FORSCHUNGSZENTRUM JÜLICH GmbH
Zentralinstitut für Angewandte Mathematik
D-52425 Jülich, Tel. (02461) 61-6402

Interner Bericht

Robuste statistische Verfahren für das Data Mining

Claudia Druska

FZJ-ZAM-IB-2004-04

März 2004

(letzte Änderung: 31.03.2004)

Inhaltsverzeichnis

1 Einleitung **2**

2 Maßzahlen für Lage und Streuung (univariat) **3**
2.1 Lagemaßzahlen . . . 3
2.2 Streuungsmaßzahlen . . . 5

3 Kriterien für Robustheit **6**
3.1 Die Sensitivitätskurve . . . 6
3.2 Die Einflusskurve . . . 8
3.3 Der Bruchpunkt . . . 14

4 L-Schätzer **15**
4.1 L-Schätzer für Lageparameter . . . 15
4.1.1 Das α-gestutzte Mittel . . . 15
4.1.2 Das α-winsorisierte Mittel . . . 16
4.2 L-Schätzer für die Streuung . . . 16
4.2.1 Der Interquartilabstand . . . 16
4.2.2 Der Median der absoluten Abweichung vom Median . . . 17
4.2.3 Die α-gestutzte Varianz . . . 18
4.2.4 Die α-winsorisierte Varianz . . . 18

5 M-Schätzer **19**
5.1 Theoretischer Hintergrund . . . 19
5.1.1 Grundlagen . . . 20
5.1.2 Asymptotische Eigenschaften von M-Schätzern . . . 21
5.2 M-Schätzer für Lageparameter . . . 23
5.2.1 Das arithmetische Mittel . . . 24
5.2.2 Der Median . . . 25
5.2.3 Der Huber-Schätzer . . . 25
5.2.4 Der Hampel-Schätzer . . . 26
5.2.5 Andrew's wave . . . 27
5.2.6 Tukey's biweight . . . 28
5.3 Diskussion der Eigenschaften verschiedener M-Schätzer . . . 28
5.4 Berechnung von M-Schätzern . . . 29
5.4.1 Berechnung mittels Newton-Raphson-Verfahren . . . 29
5.4.2 1-Schritt-Verfahren . . . 30
5.5 M-Schätzer für Streuungsparameter . . . 30

6 Robuste Schätzer für multivariate Lage und Streuung **33**
6.1 Einleitung und Motivation . . . 33
6.2 Separate robuste Schätzer für die Elemente der Kovarianz- bzw. Korrelationsmatrix 34

6.3 M-Schätzer für multivariate Lage- und Skalierungsparameter 35
6.4 ROBETH 38
6.4.1 Theoretische Grundlagen 38
6.4.2 Mögliche Gewichtsfunktionen 38

7 Robuste Hauptkomponentenanalyse **40**
7.1 Motivation 40
7.2 Herleitung der Hauptkomponenten 40
7.3 Robuste Hauptkomponentenanalyse 42

8 Zusammenfassung und Ausblick **45**

A Verwendete Symbole **47**

B Verwendeter Datensatz **48**
B.1 Datensatz A 49

Abbildungsverzeichnis

3.1 Sensitivitätskurve für das arithm. Mittel (links) und den Median (rechts) 7
3.2 Einflusskurve des arithmetischen Mittels für $\mu = 0$ 8
3.3 Einflusskurve des Medians für $\mu = 0$ und $\sigma^2 = 1$ 10

5.1 Arithmetische Mittel $\rho(u)$ (links) und $\psi(u)$ (rechts) 24
5.2 Median $\rho(u)$ (links) und $\psi(u)$ (rechts) 25
5.3 Huber-Schätzer $\rho(u)$ (links) und $\psi(u)$ (rechts) 26
5.4 Hampel-Schätzer $\rho(u)$ (links) und $\psi(u)$ (rechts) 27
5.5 Andrew's wave $\rho(u)$ (links), $\psi(u)$ (rechts) 27
5.6 Tukey's biweight $\rho(u)$ (links), $\psi(u)$ (rechts) 28

6.1 Streubild zweier Variablen: 109 Datenpaare (links), 106 Datenpaare (rechts) 38

7.1 Klassische Hauptkomponentenanalyse für 109 Beobachtungen 42
7.2 108 von 109 Beobachtungen zweier Variablen und beide Hauptkomponentenachsen 44

Tabellenverzeichnis

5.1 Eigenschaften der wichtigsten Schätzer 29

6.1 Daten zur Korrelation bei verschiedenen M-Schätzern 37

A.1 Übersicht der verwendeten Symbole 47

B.1 Datensatz aus Industrieprojekt 49

Zusammenfassung

Heutzutage werden immer größere Datenbestände bei der Planung neuer oder der Verbesserung bereits vorhandener Produkte in der Industrie erhoben. Eine wichtige Aufgabe besteht nun darin, die verborgenen Informationen innerhalb dieser Daten zu erkennen und auszuwerten. Der Begriff des Data Minings umfasst in diesem Zusammenhang halb automatische Methoden zur Bestimmung statistisch signifikanter Strukturen (Modelle, Muster, Unregelmäßigkeiten). Jedoch wird so gut wie jede Datenanalyse von verfälschenden Beobachtungen, so genannten Ausreißern, beeinflusst. Zur Behandlung dieses Ausreißerproblems kann man zum einen Ausreißer zu Beginn der Untersuchung identifizieren und eliminieren oder man setzt robuste statistische Verfahren ein, die den Einfluss von Ausreißern reduzieren.

Im Rahmen eines Projektes, initiiert von einem pharmazeutischen Unternehmen aus Aachen, wurden in dieser Diplomarbeit die Robustheitseigenschaften von Schätzern für univariate und multivariate Lage- und Streuungsparameter untersucht. Verschiedene robuste Verfahren aus der Klasse der L- und M-Schätzer sowie ein Algorithmus zur robusten Hauptkomponentenanalyse wurden vorgestellt und implementiert.

Abstract

Nowadays in more and more industrial sectors increasing amounts of data are collected when designing and developing innovative or improved products. An important task in this context is to uncover hidden information contained in the data. Under Data Mining semi-automatic procedures for extracting statistically significant structures such as models, patterns or anomalies from large data sets are subsumed. Almost every analysis is concerned by unrepresentative observations, so-called outliers, reducing and distorting the information about the data structure. One approach to the outlier problem is to detect and reject outliers at the beginning of the data analysis. Another approach is to use robust statistical procedures, which reduce the influence of outliers.

Within a project initiated by a pharmaceutical company from Aachen (Germany) this diploma thesis investigates the robustness properties of estimators for univariate and multivariate location and scale parameters. Several robust procedures from the classes of L-estimators and M-estimators as well as an algorithm for robust principal component analysis were examined and implemented.

Kapitel 1

Einleitung

In immer mehr Bereichen der Industrie werden heutzutage immer größere Datenbestände bei der Entwicklung neuer oder verbesserter Produkte erhoben. Dabei spielt die Erfassung und Archivierung dieses Datenmaterials aufgrund der Entwicklung von Speicherkapazitäten und leistungsstarken Rechnern nur noch eine untergeordnete Rolle. Vielmehr steht die Suche nach Informationen innerhalb dieser Daten im Vordergrund.
Mit Data Mining wird eine Untersuchungsmethode bezeichnet, aus einer Fülle von Daten und Beobachtungen mit Hilfe statistischer Verfahren die Informationen zu filtern, die für eine bestimmte Fragestellung von Bedeutung sind.
Bei der Betrachtung der Daten kann man zwei grundlegende Aufgabenbereiche unterscheiden. Zum einen sollte man Regelmäßigkeiten und Abhängigkeiten innerhalb des Datenmaterials erkennen können, zum anderen sollten sogenannte Ausreißer identifiziert und ihr Einfluss auf die gewonnenen Informationen minimiert werden.
Ausreißer können verschiedene Ursachen haben:

- Es wurden fehlerhafte Messungen durchgeführt.
- Einige wenige Beobachtungen besitzen überdurchschnittlich günstige oder ungünstige Eigenschaften.

Nun besitzen einige statistische Verfahren zum Data Mining die Eigenschaft, dass ihre Ergebnisse sehr stark von solchen Ausreißern beeinflusst werden und die so gewonnenen Informationen nur eine geringe bis keine Aussagekraft besitzen. An diesem Punkt angelangt, bieten sich zwei Lösungsmöglichkeiten an. Entweder nutzt man Verfahren zur Ausreißererkennung und eliminiert diese bevor Informationen gewonnen werden, oder es sollten Verfahren genutzt werden, die weitgehend unempfindlich gegenüber Ausreißern sind. Genau an dieser Stelle werden robuste Verfahren wirksam. Sie erlauben, Aussagen über das Datenmaterial zu treffen, die nicht von grenzwertigen Beobachtungen verfälscht werden.
Ziel der vorliegenden Arbeit ist es daher, einen Überblick über robuste Verfahren zu geben. Dabei wird auf univariate und multivariate Problemstellungen eingegangen, und es werden Kriterien zur Verfügung gestellt, die eine qualitative Beurteilung dieser Verfahren ermöglichen. Diese Arbeit erhebt nicht den Anspruch der Vollständigkeit.
Die Implementierung einiger ausgewählter robuster Verfahren erfolgte in der Programmiersprache C und findet bereits Anwendung in der Forschungsabteilung eines großen Industrieunternehmens.

Kapitel 2

Maßzahlen für Lage und Streuung (univariat)

Hat man eine Beobachtungsreihe mit Datenmaterial vorliegen, so möchte man diese Daten meist mit wenigen charakteristischen Größen kennzeichnen. Häufig verwendete Charakteristika sind die Lage und die Streuung der Daten. Eine Lagemaßzahl beschreibt in geeigneter Weise das Zentrum der beobachteten Werte $x_1, ..., x_n$, wohingegen eine Streuungsmaßzahl angibt, wie die Daten von einem Zentrum im Mittel abweichen. Einige dieser Lage- und Streuungsmaßzahlen sollen im Folgenden vorgestellt werden.

2.1 Lagemaßzahlen

Definition 2.1.1. (Arithmetisches Mittel)
Das arithmetische Mittel $\bar{x}_n$ einer Stichprobe $x_1, ..., x_n$ vom Umfang n ist definiert als

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

Aufgrund der leichten Berechenbarkeit dieser Lagemaßzahl erscheint das arithmetische Mittel als ein einfaches und leicht anzuwendendes Hilfsmittel, um das Zentrum einer Beobachtungsreihe zunächst einmal zu beschreiben. Untersucht man das arithmetische Mittel hingegen auf seine Empfindlichkeit gegenüber Ausreißern, also gegenüber Werten, die sehr weit entfernt von diesem berechneten Zentrum liegen, so stellt man schnell fest, dass das arithmetische Mittel nicht das optimale Werkzeug zur Beschreibung der Lage ist. Das folgende Beispiel aus [7] soll dies verdeutlichen.

Beispiel 2.1.2.
Von einer Holzlatte der Länge 5m werden nacheinander fünf Stücke der Länge 1m abgesägt. Anschliessend werden die Abweichungen x_i von der Solllänge 1m gemessen und in mm angegeben, d.h. $x_i = (\xi_i - 1) \cdot 10^3$, wobei ξ_i die Länge des i-ten Stückes in m angibt. Dabei ergeben sich folgende Werte für $x_1, \ldots, x_5$:

$$-0,5;\ 1,5;\ -1,0;\ 0,5;\ -9,5 \quad .$$

Berechnet man das arithmetische Mittel auf der Grundlage aller fünf Beobachtungen, so ergibt sich $\overline{x}_5 = -1,8$. Dieser Durchschnittswert legt die Vermutung nahe, dass die Säge bei jedem abgesägten Stück einen Fehler von ca. $1,8$mm macht. Die ersten vier Lattenstücke wurden jedoch annähernd korrekt mit der gewünschten Länge von 1m gesägt. Da aber bei jedem Sägevorgang 2mm durch Sägespäne verloren gingen, kann das letzte Lattenstück keine Länge von 1m mehr haben und weist die angegebene relativ große Abweichung vom Sollwert auf. Nimmt man hingegen nur die ersten

vier Werte zur Berechnung des Mittels, dann ergibt sich ein Wert für das arithmetische Mittel von $\overline{x}_4 = 0,125$. Dieser Wert zeigt offensichtlich eine adäquatere Beschreibung der Sachlage.

Bei diesem einfachen Beispiel war es unproblematisch, den Ausreißerwert zu erkennen und für weitere Berechnungen zu eliminieren. In den allermeisten Fällen ist diese Identifikation jedoch nicht so schnell und sicher zu bewerkstelligen und man sollte daher zu einer Lagemaßzahl übergehen, die nicht so empfindlich auf extreme Werte in den Rohdaten reagiert. Vorbereitend wird dazu zunächst der Begriff der geordneten Stichprobe eingeführt.

Definition 2.1.3. (Geordnete Stichprobe, Ordnungsgröße)
Hat man n Beobachtungen $x_1, \ldots, x_n$ aus einer Stichprobe vom Umfang n und werden die Daten der Größe nach sortiert, so bezeichnet $x_{(i)}$ den i-kleinsten Wert, also insbesondere

$$x_{(1)} = \min_{1 \le i \le n} x_i, \qquad x_{(n)} = \max_{1 \le i \le n} x_i \quad .$$

Die der Größe nach sortierte Reihe

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n-1)} \le x_{(n)}$$

heißt die geordnete Stichprobe von $x_1, \ldots, x_n$ und $x_{(i)}$ wird die i-te Ordnungsgröße genannt.

Damit lassen sich nun weitere Lagemaßzahlen definieren:

Definition 2.1.4. (Empirisches α-Quantil)
Das empirische α-Quantil $\tilde{x}_\alpha$ einer Stichprobe $x_1, ..., x_n$ vom Umfang n ist definiert als

$$\tilde{x}_\alpha = \begin{cases} x_{([n\cdot\alpha]+1)} & \text{, falls } n \cdot \alpha \text{ keine ganze Zahl ist} \\ \frac{1}{2}\left(x_{(n\cdot\alpha)} + x_{(n\cdot\alpha+1)}\right) & \text{, falls } n \cdot \alpha \text{ eine ganze Zahl ist} \end{cases} .$$

Als Spezialfall dieser Definition stellt sich der empirische Median dar.

Definition 2.1.5. (Empirischer Median)
Unter dem empirischen Median m_n einer Stichprobe $x_1, ..., x_n$ vom Umfang n versteht man

$$m_n := \tilde{x}_{0,5} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{, falls } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n+2}{2}\right)}\right) & \text{, falls } n \text{ gerade} \end{cases} .$$

Beispiel 2.1.6.
Betrachtet man nun noch einmal das Beispiel 2.1.2 und berechnet den empirischen Median, so erhält man als geordnete Stichprobe zunächst

$$x_{(1)} = -9,5;\ x_{(2)} = -1,0;\ x_{(3)} = -0,5;\ x_{(4)} = 0,5;\ x_{(5)} = 1,5 \quad .$$

Damit ergibt sich der Median zu $m_5 = \tilde{x}_{0,5} = -0,5$, welcher eine „robustere“ Beschreibung der Situation liefert als das arithmetische Mittel.

Bei symmetrisch verteilten Daten ist das midquartile ein geeignetes Lagemaß.

Definition 2.1.7. (midquartile)
Das midquartile einer Stichprobe $x_1, ..., x_n$ vom Umfang n ist definiert als

$$\frac{1}{2}\left(\tilde{x}_{0,25} + \tilde{x}_{0,75}\right) \quad .$$

Beispiel 2.1.8.
In Beispiel 2.1.2 ergibt sich so

$$\frac{1}{2}\left(\tilde{x}_{0,25}+\tilde{x}_{0,75}\right)=\frac{1}{2}(x_{(2)}+x_{(4)})=-0,25 \quad .$$

Diese kleine Auswahl an möglichen Lagemaßzahlen zeigt schon am simplen Beispiel, wie unterschiedlich auf Ausreißer reagiert wird, und es wird deutlich, dass man bei der Untersuchung experimentell erzeugter oder auch pseudo-zufällig erzeugter Daten am Rechner auf Verfahren zurückgreifen sollte, die möglichst unempfindlich gegenüber Ausreißern sind.

2.2 Streuungsmaßzahlen

Das wohl bekannteste Verfahren zur Bestimmung der Streuung innerhalb einer Stichprobe ist die empirische Standardabweichung.

Definition 2.2.1. (Empirische Standardabweichung, empirische Varianz)
Die empirische Standardabweichung einer Beobachtungsreihe $x_1, ..., x_n$ vom Stichprobenumfang n ist definiert als

$$s_n=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}_n\right)^2} \quad .$$

Die quadrierte empirische Standardabweichung wird empirische Varianz genannt.

Beispiel 2.2.2.
Angewendet auf die Daten aus Beispiel 2.1.2 erhält man hier für den arithmetischen Mittelwert $\overline{x}_5=-1,8$ und eine empirische Standardabweichung von $s_5=4,41$, bei $\overline{x}_4=0,125$ ergibt sich ein Wert von $s_4=1,11$.
Die Werte des Beispiels zeigen, dass auch die empirische Standardabweichung empfindlich auf Ausreißerwerte reagiert. Dies lässt sich damit begründen, dass große, stark vom Zentrum abweichende Werte durch das Quadrieren bei der Berechnung der empirischen Standardabweichung einen noch größeren Einfluss gewinnen und so die Streuung der Daten verfälscht wiedergeben. Auch die Verwendung einer ausreißerunempfindlicheren Lagemaßzahl kann dies nicht unterbinden, da das arithmetische Mittel $\bar{x}_n$ gerade der Wert ist, der den Ausdruck $\sum(x_i-a)^2$ für a minimiert.

