körper, S. 22
$\lfloor x \rfloor$	größte ganze Zahl kleiner oder gleich x, S. 27
$(\Omega, \mathcal{A}, P)$	Wahrscheinlichkeitsraum, S. 32
$\mathcal{F}_0$	Menge der Verteilung, deren erstes Moment existiert, S. 36
$Z(F)$	Zonoid von Verteilung F, S. 36
$I_{\{\ldots\}}$	Indikatorfunktion, S. 37
$LZ(F)$	Liftzonoid von Verteilung F, S. 41
$\stackrel{d}{=}$	Gleichheit der Verteilung auf beiden Seiten des Zeichens, S. 44
χ	Stichprobenraum, S. 52
$\mathcal{D}^d$	Menge aller konvexen Mengen des $\mathbf{R}^d$, S. 53
$D_F(\cdot)$	Kontur-Toleranzbereich von F, S. 53
$det(A)$	Determinante von A, S. 54
$DT_F(\cdot)$	Datentiefe-Funktion von F, S. 54
$vol(K)$	Volumen des konvexen Körpers K, S. 73
$rang(A)$	Rang der Matrix A, S.77

Abbildungsverzeichnis

Literaturverzeichnis

[1] Artstein, Z. (1974), On the Calculus of Closed Set-valued Functions, *Indiana University Mathematics Journal,* **24**, 433 − 441.

[2] Artstein, Z., Vitale, R.A (1975), A strong Law of large Numbers for Random Compact Sets, *The Annals of Probability,* **3**, 879 − 882.

[3] Becker, C. (2001), *Robustness Concepts for Analyzing Structured and Complex Data Sets,* Habilitationsschrift, Fachbereich Statistik, Universität Dortmund.

[4] Bentz, H. J. (1980), Das Buffon-Nadelproblem (1777), *Praxis Math.,* **22**, 167−171.

[5] Bernholt, T., Fischer, P. (2001), The Complexity of Computing the MCD-Estimator, *Technical Report 45/01, University of Dortmund.*

[6] Bolker, E. D. (1969), A Class of Convex Bodies, *Transactions of the American Mathematical Society,* **145**, 323 − 345.

[7] Blaschke, W. (1956), *Kreis und Kugel*, de Gruyter, Berlin.

[8] Davies, P. L. (1987), Asymptotic Behaviour of S-Estimates of Multivariate Location Parameters and Dispersion Matrices, *The Annals of Statistics,* **15***,* 1269 − 1292.

[9] Davies, P. L. (1992), The Asymptotics of Rousseeuw's Minimum Volume Ellipsoid Estimator, *The Annals of Statistics,* **20***,* 1828 − 1843.

[10] Davies P. L., Gather, U. (2002), Breakdown and Groups, *Preprint.*

[11] Donoho, D. L., Huber, P.J. (1983), The Notion of Breakdown Point, in: Bickel, P. J., Doksum, K. A., Hodges, J. L. (eds.), *A Festschrift for Erich Lehmann*, Wadsworth, Belmond, CA, 157 – 185.

[12] Donoho, D. L., Gasko, M. (1992), Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness, *The Annals of Statistics,* **20**, 1803 – 1827.

[13] Fraiman, R., Liu, R. Y., Meloche, J. (1997), Multivariate Density Estimation by Probing Depth, *L-1 Statistical Procedures and Related Topics,* **31**, IMS Lecture Notes - Monograph Series, 415 – 430.

[14] Grünbaum, P. (1967), *Convex Polytopes*, John Wiley & Sons, London-New York-Sydney.

[15] Leichtweiß, K. (1979), *Konvexe Mengen*, Springer-Verlag, Berlin.

[16] Liu, R. Y. (1990), On a Notion of Data Depth based on Random Simplices, *The Annals of Statistics,* **18**, 405 – 414.

[17] Liu, R. Y., Singh, K., (1993), A Quality Index based on Data Depth and Multivariate Rank Tests, *Journal of the American Statistical Association,* **99**, 257 – 260.

[18] Liu, R. Y. (1992), Data Depth and Multivariate Rank Tests, *L-1 Statistics and Related Methods* (Y. Dodge, ed.), 279 – 294, North Holland, Amsterdam.

[19] Liu, R. Y. (1995), Control Charts for Multivariate Processes, *Journal of the American Statistical Association,* **90**, 1380 – 1387.

[20] Liu, R. Y., Parelius, J. M., Singh, K. (1999), Multivariate Analysis by Data Depth: Descriptive Statistics, Graphics and Inference, *The Annals of Statistics,* **27**, 783 – 858.

[21] Lopuhaä, H. P., Rousseeuw, P. J. (1991), Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices, *The Annals of Statistics,* **19***,* 229 – 248.

[22] Koshevoy, G., Mosler, K. (1998), Lift Zonoids, Random Convex Hulls and the Variability of Random Vectors, *Bernoulli,* **4***,* 377 – 399.

[23] Koshevoy, G., Mosler, K. (1997), Zonoid Trimming for Multivariate Distributions, *The Annals of Statistics,* **9**, 1998 – 2017.

[24] Mardia, K.V., Kent J.T., Bibby, J.M. (1979), *Multivariate Analysis*, Academic Press, London.

[25] Pison, G., Van Aelst, S., Willems, G. (2002), Small sample corrections for LTS and MCD, *Metrika,* **55***,* 111 – 123.

[26] Rousseeuw, P. J. (1985), Multivariate Estimation with High Breakdown Point, in: Grossmann, W., Pflug, G., Vincze, I., Wertz, W. (eds), *Mathematical Statistics and Applications,* **8**, Reidel, Dordrecht, 283 – 297.

[27] Rousseeuw, P.J., van Zomeren B.C. (1990), Unmasking Multivariate Outliers and Leverage Points, *Journal of the American Statistical Association,* **85**, 633 – 639.

[28] Schneider, R., Weil, W. (2000), *Stochastische Geometrie*, B.G. Teubner Stuttgart, Leipzig.

[29] Schott, J. R. (1997), *Matrix Analysis for Statistics*, John Wiley & Sons, New York.

[30] Shepard, G. C. (1974), Combinatorial Properties of Associated Zonotopes, *Canad. J. Math.,* **26**, 302-321.

[31] Stoyan, D., Mecke, J. (1983), *Stochastische Geometrie*, Akademie-Verlag, Berlin.

[32] Tukey, J. W. (1975), Mathematics and the Picturing of Data, *Proceedings of the* 1975 *International Congress of Mathematics,* **2***,* 523 – 531.

[33] Tyler, D. E. (1994), Finite Sample Breakdown Points of Projection based Multivariate Location and Scatter Statistics, *The Annals of Statistics,* **22**, $1024 - 1044$.

[34] Valentine, F. A. (1964), *Convex Sets*, McGraw-Hill, New-York.

[35] Vitale, R. A. (1991), Expected absolute Random Determinants and Zonoids, *The Annals of Statistics,* **1**, $293 - 300$.

[36] Weil, W., Wieacker, J. A. (1993), Stochastic Geometry, *Handbook of Convex Geometry,* $1391 - 1438$.

[37] Zurmühl, R., Falk, S. (1986), *Matrizen und ihre Anwendungen*, 5. Aufl., Springer, Berlin.

[38] Zuo, Y., Serfling, R. (2000), General Notions of Statistical Depth Functions, *The Annals of Statistics,* **28**, $461 - 482$.

[39] Zuo, Y. (2001), Some Quantitative Relationships between Two Types of Finite Sample Breakdown Point, *Statistics and Probability Letters,* **51**, $369 - 375$.