Eine gegenüber Ausreißern robustere Streuungsmaßzahl ist der Quartilabstand.

Definition 2.2.3. (Quartilabstand)
Als Quartilabstand (Interquartilabstand, interquartile range) einer Stichprobe $x_1, \ldots, x_n$ bezeichnet man die Differenz der beiden Größen $\tilde{x}_{0,75}$ und $\tilde{x}_{0,25}$, also

$$\mathrm{IQR}_n=\tilde{x}_{0,75}-\tilde{x}_{0,25} \quad .$$

Auf den Begriff des Interquartilabstandes wird in Kapitel 4 noch näher eingegangen.
Für die Stichprobe des Beispiels 2.1.2 ergibt sich $\mathrm{IQR}_5=1,5$. Dieser Wert zeigt einen robusteren Umgang mit dem Ausreißerwert von $x_{(1)}=-9,5$ und ist daher hier besser geeignet, die Streuung der Daten um ein Lagezentrum zu beschreiben als der der empirischen Standardabweichung.

Kapitel 3

Kriterien für Robustheit

Wie man in dem vorangegangenen Kapitel an einfachen Beispielen bereits festgestellt hat, reagieren unterschiedliche Verfahren zur Bestimmung von Lage und Streuung unterschiedlich auf Ausreißer. Was man in diesem Zusammenhang unter einem „robusten“ Verfahren versteht, ist intuitiv klar, aber wie kann man diese Robustheit mathematisch beschreiben und überprüfen? Dieser Frage soll im vorliegenden Kapitel nachgegangen werden. Als Kriterien zur Beurteilung von Robustheit werden an dieser Stelle die Sensitivitätskurve, die Einflusskurve und der Bruchpunkt benutzt.
Bezeichne im Weiteren

$$T_n = T_n(x_1, \ldots, x_n)$$

eine Stichprobenfunktion, auch Statistik genannt, zur Stichprobe $x_1, \ldots, x_n$ vom Umfang n.

Beispiel 3.0.1. (Stichprobenfunktionen)

1. Arithmetisches Mittel:
$$T_n(x_1, \ldots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$
2. Median:
$$T_n(x_1, \ldots, x_n) = \begin{cases} x_{((n+1)/2)} & \text{, falls } n \text{ ungerade} \\ \frac{1}{2}(x_{(n/2)} + x_{((n+2)/2)}) & \text{, falls } n \text{ gerade} \end{cases}$$
3. Standardabweichung:
$$T_n(x_1, \ldots, x_n) = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x}_n)^2}$$

3.1 Die Sensitivitätskurve

Definition 3.1.1. (Sensitivitätskurve, sensitivity curve)
Sei $T_n = T_n(x_1, \ldots, x_n)$ eine Stichprobenfunktion basierend auf n Beobachtungen. Dann ist die Sensitivitätskurve definiert als

$$\mathrm{SC}(x; x_1, \ldots, x_{n-1}, T_n) = n \cdot (T_n(x_1, \ldots, x_{n-1}, x) - T_{n-1}(x_1, \ldots, x_{n-1})) \quad .$$

Die Sensitivitätskurve beschreibt die Reaktion einer Stichprobenfunktion, falls zu einer Stichprobe $x_1, \ldots, x_{n-1}$ eine weitere Beobachtung x hinzugefügt wird. Ihr Wert entspricht der Differenz von T_n und T_{n-1} multipliziert mit der Anzahl n der Beobachtungen insgesamt.

Beispiel 3.1.2. (Sensitivitätskurve von arithmetischem Mittel und Median)([7])
Es seien folgende Daten gegeben:

$$x_1 = -0,5;\ x_2 = 1,5;\ x_3 = -1,0;\ x_4 = 0,5 \quad \Rightarrow n - 1 = 4$$

1. Sensitivitätskurve des arithmetischen Mittels

$$\begin{aligned} \mathrm{SC}(x; x_1, \ldots, x_4, T_5) &= 5 \cdot \left(\frac{1}{5}(x_1 + \ldots + x_4 + x) - \frac{1}{4}(x_1 + \ldots + x_4) \right) \\ &= 5 \cdot (0,1 + \frac{1}{5}x - 0,125) \\ &= x - 0,125 \end{aligned}$$

2. Sensitivitätskurve des Medians

$$\mathrm{SC}(x; x_1, \ldots, x_4, T_5) = \begin{cases} -2,5 & \text{, falls } x < -0,5 \\ 5x & \text{, falls } -0,5 \leq x \leq 0,5 \\ 2,5 & \text{, falls } 0,5 < x \end{cases}$$

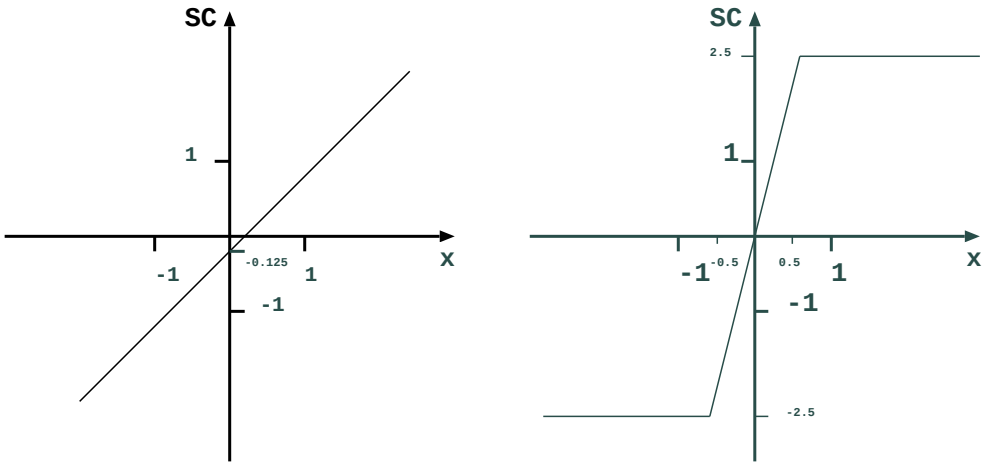

Abbildung 3.1: Sensitivitätskurve für das arithm. Mittel (links) und den Median (rechts)

Im letzten Beispiel zeigt sich deutlich am Steigungsverhalten der Kurven, dass das arithmetische Mittel bei Hinzunahme eines weiteren beliebigen Wertes x die Sensitivitätskurve und damit auch den Wert für das Zentrum der Datenpunkte so verändern kann, dass der neu berechnete Wert für die Lage der Daten ein stark verfälschtes Bild der Stichprobe liefert. Dies zeigt sich darin, dass die Sensitivitätskurve des arithmetischen Mittels unbeschränkt ist. Im Gegensatz dazu ist die Sensitivitätskurve des Medians beschränkt, so dass auch ein zusätzlicher weit ausserhalb des Zentrums liegender Beobachtungswert die geschätzte Lage des Zentrums nicht maßgeblich verändern kann.

Bevor nun weitere Kriterien zur Beurteilung von Robustheitseigenschaften von Maßzahlen eingeführt werden, wird die bisherige Betrachtungsweise aus der Sicht der rein beschreibenden (deskriptiven) Statistik verlassen und ein für das weitere Vorgehen zugrundeliegendes mathematisches Modell vorgestellt [12]. Dabei bezeichne X eine Zufallsvariable (Zufallsgröße), die auf einem geeigneten Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$ als eine Abbildung von Ω in die Menge der reellen Zahlen definiert ist, und x bezeichnet im Folgenden eine Realisation einer Zufallsvariablen X. Des Weiteren sei

$$F : \mathbb{R} \to [0; 1] \quad \text{mit} \quad F(x) := \mathbb{P}(X \leq x) \qquad \forall x \in \mathbb{R}$$

die Verteilungsfunktion der Zufallsvariablen X mit den bekannten Eigenschaften.

Um sich bei unbekannter Verteilungsfunktion trotzdem ein Bild von der Verteilung der Zufallsgröße X machen zu können, ist die Konstruktion der empirischen Verteilungsfunktion F_n hilfreich. Für eine unabhängige Folge $X_1, \ldots, X_n$ von Zufallsgrößen mit der Verteilungsfunktion F gibt

$$F_n(t, \omega) := \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,t]} \left(X_i(\omega) \right)$$

die relative Häufigkeit der X_i mit den Werten $\leq t$ an, wobei mit $1_{(-\infty,t]}$ die Indikatorfunktion [1] auf dem Intervall $(-\infty, t]$ bezeichnet wird. Außerdem heißt eine Stichprobenfunktion, deren Wert einen unbekannten Parameter der den Beobachtungen zugrunde liegenden Verteilung schätzt, ein Schätzer bzw. eine Schätzfunktion. Alle bisher betrachteten Stichprobenfunktionen sind Schätzer von Lage- bzw. Streuungsparametern der zugrunde liegenden Verteilung.

3.2 Die Einflusskurve

Bei der Berechnung der Sensitivitätskurve stellt man fest, dass diese nicht ausschließlich von der jeweiligen Stichprobenfunktion $T_n(\cdot)$ abhängt, sondern vor allem auch von den konkret vorliegenden Stichprobendaten und deren Anzahl. Man kann versuchen, diese Abhängigkeit zu eliminieren, indem man eine Grenzwertbetrachtung für $n \to \infty$ durchführt. Dazu betrachtet man zum Beispiel die SC-Kurve des arithmetischen Mittels:

$$\begin{aligned} \mathrm{SC}(x; x_1, \ldots, x_{n-1}, \bar{x}_n) &= n \cdot \left[\frac{1}{n} \left(\sum_{i=1}^{n-1} x_i + x \right) - \frac{1}{n-1} \sum_{i=1}^{n-1} x_i \right] \\ &= \sum_{i=1}^{n-1} x_i + x - \frac{n}{n-1} \sum_{i=1}^{n-1} x_i \\ &= -\frac{1}{n-1} \sum_{i=1}^{n-1} x_i + x \end{aligned}$$

Sind die Beobachtungen unabhängige Realisationen einer Zufallsgröße mit existierendem Erwartungswert μ, dann konvergiert aufgrund des Starken Gesetzes Großer Zahlen (SGGZ) die Sensitivitätskurve SC fast sicher [2] gegen $x - \mu$.

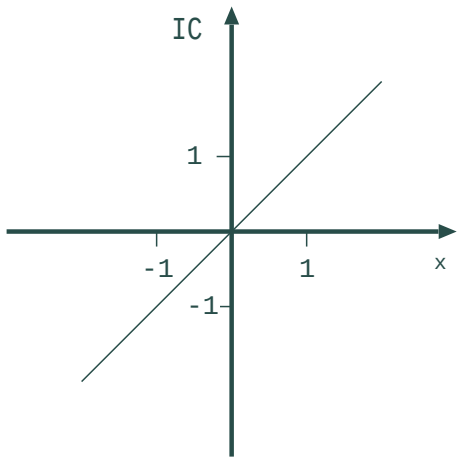

Abbildung 3.2: Einflusskurve des arithmetischen Mittels für $\mu = 0$

[1] $1_A(\omega) = \begin{cases} 1 \text{ , falls } \omega \in A \\ 0 \text{ , falls } \omega \notin A \end{cases}$

[2] Der Begriff der fast sicheren Konvergenz wird in 5.1.3 definiert

Betrachtet man jetzt allgemein eine Stichprobenfunktion, die wie das arithmetische Mittel eine Funktion der empirischen Verteilungsfunktion F_n ist, $T_n = T(F_n)$, dann ergibt sich für die SC-Kurve:

$$\begin{aligned} \mathrm{SC}(x; x_1, \ldots, x_{n-1}, T_{n-1}) &= n\left[T\left(\frac{n-1}{n}F_{n-1} + \frac{1}{n}\delta_x\right) - T(F_{n-1})\right] \\ &= \frac{T\left(\frac{n-1}{n}F_{n-1} + \frac{1}{n}\delta_x\right) - T(F_{n-1})}{\frac{1}{n}} \quad , \end{aligned}$$

wobei δ_x die sogenannte Einheitsmasse im Punkt x ist. Für $n \to \infty$ strebt $F_{n-1} \to F$, und man erhält für ein hinreichend reguläres T, von dem man voraussetzt, dass es auch in F definiert ist, in vielen Fällen (s. [6]):

$$\mathrm{SC}(x; X_1, \ldots, X_{n-1}, T_{n-1}) \xrightarrow{f.s.} \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon} \quad . \tag{3.1}$$

In diesem Fall ist $T(F_n)$ ein naturlicher Schätzwert für $T(F)$.

Definition 3.2.1. (Einflusskurve, influence function)
Der Grenzwert (3.1) wird als die Einflusskurve (influence curve) bezeichnet, d.h.

$$\mathrm{IC}(x; F, T) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)F + \varepsilon\delta_x) - T(F)}{\varepsilon} \quad .$$

Die Einflusskurve gibt also Auskunft über die Reaktion einer Schätzfunktion auf einen hinzugefügten Wert unter der Annahme, dass die Stichprobe aus der zugrundeliegenden Verteilung von großem Umfang ist. Da bei der Grenzwertbildung wegen des Starken Gesetzes Großer Zahlen die Stichprobe nicht mehr betrachtet wird, hängt die Einflusskurve nur noch von x ab und wird daher auch als lokale Kenngröße bezeichnet. Ausserdem kann man den maximalen Einfluss, den eine einzelne Beobachtung auf den zu schätzenden Wert ausüben kann, mit dem maximalen Betrag der Einflusskurve angeben. Diese Größe wird als Ausreißerempfindlichkeit (gross-error sensitivity) bezeichnet. Untersucht man das arithmetische Mittel und den Median hinsichtlich dieser Größe, so stellt man fest, dass das arithmetische Mittel im Gegensatz zum Median einen unendlichen Wert annnehmen kann und deshalb bei ausreißerbehafteten Daten nicht geeignet ist. Auch die Steigung der Einflusskurve kann man im Bezug auf die Robustheit interpretieren. Sie liefert einen Anhaltspunkt für das Verhalten der Schätzfunktion bei nur geringen Änderungen, wie sie zum Beispiel bei Rundungen oder Klassenbildung der Rohdaten auftreten können. Solche Änderungen haben einen stärkeren Einfluss auf den zu schätzenden Wert je steiler die Steigung der Einflusskurve ist. Je größer der maximale Wert des Absolutbetrages der Steigung (local-shift sensitivity) ist, desto stärker können solche Änderungen auf den Schätzwert durchschlagen. Für den Median ergibt sich hier ein unendlicher Wert, während die Größe beim arithmetischen Mittel den endlichen Wert 1 annimmt. Wie in Kapitel 5 noch genauer erläutert wird, gibt es auch robuste Schätzer, die weit außerhalb liegende Stichprobenwerte vollständig ignorieren, d.h. die Einflusskurve hat ab einem bestimmten Abstand vom zu schätzenden Lageparameter den Wert identisch Null (s. die Beispiele in Kapitel 5.2).

Beispiel 3.2.2. Einflusskurve des Medians (s. [7])
Es wird angenommen, dass eine Normalverteilung $\mathcal{N}(\mu, \sigma^2)$ mit Mittelwert μ und Varianz σ^2 vorliegt. Dann gilt

$$\begin{aligned} \mathrm{IC}(x; F, m_n) &= \frac{1}{2}\sqrt{2 \cdot \pi} \cdot \sigma \cdot \mathrm{sgn}(x - \mu) \\ &= \begin{cases} -\sqrt{\frac{\pi}{2}} \cdot \sigma & \text{, falls} \quad x - \mu < 0 \\ 0 & \text{, falls} \quad x - \mu = 0 \\ \sqrt{\frac{\pi}{2}} \cdot \sigma & \text{, falls} \quad x - \mu > 0 \end{cases} \end{aligned}$$

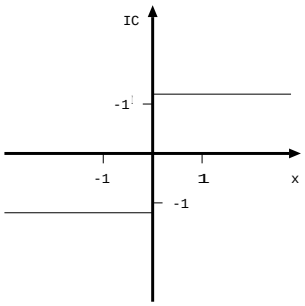

Abbildung 3.3: Einflusskurve des Medians für $\mu = 0$ und $\sigma^2 = 1$

Der Grenzwert, der die IC-Kurve definiert, kann als Ableitung des Funktionals T interpretiert werden. Man bezeichnet diese Ableitung als Gâteaux-Differential, das im Folgenden vorgestellt wird.

Definition 3.2.3. (Gâteaux-Differential)
Es seien F und G Verteilungen aus der Menge $\mathcal{F}$ aller Verteilungsfunktionen und

$$\{(1-\lambda)F + \lambda G,\ 0 \leq \lambda \leq 1\}$$

sei die Menge der Konvexkombinationen aus F und G. Das Funktional T sei definiert auf $F + \lambda(G - F)$ für alle hinreichend kleinen λ. Existiert der Grenzwert

$$d_1T(F; G - F) = \lim_{\lambda \to 0+} \frac{T(F + \lambda(G - F)) - T(F)}{\lambda} \quad ,$$

so wird dieser als das Gâteaux-Differential des Funktionals T an der Stelle F in Richtung von G bezeichnet.

Man erkennt, dass der Ausdruck $d_1T(F; G - F)$ die rechtsseitige Ableitung der reellen Funktion $Q(\lambda) = T(F + \lambda(G - F))$ nach λ an der Stelle $\lambda = 0$ ist.

Als nächstes wird die Abweichung der Schätzung $T(F_n)$ von dem wahren Wert $T(F)$ mit Hilfe des Gâteaux-Differentials von T an der Stelle F in Richtung F_n untersucht. Es sei

$$Q(\lambda) = T(F + \lambda(F_n - F)) \quad .$$

Die lineare Taylorapproximation von Q an der Stelle $\lambda = 0$ ergibt

$$\begin{aligned} T(F_n) - T(F) &= Q(1) - Q(0) \\ &\approx Q'(0+) \\ &= d_1T(F; F_n - F) \quad . \end{aligned}$$

Ist das Gâteaux-Differential im zweiten Argument linear, erhält man weiter

$$T(F_n) - T(F) \approx \frac{1}{n}\sum_{i=1}^{n} d_1T(F; \delta_{x_i} - F) \tag{3.2}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}(x_i; F, T) \tag{3.3}$$

Dieses Ergebnis liefert eine weitere Begründung für die Bezeichnung Einflusskurve. $\mathrm{IC}(\cdot)$ approximiert den Beitrag der Beobachtung x_i zu dem Schätzfehler $T(F_n) - T(F)$. Darüber hinaus zeigt

sich aus der Darstellung des Fehlers in (3.2), dass dieser näherungsweise als arithmetisches Mittel der IC-Kurve verstanden werden kann, das aufgrund des Zentralen Grenzwertsatzes bei existierender positiver Varianz von $\mathrm{IC}(X; F, T)$ asymptotisch normalverteilt ist. Damit kann man schliessen, dass der Fehler, d.h. die linke Seite der Gleichung (3.2) unter bestimmten Voraussetzungen (s. [16], S. 225f) asymptotisch normalverteilt ist:

Satz 3.2.4.
Sind die Bedingungen

- $\mathrm{Var}(IC(X; F, T)) > 0$ und
- $\sqrt{n}\Big((T(F_n) - T(F)) - \frac{1}{n}\sum_{i=1}^{n} \mathrm{IC}(X; F, T)\Big) \xrightarrow{st.} 0$

erfüllt [3], so gilt

$$T(F_n) - T(F) \overset{as.}{\sim} \mathcal{N}\left(\mathbb{E}(\mathrm{IC}(X; F, T)),\ \frac{1}{n}\mathrm{Var}(\mathrm{IC}(X; F, T))\right) \quad ,$$

wobei die Bezeichnung $\overset{as.}{\sim} \mathcal{N}(\ldots)$ bedeutet, dass der Ausdruck

$$\frac{(T(F_n) - T(F)) - (\mathbb{E}(\mathrm{IC}(X; F, T)))}{\sqrt{\frac{1}{n}\mathrm{Var}(\mathrm{IC}(X; F, T))}}$$

asymptotisch standardnormalverteilt ist, d.h. die Folge der $(T(F_n) - T(F))$ konvergiert der Verteilung nach gegen $\mathcal{N}(0; 1)$ (vgl. Definition 5.1.4).

Bemerkung 3.2.5.
Im Allgemeinen liegt asymptotische Erwartungstreue vor, d.h. der Erwartungswert der Einflusskurve ist gleich Null und für die Varianz gilt

$$\mathrm{Var}(\mathrm{IC}(X; F, T)) = \mathbb{E}(\mathrm{IC}(X; F, T)^2)$$

Bemerkung 3.2.6.
In [16] wird für viele Beispielfälle gezeigt, dass die Bedingungen aus Satz 3.2.4 erfüllt sind.

Abschliessend zwei Beispiele zum Begriff des Gâteaux-Differentials:

Beispiel 3.2.7. Zentrale Stichprobenmomente
Das k-te zentrale Moment einer Verteilung F kann mit Hilfe von

$$\mu_k = T(F) = \int_{-\infty}^{\infty} \left[x - \int_{-\infty}^{\infty} y\, dF(y)\right]^k dF(x)$$

als Funktional geschrieben werden und das k-te zentrale Stichprobenmoment lautet dann

$$m_k = T(F_n) = \int_{-\infty}^{\infty} (x - \bar{x}_n)^k\, dF_n(x) \quad .$$

Setzt man $\mu_F = \int x\, dF(x)$ und $F_\lambda = F + \lambda(G - F)$, dann gilt

$$\mu_{F_\lambda} = \mu_F + \lambda(\mu_G - \mu_F)$$

und man erhält

$$T(F_\lambda) = \int (x - \mu_{F_\lambda})^k\, dF(x) + \lambda \int (x - \mu_{F_\lambda})^k\, d[G(x) - F(x)] \quad .$$

[3] Der Begriff der stochastischen Konvergenz wird in 5.1.2 definiert.

Differenziert man diesen Ausdruck nach λ, so ergibt sich

$$\begin{aligned}\frac{dT(F_\lambda)}{d\lambda} &= -k(\mu_G - \mu_F)\int (x-\mu_{F_\lambda})^{k-1}\, dF_\lambda(x) \\ &+ \int (x-\mu_{F_\lambda})^k\, d[G(x)-F(x)] \quad .\end{aligned}$$

Da nach [16]

$$d_1T(F;G-F) = \frac{d}{d\lambda}T(F+\lambda(G-F))\bigg|_{\lambda=0+} \quad ,$$

ergibt sich mit $\mu = \mu_F = \mu_{F_0}$ für das Gâteaux-Differential

$$d_1T(F;G-F) = \int \left[(x-\mu)^k - k\mu_{k-1}x\right]\, d[G(x)-F(x)] \quad .$$

Für $G = \delta_x$ folgt daraus die Einflusskurve

$$\begin{aligned}\mathrm{IC}(x;F,T) &= \frac{dT[F+\lambda(\delta_x-F)]}{d\lambda}\bigg|_{\lambda=0+} \\ &= (x-\mu)^k - k\mu_{k-1}x - \mathbb{E}[(X-\mu)^k - k\mu_{k-1}X]\end{aligned}$$

Beispiel 3.2.8. Maximum-Likelihood-Schätzung
Seien $X_1,\ldots,X_n$ unabhängig und identisch verteilte Zufallsgrößen mit einer Verteilung F_θ, die zu einer Familie $\{F_\theta, \theta \in \Theta \subset \mathbb{R}\}$ gehört. Außerdem sei vorausgesetzt, dass die Verteilung F_θ eine Dichtefunktion $f(x;\theta)$ besitze. Die Likelihood-Funktion der Stichprobe $x_1,\ldots,x_n$ ist definiert als

$$L(t;x_1,\ldots,x_n) = \prod_{i=1}^n f(x_i;t) \quad .$$

Die Maximum-Likelihood-Methode bietet als Schätzer für das unbekannte t jeden Wert $\hat{\theta}$ an, der die Likelihood-Funktion über Θ maximiert. Dabei stellt es eine Erleichterung dar, wenn man zum Negativen der logarithmierten Likelihood-Funktion

$$-\log\left(L(t;x_1,\ldots,x_n)\right) = -\sum_{i=1}^n \log\left(f(x_i;t)\right)$$

übergeht und so das Produkt durch eine Summe ersetzen kann. Durch den Vorzeichenwechsel ist nun ein Minimum zu bestimmen. Dazu wird eine Nullstelle der ersten Ableitung gesucht, d.h. die Gleichung

$$\frac{\partial \log L}{\partial t}\bigg|_{t=\hat{\theta}} = 0 \tag{3.4}$$

ist zu lösen. (3.4) entspricht der Integralgleichung

$$\int g(x,t)\, dF_n(x) = 0$$

mit

$$g(x,t) = \frac{d}{dt}\log\left(f(x;t)\right) \quad ,$$

d.h.

$$g(x,t) = \frac{\frac{d}{dt}f(x;t)}{f(x;t)} \quad ,$$

wobei

$$f(x;t) = \frac{d}{dx} F_t(x)$$

Damit ist der Maximum-Likelihood-Schätzer $\hat{\theta} = T(F_n)$, wobei $T(F)$ das Funktional ist, das durch die Lösung von

$$\int g(x,t)\ dF(x) = 0 \tag{3.5}$$

definiert wird. Unter gewissen Regularitätsbedingungen (s. [16], S. 144) an $\{F_\theta, \theta \in \Theta\}$ ergibt sich

$$\int \frac{d}{dt} \log\left(f(x;t)\right)\Big|_{t=\theta}\ dF_\theta(x) = 0 \tag{3.6}$$

und

$$\int \frac{\frac{d^2}{dt^2} f(x;t)}{f(x;t)}\Big|_{t=\theta}\ dF_\theta(x) = \int \frac{d^2}{dt^2} f(x;t)\Big|_{t=\theta}\ dx = 0 \quad . \tag{3.7}$$

Aus (3.6) folgt unmittelbar

$$\int g(x,\theta)\ dF_\theta(x) = 0 \quad ,$$

d.h.

$$T(F_\theta) = \theta \quad .$$

Sei $F_\lambda = F_\theta + \lambda(\delta_{x_0} - F_\theta)$ und $H(t,\lambda) = \int g(x,t)\ dF_\lambda(x)$, dann erhält man durch implizites Differenzieren der Gleichung

$$H(T(F_\lambda), \lambda) = 0$$

nach λ an der Stelle $\lambda = 0$ die Gleichung

$$\frac{\partial H(t,\lambda)}{\partial t}\Big|_{t=\theta,\lambda=0} \cdot \frac{dT(F_\lambda)}{d\lambda}\Big|_{\lambda=0} + \frac{\partial H(t,\lambda)}{\partial \lambda}\Big|_{t=\theta,\lambda=0} = 0$$

und daraus ergibt sich

$$\frac{dT(F_\lambda)}{d\lambda}\Big|_{\lambda=0} = -\frac{\partial H(t,\lambda)}{\partial \lambda}\Big|_{t=\theta,\lambda=0} \Big/ \frac{\partial H(t,\lambda)}{\partial t}\Big|_{t=\theta,\lambda=0} \quad . \tag{3.8}$$

Setzt man nun die bis dahin gewonnenen Ergebnisse ein, so erhält man

$$\begin{aligned}
\frac{\partial H(t,\lambda)}{\partial \lambda}\Big|_{t=\theta,\lambda=0} &= \frac{\partial}{\partial \lambda}\left[\int g(x,t)\ dF_\theta(x) + \lambda \int g(x,t)\ d[\delta_{x_0} - F_\theta(x)]\right]\Big|_{t=\theta,\lambda=0} \\
&= \int g(x,\theta)\ d\delta_{x_0} - \underbrace{\int g(x,\theta)\ dF_\theta(x)}_{\overset{(3.6)}{=} 0} \\
&= \frac{\frac{d}{dt} f(x_0;t)}{f(x_0;t)}\Big|_{t=\theta}
\end{aligned}$$

und

$$\begin{aligned}
\left.\frac{\partial H(t,\lambda)}{\partial t}\right|_{t=\theta,\lambda=0} &= \int \left.\frac{d}{dt} g(x,t)\right|_{t=\theta} dF_\theta(x) \\
&= \int \left.\frac{d}{dt}\frac{\frac{d}{dt}f(x;t)}{f(x;t)}\right|_{t=\theta} dF_\theta(x) \\
&= \int \frac{\left.\frac{d^2}{dt^2}f(x;t)\right|_{t=\theta} \cdot f(x;t)|_{t=\theta} - \left(\left.\frac{d}{dt}f(x;t)\right|_{t=\theta}\right)^2}{f(x;\theta)^2}\, dF_\theta(x) \\
&= \underbrace{\int \frac{\left.\frac{d^2}{dt^2}f(x;t)\right|_{t=\theta}}{f(x;\theta)}\, dF_\theta(x)}_{\overset{(3.7)}{=}0} - \int \left(\frac{d}{dt}\log\left(f(x;t)\right)\right)^2 \Bigg|_{t=\theta} dF_\theta(x) \\
&= -\int \left(\frac{d}{dt}\log\left(f(x;t)\right)\right)^2 \Bigg|_{t=\theta} dF_\theta(x)
\end{aligned}$$

Damit folgt für das Gâteaux-Differential

$$\begin{aligned}
d_1T(F_\theta;\delta_{x_0}-F_\theta) &= \left.\frac{d}{d\lambda}T(F_\theta+\lambda(\delta_{x_0}-F_\theta))\right|_{\lambda=0} \\
&= \frac{\frac{d}{d\theta}f(x_0;\theta)}{f(x_0;\theta)} \cdot \frac{1}{\int \frac{\left(\frac{d}{d\theta}f(x;\theta)\right)^2}{f(x;\theta)}\, dx}
\end{aligned}$$

und für die Einflusskurve

$$\mathrm{IC}(x;F_\theta,T) = \frac{\frac{d}{d\theta}\log\left(f(x;\theta)\right)}{\int \left(\frac{d}{d\theta}\log\left(f(x;\theta)\right)\right)^2 dF_\theta(x)} \quad .$$

3.3 Der Bruchpunkt

Definition 3.3.1. (Bruchpunkt)
Der Bruchpunkt (breakdown point) ε^* gibt diejenige Grenze an, bis zu welcher der Anteil von Ausreißern in einer Stichprobe ansteigen darf, ohne dass sich dadurch der Schätzwert unbeschränkt verändern kann.

Im Gegensatz zur Einflusskurve ist der Bruchpunkt eine globale Kenngröße der Robustheit.

Beispiel 3.3.2. Bruchpunkt von arithmetischem Mittel und Median

(a) Arithmetisches Mittel
Beim arithmetischen Mittel kann bereits ein einzelner Wert innerhalb einer Beobachtungsreihe den Schätzwert für das Zentrum der Stichprobe über alle Grenzen wachsen lassen, daher gilt

$$\varepsilon^*_{\text{arith. Mittel}} = 0 \quad .$$

(b) Median
Liegt der Anteil der Ausreißer unter 50%, so kann keine beliebige Veränderung des Schätzwertes vorkommen. Der Median hat daher als Bruchpunkt den Wert

$$\varepsilon^*_{\text{Median}} = 0,5 \quad .$$

Kapitel 4

L-Schätzer

Die in Kapitel 2 vorgestellten Schätzer für Lage- und Streuungsmaßzahlen beruhen alle auf Linearkombinationen der Ordnungsstatistiken $x_{(1)}, \ldots, x_{(n)}$ einer Stichprobe vom Umfang n. Daher werden sie L-Schätzer genannt.

Definition 4.0.1. (L-Schätzer)
Seien $x_{(1)}, \ldots, x_{(n)}$ die Ordnungsstatistiken einer Stichprobe vom Umfang n und seien $a_1, \ldots, a_n$ reelle Zahlen mit $0 \leq a_i \leq 1, i = 1, \ldots, n$, so dass $\sum_{i=1}^{n} a_i = 1$. Dann ist

$$T = \sum_{i=1}^{n} a_i \, x_{(i)}$$

ein L-Schätzer mit Gewichten $a_1, \ldots, a_n$.

Das arithmetische Mittel ist also ebenso ein L-Schätzer wie der Median oder der Interquartilabstand. Die Bezeichnung „L-Schätzer“ sagt also noch nichts über die Eigenschaft der Robustheit aus. Ein Vorteil der L-Schätzer liegt in der leichten Berechenbarkeit, die in der Regel lediglich eine Vorsortierung der Stichprobenelemente erfordert.

4.1 L-Schätzer für Lageparameter

4.1.1 Das α-gestutzte Mittel

Es sei vorausgesetzt, dass die zugrundeliegende Verteilung symmetrisch und stetig ist, d.h. sie besitzt eine Dichte. Beim α-gestutzten Mittel als Schätzer für den Median bzw., sofern er existiert, den Erwartungswert der zugrunde liegenden Verteilung werden aus der geordneten Stichprobe am unteren und oberen Ende eine gleiche Anzahl von Beobachtungen weggelassen und über die verbliebenen Werte das arithmetische Mittel berechnet. Der Anteil der zu entfernenden Werte wird mit α angegeben. Unter dem α-gestutzten Mittel versteht man daher den Wert

$$\overline{x}_\alpha = \frac{1}{n-2h} \sum_{i=h+1}^{n-h} x_{(i)} \quad ,$$

wobei $0 < \alpha < 0,5$ ist und $h = [n\alpha]$ die größte ganze Zahl kleiner oder gleich $n\alpha$ bezeichnet. Dieser Ansatz stellt damit einen Kompromiss zwischen dem arithmetischen Mittel und dem Median dar, die sich für die Werte $\alpha \to 0$ bzw. $\alpha \to 0,5$ ergeben würden.
Eine asymptotisch äquivalente Version des α-getrimmten Mittels ist durch

$$\overline{x}_\alpha = T(F_n)$$

gegeben, wobei für eine beliebige Verteilungsfunktion F

$$\begin{aligned} T(F) &= \frac{1}{1-2\alpha}\int_{F^{-1}(\alpha)}^{F^{-1}(1-\alpha)} x\, dF(x) \\ &= \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha} F^{-1}(p)dp \end{aligned}$$

gilt mit $F^{-1}(\alpha) = \inf\{x|F(x) \geq \alpha\}$. Zur Berechnung der Einflusskurve ist zunächst nach [16] für eine stetige Verteilung

$$\begin{aligned} \left.\frac{dT[F+\lambda(\delta_{x_0}-F)]}{d\lambda}\right|_{\lambda=0} &= \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha}\frac{p - I(x_0 \leq F^{-1}(p))}{f\left(F^{-1}(p)\right)}\, dp \\ &= \frac{1}{1-2\alpha}\int_{\alpha}^{1-\alpha}\frac{p - I(F(x_0) \leq p)}{f\left(F^{-1}(p)\right)}\, dp \quad , \end{aligned}$$

wobei $I(\cdot)$ die Indikatorfunktion bezeichnet. Nun muss eine Fallunterscheidung vorgenommen werden für die Fälle $F(x_0) < \alpha, \quad F(x_0) > 1-\alpha$ und $\alpha \leq F(x_0) \leq 1-\alpha$. Schließlich erhält man

$$\left.\frac{dT[F+\lambda(\delta_{x_0}-F)]}{d\lambda}\right|_{\lambda=0} = \begin{cases} \frac{1}{1-2\alpha}[F^{-1}(\alpha) - F^{-1}(1/2)] & \text{, falls} \quad x < F^{-1}(\alpha) \\ \frac{1}{1-2\alpha}(x - F^{-1}(1/2)) & \text{, falls} \quad F^{-1}(\alpha) \leq x \leq F^{-1}(1-\alpha) \\ \frac{1}{1-2\alpha}[F^{-1}(1-\alpha) - F^{-1}(1/2)] & \text{, falls} \quad x > F^{-1}(1-\alpha) \quad . \end{cases}$$

4.1.2 Das α-winsorisierte Mittel

Beim α-winsorisierten Mittelwert wird im Gegensatz zum gestutzten Mittelwert die Anzahl der Beobachtungen beibehalten. Jedoch werden die ersten h Beobachtungen durch den $(h+1)$ten Wert ersetzt und die h letzten Werte durch den $(n-h)$ten Wert, dann wird das arithmetische Mittel gebildet, d.h.

$$\bar{x}_{\alpha,win} = \frac{1}{n}\left(\sum_{i=h+1}^{n-h} x_{(i)} + h(x_{(h+1)} + x_{(n-h)})\right) \quad .$$

4.2 L-Schätzer für die Streuung

4.2.1 Der Interquartilabstand

Der Begriff des Interquartilabstandes wurde bereits in Kapitel 2 definiert. Er dient an dieser Stelle der Motivation des robusten Skalenschätzers MAD, der im nächsten Unterabschnitt eingeführt wird. Der Bruchpunkt des Interquartilabstandes beträgt $\varepsilon^*_{\text{IQR}} = 0,25$.

Beispiel 4.2.1. Interquartilabstand
Mit den bekannten Werten aus Beispiel 2.1.2 erhält man wie in Kapitel 2 bereits angegeben

$$\tilde{x}_{0,25} = -1,0\ ,\ \tilde{x}_{0,75} = 0,5 \Rightarrow \text{IQR} = 1,5 \quad .$$

Ergänzt man $x_6 = 10,5$, so ergibt sich der Wert IQR $= 2,5$, jedoch bei $\hat{x}_6 = -10,5$ ergibt sich IQR $= 10,0$. Ursache für diesen auffälligen Unterschied ist die Lage der beiden Ausreißer am gleichen Ende der Stichprobe. Ihr Anteil an der Stichprobe beträgt $\frac{2}{6} = 0,33$ und übersteigt damit den Bruchpunkt.

4.2.2 Der Median der absoluten Abweichung vom Median

Sei F eine beliebige Verteilungsfunktion, dann ist

$$T(F) = \frac{1}{2}\left[F^{-1}\left(\frac{3}{4}\right) - F^{-1}\left(\frac{1}{4}\right)\right]$$

das Funktional, das die Hälfte des Interquartilabstandes angibt. Zu diesem Funktional kann man nun eine „symmetrische Version“ $\tilde{T}$ konstruieren, indem man

$$\bar{F}(x) = 1 - F(-x + F^{-1}(1/2))$$

und

$$\tilde{F}(x) = \frac{1}{2}[F(x) + \bar{F}(x)]$$

setzt. $\tilde{F}(x)$ entsteht dabei aus $F(x)$ durch eine Symmetrisierung am Median m_n. Definiert man anschließend

$$\tilde{T}(F) := T(\tilde{F}) \quad ,$$

so ist $\tilde{T}$ der Median der absoluten Abweichung vom Median (MAD).

Definition 4.2.2. (Median der absoluten Abweichung vom Median)
Als Median der absoluten Abweichung vom Median (MAD) einer Stichprobe vom Umfang n bezeichnet man

$$\text{MAD}_n = \text{med}\{|x_i - m_n|\} \quad ,$$

wobei

$$\text{med}\{x_i\} = m_n \quad .$$

Der Vorteil der symmetrisierten Version des halben Interquartilabstandes liegt darin, dass sich der Wert für den Bruchpunkt verdoppelt, d.h.

$$\varepsilon^*_{\text{MAD}} = 0,5 = 2 \cdot 0,25 = 2 \cdot \varepsilon^*_{\text{IQR}} \quad .$$

Darüber hinaus ist die Einflusskurve beschränkt, so dass einzelne Beobachtungswerte nur begrenzten Einfluss auf den Schätzwert haben.

Beispiel 4.2.3.
Es seien die fünf Beobachtungswerte wie in Beispiel 2.1.2 gegeben. Zur Berechnung des MAD benötigt man zunächst den Wert des Medians für diese Stichprobe. Der Median beträgt hier $m_5 = -0,5$. Damit lassen sich die Grössen $y_i = |x_i - m_5|$, $i = 1, \ldots, 5$ berechnen und man erhält:

$$y_1 = 0,0;\ y_2 = 2,0;\ y_3 = 0,5;\ y_4 = 1,0;\ y_5 = 9,0 \quad .$$

Berechnet man auf Grundlage dieser Werte den Median erneut, so erhält man schliesslich als MAD dieser Stichprobe den Wert $\text{MAD}_5 = 1,0$.
Eliminiert man $x_5 = -9,5$ als Ausreißer, so erhält man $\text{MAD}_4 = 0,75$. Der Wert der Sensitivitätskurve an der Stelle $x = -9,5$ ergibt sich zu:

$$\text{SC}(-9,5; x_1, \ldots, x_4, \text{MAD}_5) = -1,25 \quad .$$

Bei der Berücksichtigung von $x_1, \ldots, x_5$ ergibt sich eine Standardabweichung von $s_5 = 4,41$, wohingegen bei $x_1, \ldots, x_4$ sich der Wert $s_4 = 1,11$ ergeben würde, d.h.

$$\text{SC}(-9,5; x_1, \ldots, x_4, s_5) = 16,5 \quad .$$

Mittels einer Multiplikation mit einer Konstanten kann man einen Schätzwert korrigieren.

Beispiel 4.2.4. Schätzer für die Standardabweichung einer Normalverteilung
Sind $x_1, \ldots, x_n$ Realisationen aus einer $\mathcal{N}(\mu, \sigma^2)$-Verteilung, so ist der MAD ein Schätzer für

$$F^{-1}(0.75) \cdot \sigma \quad ,$$

so dass $\frac{1}{0{,}6745} \cdot \text{MAD}$ als Schätzer für σ selbst gilt, wobei $F^{-1}(0.75) = 0,6745$ das $0,75$-Quantil der Standardnormalverteilung ist (s. [7]). Verwendet man den Interquartilabstand, so ergibt sich entsprechend, dass $\frac{1}{2} \cdot \frac{1}{0{,}6745} \cdot \text{IQR} \approx 0,7413 \cdot \text{IQR}$ ein Schätzer für σ ist.

4.2.3 Die α-gestutzte Varianz

Die α-gestutzte Varianz ist definiert als die skalierte Varianz einer α-gestutzten Stichprobe:

$$s^2_{gest} = \frac{1}{n - 2[n\alpha]} \sum_{i=[n\alpha]+1}^{n-[n\alpha]} (x_{(i)} - \bar{x}_\alpha)^2 \quad .$$

Jedoch unterschätzt dieser Schätzer die Varianz und ist keine konsistente Schätzung für die Varianz. Daher benutzt man einen Korrekturterm

$$\frac{1 - 2\alpha}{1 - 2\alpha - 2\tilde{x}_{1-\alpha}\phi(\tilde{x}_{1-\alpha})} \quad ,$$

wobei $\tilde{x}_{1-\alpha}$ das $(1-\alpha)$-Quantil der Standardnormalverteilung ist. Damit ist

$$\hat{\sigma}^2_{gest} = \frac{1 - 2\alpha}{1 - 2\alpha - 2\tilde{x}_{1-\alpha}\phi(\tilde{x}_{1-\alpha})} \cdot s^2_{gest}$$

eine konsistente Schätzung für σ^2 unter der Normalverteilung.

4.2.4 Die α-winsorisierte Varianz

Die α-winsorisierte Varianz ist definiert als

$$s^2_{win} = \frac{1}{n-1}\Bigg(\sum_{i=[n\alpha]+1}^{n-[n\alpha]} (x_{(i)} - \bar{x}_{\alpha,win})^2 + [n\alpha](x_{([n\alpha]+1)} - \bar{x}_{\alpha,win})^2 + [n\alpha](x_{(n-[n\alpha])} - \bar{x}_{\alpha,win})^2 \Bigg) \quad ,$$

wobei auch dies zunächst kein konsistenter Schätzer für die Varianz ist. Analog zur α-gestutzten Varianz kann man auch für diesen Fall einen Korrekturterm angeben, so dass

$$\hat{\sigma}^2_{win} = \frac{1}{1 - 2\alpha - 2\tilde{x}_{1-\alpha}\phi(\tilde{x}_{1-\alpha}) + 2\alpha\tilde{x}^2_{1-\alpha}} \cdot s^2_{win}$$

insgesamt ein konsistenter Schätzer für σ^2 unter der Normalverteilung ist.

Kapitel 5

M-Schätzer

5.1 Theoretischer Hintergrund

Die Theorie der M-Schätzer basiert auf einer Verallgemeinerung der Maximum-Likelihood-Methode (s. Beispiel 3.2.8). Wie dort sei $F_\theta(x) := F(x;\theta)$ eine Verteilung aus der Verteilungsfamilie $\{F_\theta |\ \theta \in \Theta \subset \mathbb{R}\}$ mit einer Dichte $f_\theta(x) := f(x;\theta)$. Die Maximum-Likelihood-Methode liefert einen Schätzer für den unbekannten Parameter θ, indem der Ausdruck $\sum_{i=1}^{n}(-\log(f(x_i;\theta)))$ minimiert wird. Ist f differenzierbar, so führt dies zur Lösung des Gleichungssystems

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta}(-\log(f(x_i;\theta))) = 0 \quad .$$

Beispiel 5.1.1. Normalverteilung
Der oben beschriebene Ansatz führt unter der Voraussetzung, dass eine Normalverteilung mit Erwartungswert μ und Varianz σ^2 vorliegt, zu folgender Problemstellung bei der Suche nach einem Schätzer für den Erwartungswert:

$$\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2} \to \min_{\mu} \quad ,$$

d.h.

$$\sum_{i=1}^{n}(x_i-\mu) = 0 \quad .$$

Das Ergebnis ist das arithmetische Mittel $\bar{x}_n$.

Probleme bei diesem Vorgehen tauchen dann auf, wenn Ausreißer in den Daten vorhanden sind. D.h. die Ausreißerempfindlichkeit der ML-Methode ist nicht befriedigend. Um diese zu verbessern, ersetzt man $-\log(f(x;\theta))$ durch eine Funktion $\rho(x,\theta)$, die weniger sensitiv auf Ausreißer reagiert. Damit lautet das verallgemeinerte Vorgehen

$$\text{minimiere den Ausdruck } \sum_{i=1}^{n} \rho(x_i,\theta) \tag{5.1}$$

bzw.

$$\text{löse das Gleichungssystem } \sum_{i=1}^{n} \psi(x_i,\theta) = 0 \quad , \tag{5.2}$$

wobei $\psi(x,\theta)$ die Ableitung von $\rho(x,\theta)$ nach θ ist. Jede Lösung von (5.1) bzw. (5.2) wird M-Schätzer genannt.

5.1.1 Grundlagen

Zu einer beliebigen Funktion $\psi(x,t)$ sei ein zugehöriges Funktional T auf einer Menge von Verteilungsfunktionen F wie folgt definiert. Es ist $T(F)$ die Lösung t_0 der Gleichung

$$\int \psi(x,t_0)\ dF(x) = 0 \tag{5.3}$$

Ein solches T wird das zu ψ gehörende M-Funktional genannt. Für eine Stichprobe $x_1,\ldots,x_n$ aus einer Verteilung F ist der zu ψ gehörende M-Schätzer T_n die Lösung der Gleichung

$$\sum_{i=1}^{n} \psi(x_i,T_n) = 0 \quad ,$$

die sich aus (5.3) mit $F = F_n$ ergibt, d.h. $T_n = T(F_n)$. Dabei ist zu beachten, dass mehrere Lösungen existieren können. Gleichung (5.3) ergibt sich in vielen Fällen aus der Minimierung des Ausdrucks

$$\int \rho(x,t_0)\ dF(x) \quad ,$$

wobei dann die Funktion ψ bestimmt ist durch

$$\psi(x,t) = c \cdot \frac{\partial}{\partial t}\rho(x,t)$$

mit einer Konstanten c, falls $\rho(x,\cdot)$ hinreichend glatt ist.
Gilt für ein Funktional T auf der Verteilungsfamilie $\{F_\theta |\ \theta \in \Theta \subset \mathbb{R}\}$

$$T(F_\theta) = \theta \quad ,$$

dann ist der zugehörige Schätzer $T_n = T(F_n)$ ein Schätzer für θ. Ein Kriterium zur Beurteilung eines Schätzers stellt die Einflusskurve (s. Abschnitt 3.2) dar, die somit zur Bestimmung einer geeigneten ψ-Funktion und somit eines geeigneten M-Schätzers herangezogen werden kann.
Ziel ist es nun, das Gâteaux-Differential eines M-Funktionals zu berechnen, d. h. es ist $d_1T(F, G-F)$ für ein Funktional T gesucht, das Lösung t_0 der Gleichung

$$\int \psi(x,t_0)dF(x) = 0$$

ist. $\psi(x,t)$ sei dabei eine beliebige Funktion. Es sei

$$H(t,\lambda) = \int \psi(x,t)dF_\lambda(x) \quad .$$

Mit

$$F_\lambda(x) = F + \lambda(G-F)$$

folgt wie in Beispiel 3.2.8 durch implizite Differentiation der Gleichung $H(T(F_\lambda), \lambda) = 0$ nach λ an der Stelle $\lambda = 0$ [1]

$$\begin{aligned}
d_1T(F, G-F) &= \frac{d}{d\lambda}T(F+\lambda(G-F))\Big|_{\lambda=0} \\
&= -\frac{\partial H(t,\lambda)}{\partial\lambda}\Big|_{t=T(F),\lambda=0} \Big/ \frac{\partial H(t,\lambda)}{\partial t}\Big|_{t=T(F),\lambda=0} \\
&= \frac{-\frac{\partial}{\partial\lambda}\left[\int \psi(x,t)\, dF(x) + \lambda \int \psi(x,t)\, d[G(x)-F(x)]\right]\big|_{t=T(F),\lambda=0}}{\frac{d}{dt}\int \psi(x,t)\, dF(x)\big|_{t=T(F)}} \\
&= -\frac{\int \psi(x,t)\, dG(x)\big|_{t=T(F)} - \overbrace{\int \psi(x,t)\, dF(x)\Big|_{t=T(F)}}^{\overset{(5.3)}{=}0}}{\frac{d}{dt}\int \psi(x,t)\, dF(x)\big|_{t=T(F)}} \\
&= -\frac{\int \psi(x,T(F))dG(x)}{\lambda'_F(T(F))}
\end{aligned}$$

vorausgesetzt, dass $\lambda'_F(T(F)) \neq 0$ mit

$$\lambda_F(t) = \int \psi(x,t)\, dF(x), \quad -\infty < t < \infty \quad .$$

Damit besitzt die Einflusskurve von ψ die Form:

$$IC(x; F, T) = -\frac{\psi(x, T(F))}{\lambda'_F(T(F))}, \quad -\infty < x < \infty \quad .$$

Aufgrund der Tatsache, dass die Einflusskurve proportional zu ψ ist, besitzen M-Schätzer die Eigenschaft, dass Bedingungen an die Einflusskurve über die Wahl von ψ gesteuert werden können. D.h. eine geeignete Wahl von ψ führt zu vorteilhaften Eigenschaften der Einflusskurve und damit des M-Schätzers.

5.1.2 Asymptotische Eigenschaften von M-Schätzern

In diesem Abschnitt werden M-Schätzer hinsichtlich Konsistenz und asymptotischer Normalität untersucht. Zunächst sei dazu an dieser Stelle auf die verschiedenen Konvergenzbegriffe in der Wahrscheinlichkeitstheorie hingewiesen.

Definition 5.1.2. (Stochastische Konvergenz)
Seien $X_1, X_2, \ldots$ und X Zufallsgrößen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$.
Dann konvergiert $(X_n)_{n\geq 1}$ stochastisch gegen X, falls

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1 \quad \forall \varepsilon > 0 \quad .$$

Bez.: $X_n \xrightarrow{st} X \quad (n \to \infty)$

[1] g in Beispiel 3.2.8 entspricht ψ hier

Definition 5.1.3. (Fast sichere Konvergenz)
Seien $X_1, X_2, \ldots$ und X Zufallsgrößen über einem Wahrscheinlichkeitsraum $(\Omega, \mathcal{A}, \mathbb{P})$.
$(X_n)_{n\geq 1}$ konvergiert fast sicher (f.s.) gegen X, falls

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1 \quad .$$

Bez.: $X_n \xrightarrow{f.s.} X \quad (n \to \infty)$

Definition 5.1.4. (Konvergenz der Verteilung nach)
Seien $F_1(\cdot), F_2(\cdot), \ldots$ und $F(\cdot)$ Verteilungsfunktionen und seien $X_1, X_2, \ldots$ sowie X Zufallsgrößen (nicht notwendig über einem gemeinsamen Wahrscheinlichkeitsraum) mit den angegebenen Verteilungen.
Dann konvergiert $(X_n)_{n\geq 1}$ der Verteilung nach gegen X, falls

$$\lim_{n\to\infty} F_n(t) = F(t) \quad \forall \text{ Stetigkeitsstellen } t \text{ von } F \quad .$$

Bez.: $X_n \xrightarrow{n.V.} X \quad (n \to \infty)$

Konsistenz eines M-Schätzers

Definition 5.1.5. (Starke und schwache Konsistenz eines Schätzers)
Eine Folge $\{T_n\}$ von Schätzern für eine parametrische Funktion $g(\theta)$ heißt konsistent, falls T_n gegen $g(\theta)$ in geeigneter Weise konvergiert. Man spricht von schwacher Konsistenz, falls

$$T_n \xrightarrow{st} g(\theta) \qquad (n \to \infty)$$

und von starker Konsistenz im Falle von

$$T_n \xrightarrow{f.s.} g(\theta) \qquad (n \to \infty) \quad .$$

Sei $\psi(x,t)$ eine Funktion und $\lambda_F(t) = \int \psi(x,t)\, dF(x)$. Außerdem besitze die Gleichung $\lambda_F(t) = 0$ eine Lösung t_0 und die „empirische“ Gleichung $\lambda_{F_n}(t) = 0$ ebenfalls eine Lösung T_n. Im Folgenden werden nun Bedingungen genannt, unter denen $T_n \xrightarrow{f.s.} t_0$ gilt. Dazu sei $x_1, \ldots, x_n$ eine Stichprobe aus einer nach F verteilten Grundgesamtheit mit empirischer Verteilungsfunktion F_n.

Lemma 5.1.6.
Sei t_0 eine isolierte Lösung der Gleichung $\lambda_F(t) = 0$ und sei $\psi(x,t)$ monoton in t. Dann ist t_0 eindeutig und jede Folge $\{T_n\}$ von Lösungen der empirischen Gleichung $\lambda_{F_n}(t) = 0$ konvergiert fast sicher gegen t_0. Ist außerdem $\psi(x,t)$ stetig in t in einer Umgebung von t_0, so existiert eine solche Folge von Lösungen.

Für den Beweis sei an dieser Stelle auf [16], Abschnitt (7.2.1) verwiesen.

Bemerkung 5.1.7.
t_0 muss nicht notwendigerweise eine Nullstelle von $\lambda_F(t)$ sein. Es reicht aus, dass die Funktion in jeder hinreichend kleinen Umgebung genau einmal das Vorzeichen wechselt.

Die Beispiele in 5.2.1, 5.2.2 und 5.2.3 zeigen konsistente Schätzer der jeweiligen Lageparameter. Für den Hampel-Schätzer (s. Beispiel in 5.2.4) reicht dieses Ergebnis jedoch nicht aus und man benötigt

Lemma 5.1.8.
Sei t_0 eine isolierte Nullstelle der Gleichung $\lambda_F(t) = 0$ und sei $\psi(x,t)$ stetig in t und beschränkt. Dann besitzt die empirische Gleichung $\lambda_{F_n}(t) = 0$ eine Folge $\{T_n\}$ von Lösungen, die mit Wahrscheinlichkeit 1 gegen t_0 konvergiert.

Der Beweis hierzu ist ebenfalls in [16], Abschnitt (7.2.1) zu finden.

Asymptotische Normalverteilung

Sei $\psi(x,t)$ gegeben und $\lambda_F(t) = \int \psi(x,t)\, dF(x)$. Außerdem sei $t_0 = T(F)$ eine Lösung der Gleichung $\lambda_F(t) = 0$. $\{X_i\}$ ist eine Folge von unabhängig und identisch verteilten Zufallsgrößen mit einer Verteilung F und $T_n = T(F_n)$ eine zu t_0 konsistente Folge von Lösungen von $\lambda_{F_n}(t) = 0$. In dem vorliegenden Abschnitt sollen nun Bedingungen angegeben werden, unter denen

$$\sqrt{n}(T_n - t_0) \xrightarrow{n.V.} \mathcal{N}(0, \sigma^2(T,F)) \tag{5.4}$$

gilt. Dabei ist $\sigma^2(T,F)$ abhängig von den Annahmen an $\psi(x,t)$ entweder gegeben durch

$$\frac{\int \psi^2(x,t_0)\; dF(x)}{(\lambda_F'(t_0))^2}$$

oder durch

$$\frac{\int \psi^2(x,t_0)\; dF(x)}{\left(\int \left(\frac{\partial\psi(x,t)}{\partial t}\right)_{t=t_0} dF(x)\right)^2} \quad .$$

Das folgende Ergebnis behandelt den Fall, dass die Funktion $\psi(x,t)$ monoton in t ist und wird in [16], Abschnitt (7.2.2) bewiesen.

Satz 5.1.9.
Sei t_0 eine isolierte Nullstelle von $\lambda_F(t) = 0$ und $\psi(x,t)$ monoton in t. Angenommen, $\lambda_F(t)$ ist differenzierbar in $t = t_0$ mit $\lambda_F'(t_0) \neq 0$ und $\int \psi^2(x,t)\; dF(x)$ ist endlich für t in einer Umgebung von t_0 und stetig in $t = t_0$. Dann erfüllt jede Folge T_n von Lösungen der empirischen Gleichung $\lambda_{F_n}(t) = 0$ die Bedingung (5.4), wobei $\sigma^2(T,F)$ gegeben ist durch

$$\frac{\int \psi^2(x,t_0)\; dF(x)}{[\lambda_F'(t_0)]^2} \quad .$$

Darüber hinaus gilt $T_n \xrightarrow{f.s.} t_0$ nach Lemma 5.1.6.

5.2 M-Schätzer für Lageparameter

Sind die das M-Funktional bestimmenden Funktionen ψ und ρ von der Form $\psi(x,t) = \tilde{\psi}(x-t)$ bzw. $\rho(x,t) = \tilde{\rho}(x-t)$, so nennt man $T(F)$ einen Lageparameter. Ist zudem F symmetrisch zu θ, ρ gerade und ψ ungerade, so gilt $T(F) = \theta$, d.h. alle M-Schätzer mit ungeradem ψ schätzen das Symmetriezentrum θ einer symmetrischen Verteilung. Ausserdem sollten M-Schätzer translations- und skalierungsäquivariant sein, d.h. es sollte gelten:

$$T_n(x_1,\ldots,x_n) = BT_n\left(\frac{x_1 - A}{B},\ldots,\frac{x_n - A}{B}\right) + A \quad .$$

Die meisten M-Schätzer für einen Lageparameter müssen daher eine Skalierung der Daten berücksichtigen, so dass sie diese Bedingung erfüllen. Dabei stellen das arithmetische Mittel und der Median zwei Ausnahmen dar, denn für den Mittelwert gilt

$$\psi(x,t) = x - t$$

und damit folgt

$$\begin{aligned} \psi(bx+a, bt+a) &= bx + a - bt - a \\ &= bx - bt \\ &= b(x-t) \\ &= b\psi(x,t) \quad . \end{aligned}$$

Analoges gilt für den Median mit $\psi(x,t) = \text{sgn}(x-t)$. Alle anderen M-Schätzer benötigen die Skalierung der Argumente von ρ und ψ. Daher wird ein Skalierungsschätzer s_n als Funktion von $x_1, \ldots, x_n$ benutzt, der zusammen mit einer Tuning-Konstanten C den Ausdruck $x_i - t$ neu skaliert. Damit erhält man zentrierte und skalierte Daten über den Ausdruck

$$u_i = \frac{x_i - t}{C \cdot s_n} \quad .$$

5.2.1 Das arithmetische Mittel

Sucht man einen kleinste-Quadrate-Schätzer, so will man den Ausdruck

$$\sum_{i=1}^{n} (x_i - \theta)^2$$

minimieren. Als ρ und ψ ergeben sich in diesem Falle

$$\rho(u) = \frac{1}{2}u^2$$

$$\psi(u) = u, \quad -\infty < u < \infty \quad .$$

Damit ist das M-Funktional T der Erwartungswert und der M-Schätzer ist das arithmetische Mittel $\overline{x}_n$.

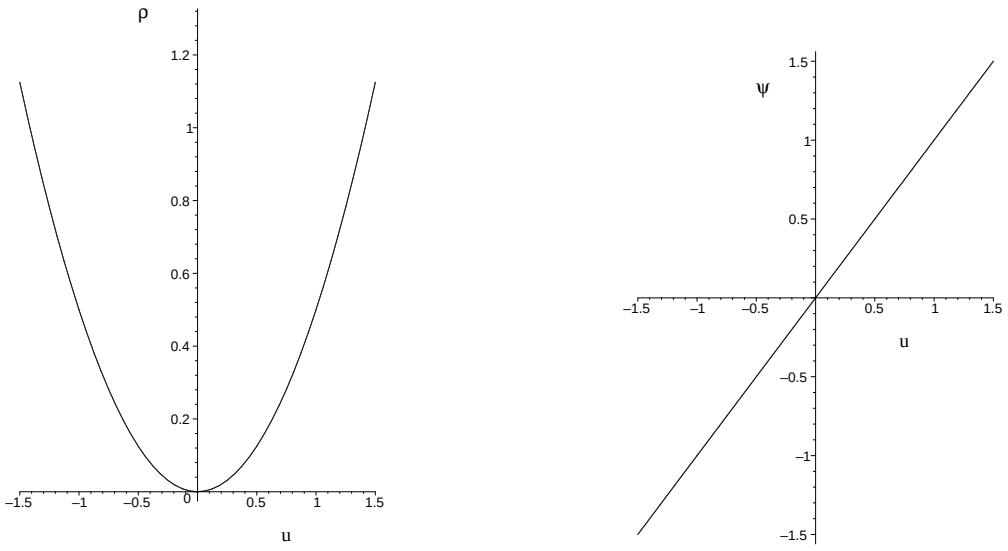

Abbildung 5.1: Arithmetische Mittel $\rho(u)$ (links) und $\psi(u)$ (rechts)

Die ψ-Funktion ist eine Gerade und nach beiden Seiten unbeschränkt, d.h. Ausreißer haben einen starken Einfluss.

5.2.2 Der Median

Möchte man die absoluten Abweichungen minimieren, also den Ausdruck

$$\sum_{i=1}^{n} |x_i - \theta| \quad ,$$

so lauten die zugehörigen Funktionen für ρ und ψ:

$$\rho(u) = |u|$$

$$\psi(u) = \text{sgn}(u)$$

Hier ist der M-Schätzer der Stichprobenmedian. Er ist unempfindlich gegenüber Ausreißern, da die ψ-Funktion beschränkt ist. Die Sprungstelle von ψ bei $u = 0$ zeigt jedoch an, dass der Median sehr stark von den „mittleren“ Werten abhängt.

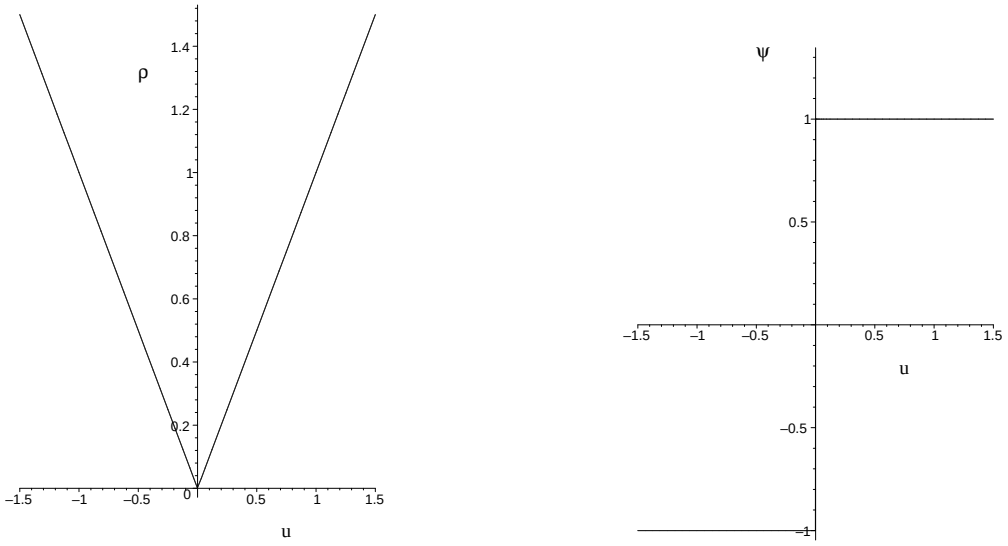

Abbildung 5.2: Median $\rho(u)$ (links) und $\psi(u)$ (rechts)

5.2.3 Der Huber-Schätzer

Als Familie der Huber-Schätzer bezeichnet man die M-Schätzer, die

$$\sum_{i=1}^{n} \rho(u_i)$$

minimieren. Dabei ist ρ von der Form

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{, falls} \quad |u| \leq k \\ k|u| - \frac{1}{2}k^2 & \text{, falls} \quad |u| > k \end{cases} .$$

Das entsprechende ψ lautet

$$\psi(u) = \begin{cases} u & \text{, falls} \quad |u| \leq k \\ k \cdot \text{sgn}(u) & \text{, falls} \quad |u| > k \end{cases} .$$

Dabei hat sich in Anwendungen gezeigt, dass sich als Skalierungsschätzer s_n der normalisierte MAD empfielt, so dass C gleich Eins gesetzt wird und für k sollte man einen Wert im Intervall $[1;2]$ wählen. Der zugehörige M-Schätzer ist von der Art eines winsorisierten Mittelwertes, d.h. es ist der Stichprobenmittelwert über die x_i's, wobei diejenigen x_i durch $T_n \pm kCs_n$ ersetzt werden, für die $|u_i| > k$ gilt. Hierbei wird der Einfluss von Ausreißern gedämpft, aber nicht vollständig ausgeräumt.

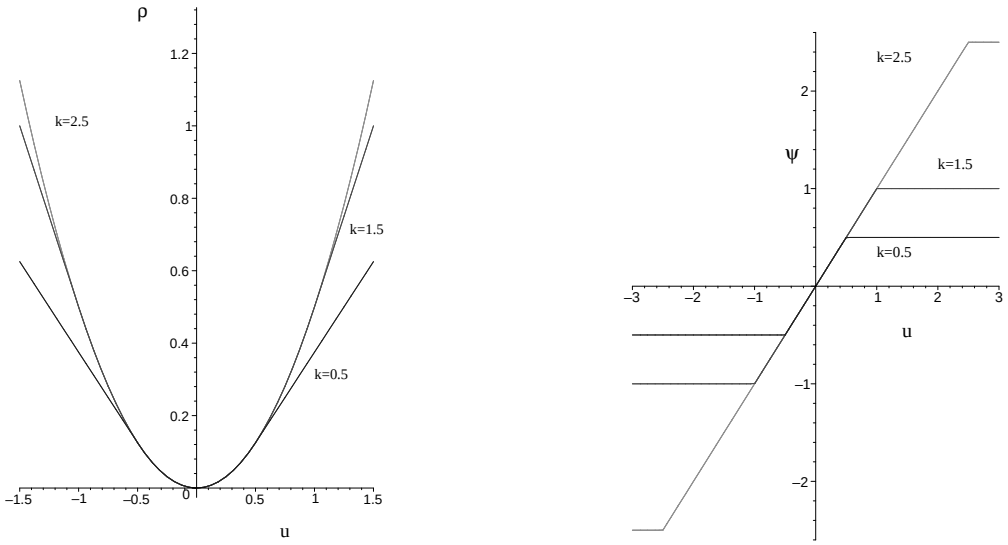

Abbildung 5.3: Huber-Schätzer $\rho(u)$ (links) und $\psi(u)$ (rechts)

5.2.4 Der Hampel-Schätzer

Hampel [5] hat den Huber-Schätzer im Bezug auf die Ausreißerempfindlichkeit (gross-error-sensitivity) modifiziert. Er verlangt, dass $\psi(u)$ für Ausreißer den Wert Null annehmen soll. Analog zu den bisherigen Fällen ist auch hier der Ausdruck

$$\sum_{i=1}^{n} \rho(u_i)$$

zu minimieren. Dabei ist

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & \text{, falls} \quad |u| \le a \\ a|u| - \frac{1}{2}a^2 & \text{, falls} \quad a < |u| \le b \\ ab - \frac{1}{2}a^2 + (c-b)\frac{a}{2}\left[1 - \left(\frac{c-|u|}{c-b}\right)^2\right] & \text{, falls} \quad b < |u| \le c \\ ab - \frac{1}{2}a^2 + (c-b)\frac{a}{2} & \text{, falls} \quad |u| > c \end{cases}$$

und

$$\psi(u) = \begin{cases} u & \text{, falls} \quad |u| \le a \\ a \cdot \text{sgn}(u) & \text{, falls} \quad a < |u| \le b \\ a \cdot \left(\frac{c-|u|}{c-b}\right) \cdot \text{sgn}(u) & \text{, falls} \quad b < |u| \le c \\ 0 & \text{, falls} \quad |u| > c \end{cases}$$

Hierbei sind a, b und c positive Konstanten mit $a < b < c$ aus dem Intervall $[2;9]$ (zum Beispiel $a = 2, b = 4$ und $c = 8$). Als Skalierungsschätzer s_n wird wie im vorangegangenen Beispiel beim Huber-Schätzer der normalisierte MAD benutzt und C gleich Eins gesetzt. Beim Hampel-Schätzer bleiben Ausreißer vollständig unberücksichtigt.

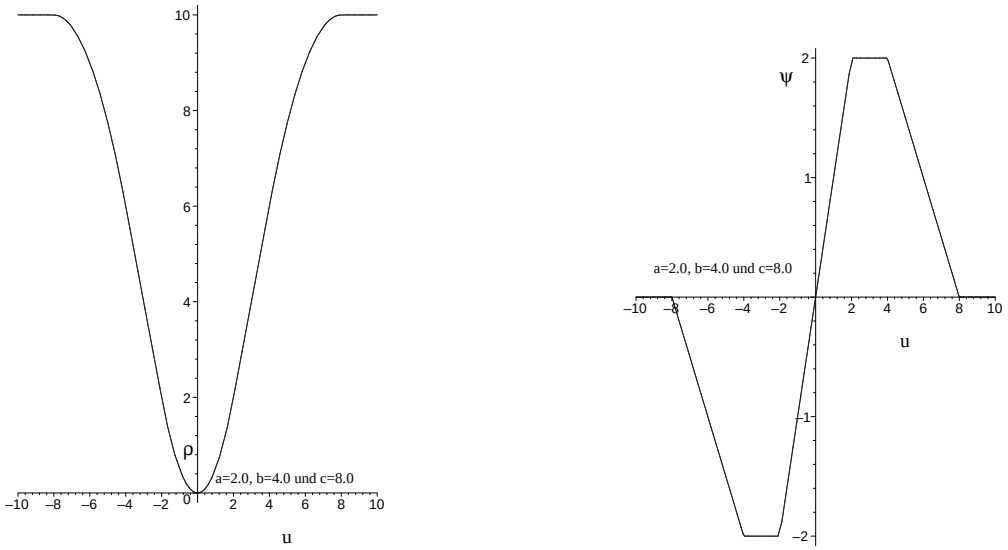

Abbildung 5.4: Hampel-Schätzer $\rho(u)$ (links) und $\psi(u)$ (rechts)

5.2.5 Andrew's wave

Man geht hier ebenfalls von dem zu minimierenden Ausdruck

$$\sum_{i=1}^{n} \rho(u_i)$$

aus mit

$$\rho(u) = \begin{cases} \frac{1}{\pi^2} \cdot (1 - \cos(\pi u)) & \text{, falls} \quad |u| \leq 1 \\ \frac{2}{\pi^2} & \text{, falls} \quad |u| > 1 \end{cases} .$$

Die zugehörige ψ-Funktion ist gegeben durch

$$\psi(u) = \begin{cases} \frac{1}{\pi} \sin(\pi u) & \text{, falls} \quad |u| \leq 1 \\ 0 & \text{, falls} \quad |u| > 1 \end{cases}$$

und eliminiert ebenfalls Ausreißer vollständig. Hier wird als Skalierungsschätzer ebenfalls $s_n = \text{MAD}$ empfohlen und C sollte im Bereich $[1,5\pi;\ 2,4\pi]$ liegen.

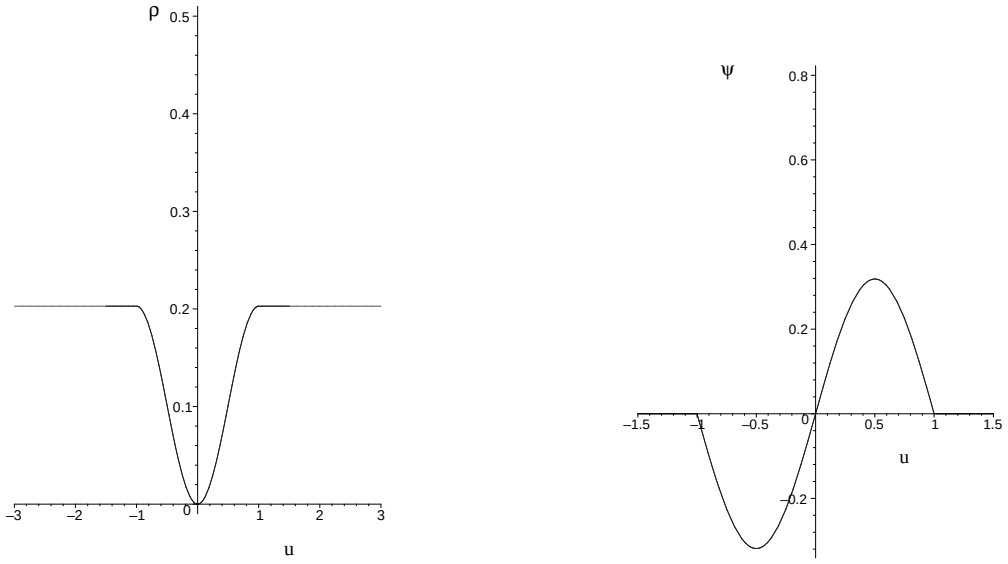

Abbildung 5.5: Andrew's wave $\rho(u)$ (links), $\psi(u)$ (rechts)

5.2.6 Tukey's biweight

Zu dem zu minimierenden Ausdruck

$$\sum_{i=1}^{n} \rho(u_i)$$

mit

$$\rho(u) = \begin{cases} \frac{1}{6} \cdot (1 - (1 - u^2)^3) & \text{, falls} \quad |u| \leq 1 \\ \frac{1}{6} & \text{, falls} \quad |u| > 1 \end{cases}$$

lautet das zugehörige ψ:

$$\psi(u) = \begin{cases} u(1 - u^2)^2 & \text{, falls} \quad |u| \leq 1 \\ 0 & \text{, falls} \quad |u| > 1 \end{cases}$$

Der MAD wird als Skalierungsschätzer s_n verwendet und für C empfehlen sich Werte aus dem Intervall $[6; 12]$.

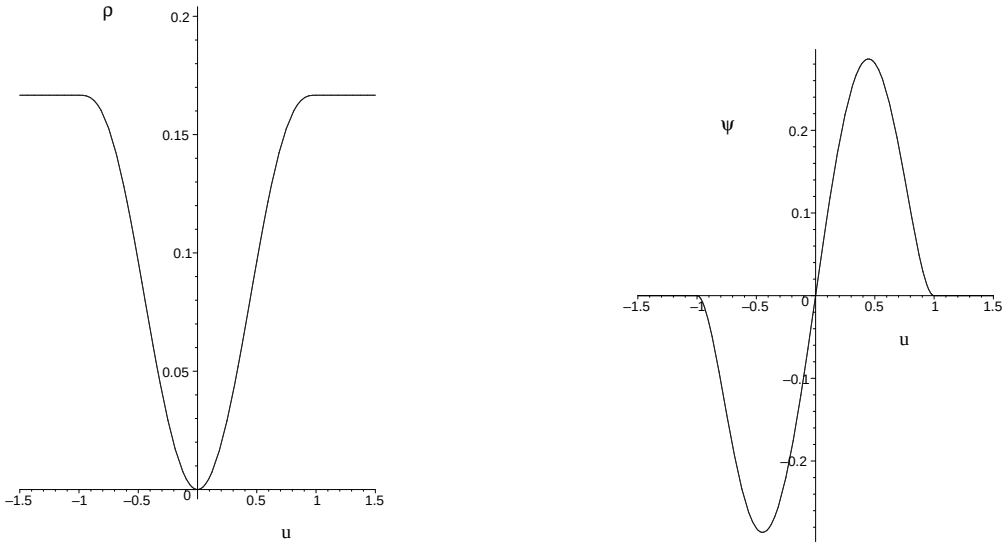

Abbildung 5.6: Tukey's biweight $\rho(u)$ (links), $\psi(u)$ (rechts)

Auch hier werden Ausreißer mit einem zu großen Abstand zum Zentrum der Datenpunkte eliminiert. Für kleine u gilt $\psi(u) \approx u$. Daher verhält sich dieser Schätzer in der Nähe des Datenzentrums ähnlich dem arithmetischen Mittel.

Von diesen sechs Schätzern wurden der Huber-Schätzer, Andrew's wave und Tukey's biweight in der die Diplomarbeit begleitend entwickelten Software implementiert und dienen dort der Berechnung von robusten Lagemaßzahlen. Auf der Grundlage dieser robusten Schätzungen können dann verlässliche Aussagen über das Zentrum der Daten gemacht werden.

5.3 Diskussion der Eigenschaften verschiedener M-Schätzer

Die Robustheitseigenschaften der im vorangegangenen Abschnitt vorgestellten M-Schätzer kann man an Hand der ψ-Funktion diskutieren. Folgende Eigenschaften sollte die ψ-Funktion eines robusten Lage-M-Schätzers einer symmetrischen Verteilung besitzen:

1. Hoher Bruchpunkt.
2. ψ ist beschränkt.

3. ψ besitzt „annähernd“ stetiges Verhalten.
4. ψ besitzt endlichen Ablehnungspunkt.
5. $\psi(u) \approx k \cdot u, k \neq 0$ für kleine Werte von u.
6. $\psi(-u) = -\psi(u)$.

Für die wichtigsten sechs M-Schätzer fasst folgende Tabelle die geforderten Eigenschaften zusammen:

Schätzer	1.	2.	3.	4.	5.	6.
Mittelwert	nein	nein	ja	nein	ja	ja
Median	ja	ja	nein	nein	nein	ja
Huber	ja	ja	ja	nein	ja	ja
Hampel	ja	ja	ja	ja	ja	ja
Andrew‘s wave	ja	ja	ja	ja	ja	ja
Tukey‘s biweight	ja	ja	ja	ja	ja	ja

Tabelle 5.1: Eigenschaften der wichtigsten Schätzer

Die Spaltennummern beziehen sich auf die Aufzählung zu Beginn dieses Abschnittes.
Die Tuning-Konstante k beim Huber-Schätzer kann bei bekanntem Ausreißeranteil ε folgendermaßen bestimmt werden:

$$\frac{2\phi(k)}{k} - 2\Phi(-k) = \frac{\varepsilon}{1-\varepsilon} \quad ,$$

wobei Φ die Standard-Normalverteilungsfunktion bezeichnet und ϕ die zugehörige Dichtefunktion.
Die Tuning-Konstanten bei der Familie der Hampel-Schätzer können auf analoge Weise bestimmt werden ([8]).
Möchte man die im vorliegenden Abschnitt vorgestellten M-Schätzer mit den L-Schätzern aus Kapitel 4 vergleichen, so fällt auf, dass die M-Schätzer in der Regel einen hohen Bruchpunkt nahe bei $\varepsilon^* = 0,5$ haben und eine hohe Effizienz in einer Umgebung der Modellverteilung aufweisen (siehe dazu [9]). L-Schätzer hingegen besitzen in diesem Falle eine geringere Effizienz oder einen kleineren Bruchpunkt.

5.4 Berechnung von M-Schätzern

5.4.1 Berechnung mittels Newton-Raphson-Verfahren

Zur Berechnung eines Lage-M-Schätzers ist T_n als Lösung der Gleichung

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - T_n}{c \cdot S_n}\right) = 0$$

zu bestimmen. Im Allgemeinen ist diese Lösung jedoch nicht explizit angebbar. Einzige Ausnahme stellt der Fall des arithmetischen Mittels dar mit $\psi(x_i, T_n) = x_i - T_n$. In allen anderen Fällen muss daher mit Hilfe eines geeigneten Iterationsverfahrens eine Lösung gesucht werden. Als geeignet hat sich das Newton-Raphson-Verfahren erwiesen. Dabei ist eine Nullstelle der Funktion $h(t)$ gesucht. Diese Nullstelle wird durch Auswertungen von h und h' an einer Näherungsstelle $t^{(k)}$ bestimmt. $t^{(k+1)}$ ist dann die Stelle, an der die Tangente an h im Punkt $t^{(k)}$ die t-Achse schneidet. Bei der Berechnung eines M-Schätzers lautet die Funktion h:

$$h(T_n) = \sum_{i=1}^{n} \psi\left(\frac{x_i - T_n}{c \cdot S_n}\right) \quad .$$

Für die Iterationsvorschrift ergibt sich daraus

$$\begin{aligned} T_n^{(k+1)} &= T_n^{(k)} - \frac{h(T_n^{(k)})}{h'(T_n^{(k)})} \\ &= T_n^{(k)} + c \cdot S_n \frac{\sum_{i=1}^n \psi(u_i^{(k)})}{\sum_{i=1}^n \psi'(u_i^{(k)})} \quad . \end{aligned}$$

Als Startwert sollte für $T_n^{(0)}$ der Median gewählt werden als erste robuste Näherung und der gesuchte M-Schätzer ergibt sich dann als Grenzwert der Folge $(T_n^{(k)})_{k\geq 1}$. Das Ergebnis ist in den meisten Fällen ein robuster Schätzer.

5.4.2 1-Schritt-Verfahren

Führt man beim Newton-Raphson-Verfahren nur einen einzigen Iterationsschritt aus, so wird das daraus resultierende Ergebnis für den gesuchten Schätzer 1-Schritt-M-Schätzer genannt. Bei diesem Vorgehen ist offensichtlich die Wahl des Startwertes für $T_n^{(0)}$ sehr wichtig. Falls als Startwert der arithmetische Mittelwert benutzt wird, ist das Ergebnis nach einem Iterationsschritt unbefriedigend. Es sollte daher der Median als Anfangswert genutzt werden, damit das Ergebnis einen robusten M-Schätzer liefert.

5.5 M-Schätzer für Streuungsparameter

Grundlage sei eine Verteilungsfamilie mit Dichten $\frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right) = f(x;\mu,\sigma)$ mit Lageparameter μ und Skalierungsparameter σ. Wendet man das Prinzip der Maximum-Likelihood-Schätzung auf diesen Fall an, so ergibt sich folgendes zu lösendes Problem

$$\sum_{i=1}^n -\log\left(\frac{1}{\sigma} f_0\left(\frac{x_i-\mu}{\sigma}\right)\right) \to \min_{\mu,\sigma} \quad .$$

Aufgrund der Rechenregeln für die Logarithmusfunktion ergibt sich daraus

$$\sum_{i=1}^n \Bigg(\log(\sigma) + \underbrace{\left(-\log\left(f_0\left(\frac{x_i-\mu}{\sigma}\right)\right)\right)}_{=:\rho_{f_0}\left(\frac{x_i-\mu}{\sigma}\right)}\Bigg) \to \min_{\mu,\sigma} \quad .$$

Zur Bestimmung des Minimums werden nun die partiellen Ableitungen nach μ und σ gebildet, diese gleich Null gesetzt und das so entstandene System der Normalengleichungen gelöst, wobei $\psi_f = \rho_f'$ ist. Man erhält

$$\sum_{i=1}^n \frac{1}{\sigma} \psi_{f_0}\left(\frac{x_i-\mu}{\sigma}\right) = 0 \quad \wedge \quad \sum_{i=1}^n \left(\frac{1}{\sigma} - \frac{x_i-\mu}{\sigma^2} \psi_{f_0}\left(\frac{x_i-\mu}{\sigma}\right)\right) = 0$$

und daraus ergibt sich

$$\sum_{i=1}^n \psi_{f_0}\left(\frac{x_i-\mu}{\sigma}\right) = 0 \quad \wedge \quad \sum_{i=1}^n \left(\frac{x_i-\mu}{\sigma} \psi_{f_0}\left(\frac{x_i-\mu}{\sigma}\right) - 1\right) = 0$$

Um nun simultane M-Schätzer für Lage und Streuung zu erhalten, wird dieses zu lösende System verallgemeinert zu

$$\sum_{i=1}^{n} \psi\left(\frac{x_i - T}{S}\right) = 0 \quad \wedge \quad \sum_{i=1}^{n} \chi\left(\frac{x_i - T}{S}\right) = 0$$

Im Allgemeinen ist dabei ψ eine ungerade und χ eine gerade Funktion. In Funktionalschreibweise ergeben sich damit folgende Integralgleichungen:

$$\int \psi\left(\frac{x - T}{S}\right)\, dF(x) = 0 \quad \wedge \quad \int \chi\left(\frac{x - T}{S}\right)\, dF(x) = 0 \quad .$$

D.h. die Schätzer für den Lageparameter und den Streuungsparameter werden durch simultanes Lösen dieser beiden Gleichungen bestimmt.

Beispiel 5.5.1. Huber-Schätzer für Lage und Streuung
Mit

$$\psi(x) = \max[-k, \min(k, x)]$$

und

$$\chi(x) = \min(c^2, x^2) - \beta \quad ,$$

wobei $\beta = \beta(c) = \int \min(c^2, x^2)\ \phi(x)\ dx$ und $0 < \beta < c^2$, ergibt sich ein unter Normalverteilung konsistenter Schätzer für die Streuung mit $\phi(\cdot)$ als der Dichtefunktion der Standardnormalverteilung.

Beispiel 5.5.2. Median und MAD
Für

$$\psi(x) = sgn(x)$$

und

$$\chi(x) = sgn(|x| - 1)$$

ergeben sich der Median als Schätzer für den Lageparameter und der MAD als Schätzung für die Streuung.

Eine weitere Möglichkeit, einen robusten Schätzer für die Varianz bzw. Streuung einer Stichprobe zu erhalten ist die Verwendung von sogenannten A-Schätzern.
Die A-Schätzer

$$s_T = \frac{C \cdot \text{MAD}\sqrt{n} \cdot \sqrt{\sum_{i=1}^{n} \psi(u_i)^2}}{|\sum_{i=1}^{n} \psi'(u_i)|}$$

erhält man aus der asymptotischen Varianz der M-Lageparameterschätzer (s. Abschnitt 5.2) mit $u_i = \frac{x_i - m_n}{C \cdot \text{MAD}}$.

Beispiel 5.5.3.
Setzt man

$$\psi(u) = \begin{cases} sin(\pi u) & \text{, falls} \quad |u| < 1 \\ 0 & \text{, falls} \quad |u| \geq 1 \end{cases} \quad ,$$

so führt dies zu dem Skalenschätzer

$$s_{wa} = \frac{(C \cdot \text{MAD})\sqrt{n}[\sum_{|u_i|<1} sin^2(\pi u_i)]^{1/2}}{\pi|\sum_{|u_i|<1} cos(\pi u_i)|} \quad ,$$

der auf Andrew‘s Wave Schätzer für einen Lageparameter (s. 5.2.5) basiert.

Ein weiteres Beispiel zeigt einen A-Schätzer für die Streuung, der auf dem Biweight-Lageschätzer (s. 5.2.6) beruht.

Beispiel 5.5.4.
Sei

$$\psi(u) = \begin{cases} u(1-u^2)^2 & \text{, falls} \quad |u| < 1 \\ 0 & \text{, falls} \quad |u| \geq 1 \end{cases},$$

dann erhält man

$$s_{bi} = \frac{\sqrt{n}\left[\sum_{|u_i|<1}(x_i - m_n)^2(1-u_i^2)^4\right]^{1/2}}{|\sum_{|u_i|<1}(1-u_i^2)(1-5u_i^2)|}$$

als Skalenschätzer.

Die Skalenschätzer aus den beiden vorangegangenen Beispielen sind in der begleitenden Software implementiert worden.

Kapitel 6

Robuste Schätzer für multivariate Lage und Streuung

6.1 Einleitung und Motivation

Bisher lagen n Beobachtungen aus einer Beobachtungsreihe vor, für die in der Regel *ein* Lageparameter und *ein* Parameter für die Streuung zu schätzen waren. Im vorliegenden Kapitel seien nun mehrere Beobachtungsreihen gegeben, d.h. es seien $x^{(1)}, \ldots, x^{(n)}$ p-variate Beobachtungsdaten. Die robuste Schätzung eines Lageparameters für diesen multivariaten Fall führt daher zur Bestimmung eines robusten Lagevektors. Ein multivariater Lagevektor kann auf zwei unterschiedliche Methoden geschätzt werden. Zum einen kann man komponentenweise für jede Beobachtungsreihe einen einzelnen robusten Lageschätzer bestimmen. Dazu eignen sich die in Kapitel 4 und Kapitel 5 vorgestellten L- bzw. M-Schätzer.
Eine weitere Möglichkeit besteht in der multivariaten, simultanen Schätzung des Lagevektors, bei der alle Beobachtungen gemeinsam eingehen. Dieses Vorgehen stellt eine Erweiterung der M-Schätzer für den multivariaten Fall dar. Im Abschnitt 6.3 werden dazu M-Schätzer für multivariate Lage und Streuung betrachtet.
Die Schätzung der Streuung läuft auf die Berechnung von Kovarianz- bzw. Korrelationsmatrizen hinaus. Auf der Grundlage der Matrizen, die sich aus diesen Kovarianzen bzw. Korrelationen aufbauen, wird zum Beispiel die Hauptkomponentenanalyse (s. Kapitel 7) oder die Faktoranalyse durchgeführt. Darüber hinaus dienen Kovarianz- und Korrelationsmatrizen dem Test auf Unabhängigkeit. Die entsprechenden empirischen Matrizen sind jedoch sehr ausreißerempfindlich, so dass robusten Methoden eine große Bedeutung zukommt.

Definition 6.1.1. (**Kovarianz und Korrelation**)
Gegeben seien zwei Zufallsgrößen X und Y. Die Kovarianz von X und Y ist definiert als

$$\text{Kov}(X,Y) = \mathbb{E}\Big((X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y)) \Big) \quad .$$

Unter der Korrelation von X und Y versteht man den Ausdruck

$$\text{Korr}(X,Y) = \frac{\text{Kov}(X,Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}} \quad .$$

Definition 6.1.2. (Kovarianz- und Korrelationsmatrix)
Seien $X_1, \ldots, X_n$ Zufallsvariablen, dann ist die Kovarianzmatrix definiert als

$$\Sigma = \begin{pmatrix} \text{Var}X_1 & \text{Kov}(X_1, X_2) & \ldots & \text{Kov}(X_1, X_n) \\ \text{Kov}(X_2, X_1) & \text{Var}X_2 & \ldots & \text{Kov}(X_2, X_n) \\ \vdots & \ddots & \ddots & \vdots \\ \text{Kov}(X_n, X_1) & \ldots & \text{Kov}(X_n, X_{n-1}) & \text{Var}X_n \end{pmatrix}$$

und als Korrelationsmatrix wird

$$R = \begin{pmatrix} 1 & \text{Korr}(X_1, X_2) & \ldots & \text{Korr}(X_1, X_n) \\ \text{Korr}(X_1, X_2) & 1 & \ldots & \text{Korr}(X_2, X_n) \\ \vdots & \ddots & \ddots & \vdots \\ \text{Korr}(X_1, X_n) & \ldots & \text{Korr}(X_{n-1}, X_n) & 1 \end{pmatrix}$$

bezeichnet.
Bei der Berechnung robuster Kovarianz- bzw. Korrelationsmatrizen unterscheidet man zwei Vorgehensweisen. Zum einen kann man jeden Eintrag in der Kovarianz- bzw. Korrelationsmatrix separat robust schätzen. Dieses Vorgehen führt zu univariaten Analysen für die Varianzen und bivariaten Analysen für die Kovarianzen bzw. Korrelationen. Dieses Verfahren hat aber den Nachteil, dass die sich ergebenden Matrizen im Allgemeinen nicht positiv semi-definit sind. Die zweite Vorgehensweise besteht in der simultanen Schätzung aller Elemente der Kovarianzmatrix und führt zu einer multivariaten Analyse. Die aus dieser Methode sich ergebenden Matrizen sind positiv semi-definit. Die Methode der separaten Schätzung bietet jedoch Vorteile, wenn ein größerer Anteil der Beobachtungen fehlen.

6.2 Separate robuste Schätzer für die Elemente der Kovarianz- bzw. Korrelationsmatrix

Ein einfacher Ansatz zur Schätzung der Kovarianz zwischen zwei Zufallsvariablen Y_1 und Y_2 beruht auf der Gleichung

$$\text{Kov}(Y_1, Y_2) = \frac{1}{4}[\text{Var}(Y_1 + Y_2) - \text{Var}(Y_1 - Y_2)] \quad .$$

Die Schätzung der Kovarianz kann somit zurückgeführt werden auf die Schätzung von Varianzen. Auf der Grundlage der robusten univariaten Streuungsschätzer aus den Abschnitten 4.2 und 5.5 ist es also möglich, robuste Schätzer für Kovarianz- und Korrelationsmatrix anzugeben.
Einen robusten Schätzer s_{12}^* für die Kovarianz zwischen Y_1 und Y_2 erhält man mittels

$$s_{12}^* = \frac{1}{4}[\hat{\sigma}_1^{*2} - \hat{\sigma}_2^{*2}] \quad ,$$

wobei $\hat{\sigma}_1^*, \hat{\sigma}_2^*$ robuste Skalenschätzer für $Y_1 + Y_2$ und $Y_1 - Y_2$ sind. Damit erhält man auch eine robuste Schätzung der Korrelation zwischen Y_1 und Y_2:

$$r_{12}^* = \frac{s_{12}^*}{\sqrt{s_{11}^* \cdot s_{22}^*}} \quad ,$$

wobei s_{ii}^* eine robuste Schätzung der Varianz von Y_i ist.
Berechnet man für zwei Zufallsgrößen X und Y die herkömmliche Korrelation

$$r_{12} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y}_n)^2}} \quad ,$$

so liegt der Wert aufgrund der Gültigkeit der Cauchy-Schwarz-Ungleichung immer im Intervall $[-1;1]$. Da aber r_{12}^* nicht auf den normalen Standardabweichungen, sondern auf robusten Schätzungen für die Varianzen basiert, muss r_{12}^* nicht zwangsläufig im Intervall $[-1;1]$ liegen. Um diese Eigenschaft der Korrelation sicher zu stellen, wendet man eine Modifikation an:
Sei

$$Z_i = \frac{Y_i}{\sqrt{s_{ii}^*}}$$

die standardisierte Form von Y_i mit einer robusten Schätzung s_{ii}^* der Varianz. Definiert man jetzt

$$\hat{\rho}_{12}^* = \frac{\hat{\sigma}_3^{*2} - \hat{\sigma}_4^{*2}}{\hat{\sigma}_3^{*2} + \hat{\sigma}_4^{*2}} \quad , \tag{6.1}$$

wobei $\hat{\sigma}_3^{*2}, \hat{\sigma}_4^{*2}$ robuste Schätzungen der Varianzen von $Z_1 + Z_2$ bzw. $Z_1 - Z_2$ sind, so ergibt sich, dass $\hat{\rho}_{12}^*$ im Intervall $[-1;+1]$ liegt. Die robusten Varianzschätzungen sind im Allgemeinen nicht konsistent. Eine konsistente Schätzung der Varianz bei Normalverteilung erhält man durch geeignete Normierungskonstanten. Diese müssen bei der Berechnung der Kovarianzen jedoch nicht berücksichtigt werden, da sie sich herauskürzen.
Aus Gleichung (6.1) lässt sich leicht auch der entsprechende robuste Kovarianzschätzer herleiten:

$$\hat{\sigma_{12}^*} = \hat{\rho}_{12}^* \cdot \sqrt{s_{11}^* s_{22}^*} \quad .$$

In der begleitenden Software wurde dieses Verfahren mit den robusten Verfahren der α-gestutzten Varianz und der α-winsorisierten Varianz (s. Abschnitt 4.2.3 und 4.2.4) implementiert.

6.3 M-Schätzer für multivariate Lage- und Skalierungsparameter

Wie im univariaten Fall ergeben sich die M-Schätzer durch Verallgemeinerung der Maximum-Likelihood-Schätzung. Zunächst wird die Klasse der Normalverteilungen, für die die zugehörigen Maximum-Likelihood-Schätzer nicht robust sind, zu einer Verteilungsklasse erweitert, die auch Verteilungsfamilien mit stärker besetzten „Schwänzen“ (Tails) enthält, deren Maximum-Likelihood-Schätzer robust sind.
Ein p-dimensionaler Zufallsvektor X mit der Verteilungsdichte

$$f(x) = (\det \Sigma)^{-\frac{1}{2}} h\left((x-t)^\top \Sigma^{-1}(x-t)\right)^1 \quad ,$$

wobei h eine nichtnegative reellwertige Funktion, $t \in \mathbb{R}^p$ und V eine positiv-definite Matrix ist, heißt elliptisch symmetrisch verteilt.

Beispiel 6.3.1. Normalverteilung und t-Verteilung

1. Für
$$h(u) = \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{u}{2}\right)$$
ergibt sich die Normalverteilung.
2. Mit
$$h(u) = \frac{\Gamma(\frac{\nu+p}{2})}{(\pi\nu)^{p/2}\Gamma(\nu/2)} \cdot \frac{1}{(1+\frac{1}{\nu}u)^{(\nu+p)/2}} \tag{6.2}$$
ergibt sich die Familie der p-dimensionalen t-Verteilung mit ν Freiheitsgraden. Wie im univariaten Fall sind die „Schwänze“ der multivariaten t-Verteilung stärker besetzt als die der Normalverteilung. Die auf der t-Verteilung basierende Maximum-Likelihood-Schätzung gewichtet stark abweichende Beobachtungen schwächer.

[1] mit $x^\top$ bzw. $A^\top$ sei im Folgenden der transponierte Vektor bzw. die transponierte Matrix bezeichnet

Die Maximum-Likelihood-Schätzung ergibt sich durch Maximierung der Likelihood-Funktion

$$\prod_{i=1}^{n} f(x^{(i)}) = \prod_{i=1}^{n} \frac{1}{\sqrt{\det(V)}} h\left((x^{(i)}-t)^\top V^{-1}(x^{(i)}-t)\right) \to \max_{t,V} \quad ,$$

wobei mit $x^{(i)}$ der Vektor der i-ten Beobachtungsreihe bezeichnet wird, d.h.

$$x^{(i)} = \left(x_1^{(i)}, \ldots, x_p^{(i)}\right), \quad 1 \le i \le n \quad .$$

Logarithmiert man die Likelihood-Funktion und maximiert bzgl. V^{-1} statt V, so ergibt sich die äquivalente Formel

$$\log \prod_{i=1}^{n} f(x^{(i)}) = \frac{n}{2} \log(\det V^{-1}) + \sum_{i=1}^{n} \log\left(h\left((x^{(i)}-t)^\top V^{-1}(x^{(i)}-t)\right)\right) \to \max_{t,V^{-1}} \quad .$$

Zur Bestimmung des Maximums werden nun die beiden partiellen Ableitungen nach t und V^{-1} gebildet:

$$\frac{\partial}{\partial t}\left[\log \prod_{i=1}^{n} f(x^{(i)})\right] = \sum_{i=1}^{n} \frac{-2h'\left((x^{(i)}-t)^\top V^{-1}(x^{(i)}-t)\right)}{h\left((x^{(i)}-t)^\top V^{-1}(x^{(i)}-t)\right)} V^{-1}(x^{(i)}-t)$$

$$\begin{aligned} \frac{\partial}{\partial v^{kl}}\left[\log \prod_{i=1}^{n} f(x^{(i)})\right] &= \frac{n}{2} \underbrace{\frac{\frac{\partial}{\partial v^{kl}} \det(V^{-1})}{\det(V^{-1})}}_{=\frac{\det(V^{kl})}{\det(V^{-1})} = v_{lk}} \\ &+ \sum_{i=1}^{n} \frac{h'\left((x^{(i)}-t)^\top V^{-1}(x^{(i)}-t)\right)}{h\left((x^{(i)}-t)^\top V^{-1}(x^{(i)}-t)\right)} \cdot (x_k^{(i)} - t_k)(x_l^{(i)} - t_l) \end{aligned}$$

wobei v^{kl} das (k,l)-te Element der inversen Matrix V^{-1} bezeichnet und V^{kl} die Matrix ist, die aus der Matrix V entsteht, indem das (k,l)-te Element gleich Eins gesetzt wird und die restlichen Elemente der k-ten Zeile und l-ten Spalte gleich Null gesetzt werden. Damit ergibt sich insgesamt für die partielle Ableitung nach der inversen Matrix V^{-1}

$$\frac{\partial}{\partial V^{-1}}\left[\log \prod_{i=1}^{n} f(x^{(i)})\right] = \frac{n}{2} V^\top + \sum_{i=1}^{n} \frac{h'\left((x^{(i)}-t)^\top V^{-1}(x^{(i)}-t)\right)}{h\left((x^{(i)}-t)^\top V^{-1}(x^{(i)}-t)\right)} \cdot (x^{(i)}-t)(x^{(i)}-t)^\top \quad .$$

Beide partiellen Ableitungen werden anschliessend gleich Null gesetzt, wobei als abkürzende Schreibweisen

$$\begin{aligned} d_i^2 &= (x^{(i)}-t)^\top V^{-1}(x^{(i)}-t) \\ w(d_i^2) &= \frac{-2h'(d_i^2)}{h(d_i^2)} \end{aligned} \tag{6.3}$$

verwendet werden und $V = V^\top$ gilt. Damit ergibt sich folgendes Gleichungssystem

$$\begin{aligned} \sum_{i=1}^{n} w(d_i^2)(x^{(i)}-t) &= 0 \\ \sum_{i=1}^{n} \left(w(d_i^2)(x^{(i)}-t)(x^{(i)}-t)^\top - V\right) &= 0 \quad . \end{aligned}$$

Durch Verallgemeinerung dieses Gleichungssystems ergeben sich die M-Schätzer für multivariate Lage und Streuung. Die Lösungen des Gleichungssystems

$$\sum_{i=1}^{n} w_1(d_i^2)(x^{(i)} - t) \quad = \quad 0 \tag{6.4}$$

$$\sum_{i=1}^{n} \left(w_2(d_i^2)(x^{(i)} - t)(x^{(i)} - t)^\top - v(d_i^2)V \right) \quad = \quad 0 \tag{6.5}$$

heißen M-Schätzer für den multivariaten Lageparameter und die Kovarianzmatrix.
Hieraus ergeben sich die Maximum-Likelihood-Schätzer der p-dimensionalen t-Verteilung direkt aus (6.3) und (6.2):

$$w_1(d) = \frac{m+\nu}{\nu + d^2} = w_2(d^2) \quad \text{und} \quad v(d_i^2) = 1 \quad ,$$

während bei den auf Hubers ψ-Funktionen basierenden Schätzern

$$w_1(d) = \begin{cases} 1 & \text{, falls} \quad d \leq a \\ \frac{a}{d} & \text{, falls} \quad d > a \end{cases}$$

und

$$w_2(d^2) = w_1(d)^2 \quad \text{und} \quad v(d_i^2) = c$$

gilt, wobei c ein Korrekturterm für die asymptotische Erwartungstreue unter der Normalverteilung ist und gemäß [14], S. 225 berechnet werden kann.

Beispiel 6.3.2. Die für dieses Beispiel genutzten Daten werden im Datensatz A im Anhang angegeben. Aus den zunächst 109 Datenpaaren werden drei Beobachtungswerte als Ausreißer identifiziert und nacheinander eliminiert. Die Angaben zur Korrelation in der folgenden Tabelle zeigen deutlich, dass bei der herkömmlichen Korrelationsberechnung mit den drei Ausreißern der Wert keine Aussage über eine Beziehung zulässt. Die Werte, die sich bei Benutzung der robusten Verfahren ergeben, ähneln sich und zeigen eine Korrelation von ca. $0,9$.

Verfahren	**Korrelation**	**Ausreißer**	**Tuning-Parameter**
einfache Korrelation	-0,35	keiner	
	-0,81	(1; 105)	
	-0,88	(1; 105), (10; 10)	
	-0,89	(1; 105), (10; 10), (14; 1)	
0,01-getrimmt	-0,81	(1; 105), (10; 10)	
0,05-getrimmt	-0,90	(1; 105), (10; 10), (14; 1)	
0,01-winsorisiert	-0,75	(1; 105), (10; 10)	
0,05-winsorisiert	-0,90	(1; 105), (10; 10), (14; 1)	
Tukey‘s biweight	-0,90	(1; 105), (10; 10), (14; 1)	9
(A-Schätzer)	-0,91	(1; 105), (10; 10), (14; 1)	6
Andrew‘s wave	-0,90	(1; 105), (10; 10), (14; 1)	6,6
(A-Schätzer)			
Huber	-0,91	(1; 105), (10; 10), (14; 1)	1,5
(A-Schätzer)			
Huber	-0.89	(1; 105), (10; 10), (14; 1)	
(Simultan)			

Tabelle 6.1: Daten zur Korrelation bei verschiedenen M-Schätzern

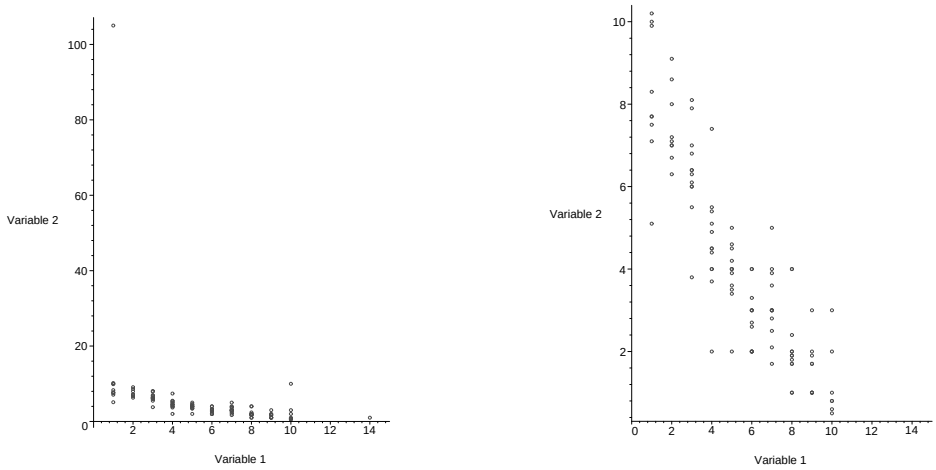

Abbildung 6.1: Streubild zweier Variablen: 109 Datenpaare (links), 106 Datenpaare (rechts)

6.4 ROBETH

ROBETH ist eine FORTRAN Subroutine Library [14]. Aus dieser Bibliothek wurden Routinen zur Berechnung robuster Kovarianzmatrizen auf der Grundlage von M-Schätzern genutzt. Die Routinen basieren auf den theoretischen Grundlagen aus den Arbeiten von Huber [9], Hampel [6] und Rousseeuw [15]. ROBETH ist in ANSI FORTRAN 77 geschrieben und sowohl auf Großrechnern wie auch auf Microcomputern installiert worden. Bekannte Bibliotheken wie etwa die NAG-Library haben Teile von ROBETH übernommen.

6.4.1 Theoretische Grundlagen

Der in ROBETH verwendete Algorithmus zur Berechnung robuster multivariater Lageparameter und Bestimmung robuster Streuungsmatrizen beruht auf den mathematischen Grundlagen wie sie im Abschnitt 6.3 angegeben sind. Das heißt, dass der Vektor der Lageparameter und die Streuungsmatrix durch simultanes Lösen von Gleichungen der Form (6.4) und (6.5) bestimmt werden.
Im folgenden Abschnitt werden vier Möglichkeiten für geeignete Gewichtsfunktionen $w_1(\cdot), w_2(\cdot)$ und $v(\cdot)$ angegeben, die in der begleitenden Software zur Auswahl stehen. Das System der Matrixgleichungen wird in ROBETH mit Hilfe des Newton-Algorithmus simultan gelöst (s. [14]).

6.4.2 Mögliche Gewichtsfunktionen

In der von ROBETH übernommenen Routine zur Berechnung robuster Kovarianzmatrizen einer multivariaten Verteilung werden folgende Gewichtsfunktionen zur Auswahl gestellt:

1. klassische Analyse

   $$w_1(s) \equiv 1$$

   $$w_2(s) \equiv 1$$

   $$v(s) \equiv 1$$

2. Huber's Minimax

$$w_1(s) = \begin{cases} 1 & \text{, falls} \quad s < c \\ \frac{c}{s} & \text{, falls} \quad s \geq c \end{cases}$$

$$w_2(s) = \begin{cases} \frac{a^2}{s^2} & \text{, falls} \quad s^2 < a^2 \\ 1 & \text{, falls} \quad a^2 \leq s^2 \leq b^2 \\ \frac{b^2}{s^2} & \text{, falls} \quad b^2 < s^2 \end{cases}$$

$$v(s) = d$$

wobei a, b, c und d gegebene Konstanten sind.

3. Hampel-Krasker

$w_1(s) \equiv 1$

$w_2(s) = 2\Phi(c/s) - 1$

$v(s) \equiv 1$

wobei c eine gegebene Konstante ist und mit $\Phi(\cdot)$ die Verteilungsfunktion der Standardnormalverteilung bezeichnet wird.

4. Krasker-Welsch

$w_1(s) \equiv 1$

$w_2(s) = t^2 + (1 - t^2)(2\Phi(t) - 1) - 2t\phi(t)$, wobei $t = c/s$

$v(s) \equiv 1$

wobei c eine gegebene Konstante ist und $\phi(\cdot)$ die Dichtefunktion der Standardnormalverteilung darstellt.

Kapitel 7

Robuste Hauptkomponentenanalyse

7.1 Motivation

Bei der Hauptkomponentenanalyse geht man der Frage nach, ob eine Reduktion der Dimension des Merkmalsraumes ohne wesentlichen Informationsverlust möglich ist. D.h. man versucht, die Originaldaten durch eine kleinere Anzahl „dahinter liegender“ Variablen zu ersetzen. Dazu nimmt man eine orthogonale Hauptachsentransformation der ursprünglichen Variablen in eine neue Menge unkorrelierter Variablen, den sogenannten Hauptkomponenten (principal components), vor. Die Hauptkomponenten werden nacheinander in absteigender Bedeutung konstruiert und sind Linearkombinationen der ursprünglichen Variablen. Man hofft dabei, dass nur wenige der ersten Variablen für den größten Teil der Variation in den Originaldaten verantwortlich sind, so dass die effektive Dimension der Daten reduziert werden kann.
Das Vorgehen erfolgt in zwei Schritten. Zunächst wird der Ursprung des neuen Koordinatensystems in den Schwerpunkt der Punktwolke der Originaldaten gelegt, dann wird das Koordinatensystem so gedreht, dass die erste Koordinate in Richtung der größten Varianz der Punktwolke zeigt. Somit erhält man die erste Hauptachse, und die Varianz in dieser Richtung ist der größte Eigenwert. Die nächste Drehung wird um diese Koordinatenachse durchgeführt unter den Bedingungen, dass die zweite Hauptachse orthogonal zur ersten liegen muss und in die Richtung der größten verbleibenden Varianz zeigen muss. Diesen Schritt wiederholt man solange, bis eine neue p-dimensionale Basis erzeugt ist.

7.2 Herleitung der Hauptkomponenten

Nach [11] sei $X = (X_1, \ldots, X_p)^\top$ ein p-dimensionaler Zufallsvektor mit Erwartungswertvektor μ und Kovarianzmatrix Σ, der die p Merkmale bezeichnet. Die Aufgabe ist es nun, neue Variablen $Z_1, \ldots, Z_p$ zu finden, die unkorreliert sind und deren Varianzen mit wachsendem Index $j = 1, \ldots, p$ fallen. Jedes Z_j ist eine Linearkombination der X_i, so dass

$$Z_j = a_{1j}X_1 + a_{2j}X_2 + \ldots + a_{pj}X_p = a_j^\top X \tag{7.1}$$

wobei $a_j^\top = (a_{1j}, \ldots, a_{pj})$ ein Vektor von Konstanten ist. Da man einen Einheitsrichtungsvektor in Richtung der Hauptachse haben möchte, wird der Vektor a_j so normiert, dass

$$a_j^\top a_j = \sum_{k=1}^{p} a_{kj}^2 = 1 \quad .$$

Die erste Hauptkomponente Z_1 wird bestimmt durch die Bedingung, dass die Varianz maximal sein soll, d.h. es soll gelten

$$\text{Var}(Z_1) = \text{Var}(a_1^\top X) = a_1^\top \Sigma a_1 \to \max$$

mit der Kovarianzmatrix

$$\Sigma = \begin{pmatrix} \text{Var}(X_1) & \text{Kov}(X_1, X_2) & \dots & \text{Kov}(X_1, X_p) \\ \text{Kov}(X_2, X_1) & \text{Var}(X_2) & \dots & \text{Kov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Kov}(X_p, X_1) & \text{Kov}(X_p, X_2) & \dots & \text{Var}(X_p) \end{pmatrix}$$

und unter der Nebenbedingung $a_1^\top a_1 = 1$. Bestimmt man nun für diese Kovarianzmatrix ihre Eigenwerte, so können mehrere Eigenwerte den gleichen Wert haben bzw. gleich Null sein. Im Folgenden stellt es jedoch keine Einschränkung dar, wenn angenommen wird, dass die Eigenwerte alle paarweise verschieden und ungleich Null sind. Außerdem lassen sich so Missverständnisse in Bezeichnungen vermeiden.
Zur Bestimmung des Maximums benutzt man das Verfahren nach Lagrange mit Nebenbedingungen (s. [3], Kapitel 3.5). Unter Berücksichtigung der Nebenbedingung $a_1^\top a_1 = 1$ lautet der zu maximierende Ausdruck

$$a_1^\top \Sigma a_1 - \lambda(a_1^\top a_1 - 1) \quad ,$$

wobei λ der Lagrange-Multiplikator ist. Leitet man diesen Ausdruck nach a_1 ab und setzt anschließend den Term gleich Null, so ergibt sich folgende zu lösende Gleichung

$$(\Sigma - \lambda I_p)\, a_1 = 0 \quad , \tag{7.2}$$

wobei I_p die p-dimensionale Einheitsmatrix bezeichnet. Diese Gleichung ergibt sich ebenso, wenn man die Eigenwerte der Matrix Σ berechnen möchte.

Definition 7.2.1. (**Eigenwert und Eigenvektor einer Matrix**)
Sei $A \in \mathbb{R}^{p\times p}$ eine quadratische Matrix und $x \in \mathbb{R}^p$ mit $x \neq 0$ ein p-dimensionaler Vektor. Erfüllt x die Gleichung

$$Ax = \lambda x \quad , \tag{7.3}$$

so heißt x der Eigenvektor von A und λ der zugehörige Eigenwert von A.

Bemerkung 7.2.2. Die Bestimmungsgleichung (7.3) für einen Eigenwert von A ist gleichbedeutend mit der Bedingung, dass $(A - \lambda I_n)\, x = 0$ gelten soll.

Damit entspricht die Suche nach der ersten Hauptkomponente der Lösung des Eigenwertproblems der Kovarianzmatrix und a_1 ist ein Eigenvektor der Kovarianzmatrix Σ und λ der zugehörige Eigenwert. Mit

$$a_1^\top \Sigma a_1 \overset{(7.2)}{=} a_1^\top \lambda a_1 = \lambda a_1^\top a_1 = \lambda$$

folgt, dass a_1 der zum größten Eigenwert λ_1 gehörende Eigenvektor ist und damit ist die erste Hauptkomponente bestimmt durch

$$Z_1 = a_1^\top X \quad \text{mit} \quad \text{Var}(Z_1) = \lambda_1 \quad .$$

Die zweite Hauptkomponente soll ebenfalls maximale Varianz unter den noch verbliebenen, zu Z_1 unkorrelierten Linearkombinationen der X_i haben, d.h.

$$\text{Kov}(Z_1, Z_2) = 0 \quad .$$

Es gilt

$$\text{Kov}(Z_1, Z_2) = \text{Kov}(a_1^\top X, a_2^\top X) = a_1^\top \Sigma a_2 = a_2^\top \Sigma a_1 = a_2^\top \lambda_1 a_1 = \lambda_1 a_2^\top a_1 = 0 \quad .$$

Daher muss $a_2^\top a_1 = 0 \quad (\lambda_1 \neq 0)$ gelten, wenn Z_1 unkorreliert zu Z_2 sein soll. Man hat also eine weitere Nebenbedingung bei der Bestimmung der zweiten Hauptkomponente mit maximaler Varianz:

$$a_2^\top \Sigma a_2 - \lambda(a_2^\top a_2 - 1) - \hat{\lambda} a_2^\top a_1 \quad ,$$

wobei $\lambda, \hat{\lambda}$ die Lagrange-Multipliaktoren für die beiden Nebenbedingungen sind. Leitet man den Ausdruck nach a_2 ab, setzt ihn gleich Null und multipliziert von links mit $a_1^\top$, so erhält man

$$a_1^\top \Sigma a_2 - a_1^\top \lambda a_2 - a_1^\top \hat{\lambda} a_1 = 0 \quad .$$

Aufgrund der Nebenbedingungen folgt, dass $\hat{\lambda} = 0$ sein muss und daher wegen

$$\Sigma a_2 - \lambda a_2 = 0 \quad \Leftrightarrow \quad (\Sigma - \lambda I_p)\, a_2 = 0$$

erneut ein Eigenwertproblem zu lösen ist. Wegen der Anforderung, dass die Varianz maximal sein soll, ist λ der zweitgrößte Eigenwert λ_2 und a_2 der zugehörige Eigenvektor, falls die Eigenwerte der Kovarianzmatrix paarweise verschieden sind. Analog lassen sich die noch verbleibenden Hauptkomponenten bestimmen:

$$Z_k = a_k^\top p \quad \text{und} \quad \mathrm{Var}(Z_k) = \lambda_k, \quad k = 1, \ldots, p \quad .$$

Beispiel 7.2.3. Führt man eine klassische Hauptkomponentenanalyse für die Daten aus dem Datensatz A durch, so ergibt sich folgendes Bild bei Berücksichtigung aller 109 Beobachtungen:

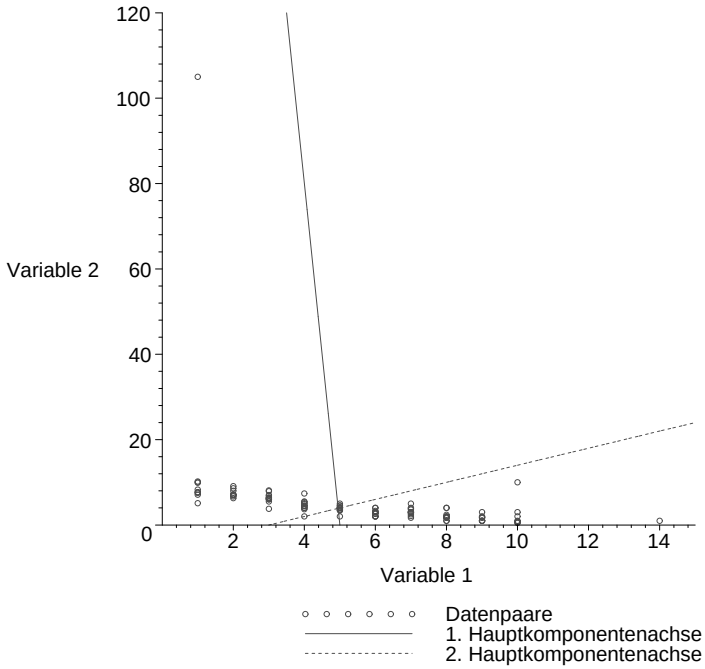

Abbildung 7.1: Klassische Hauptkomponentenanalyse für 109 Beobachtungen

Nachteil der herkömmlichen Hauptkomponentenanalyse: Ausreißer in den Daten führen zu verfälschenden Ergebnissen bei der Bestimmung des Mittelwertes, wenn keine robuste Schätzung vorgenommen wird. Gleiches gilt bei der Berechnung der Kovarianzmatrix. Insgesamt wird so das Ergebnis der Hauptkomponentenanalyse beeinflusst. Daher sollte man eine robuste Hauptkomponentenanalyse durchführen.

7.3 Robuste Hauptkomponentenanalyse

In der Arbeit von J.E. Jackson [10] werden drei unterschiedliche Methoden kurz vorgestellt, die zur Durchführung der robusten Hauptkomponentenanalyse verwendet werden können.

- Eliminiere zunächst Ausreißer aus den Rohdaten und führe anschließend eine konventionelle Hauptkomponentenanalyse auf den verbliebenen Daten durch.

- Benutze eine robuste Schätzung der Kovarianz- bzw. Korrelationsmatrix und führe eine konventionelle Hauptkomponentenanalyse durch.
- Verwende robuste Schätzungen der Eigenwerte und Eigenvektoren von Kovarianzmatrizen.

Ein Verfahren zur Bestimmung robuster Hauptkomponenten auf der Grundlage robuster Schätzungen der Eigenwerte und Eigenvektoren wird in [1] angegeben und soll im Folgenden vorgestellt werden. Die Berechnung erfolgt ebenfalls iterativ nach folgender Iterationsvorschrift:

(1) Berechne Startschätzung V für die Kovarianzmatrix und berechne den ersten Eigenvektor u_1.

(2) Initialisiere $w_m \equiv 1$.

(3) Bestimme die Hauptkomponenten-Scores $y_m = u_1^\top x_m$.

(4) Berechne M-Schätzungen für den Mittelwert und die Varianz zu y_m (wie in Kapitel 5) iterativ:

$$\bar{y} = \frac{\sum w_m^{(neu)} y_m}{\sum w_m^{(neu)}}$$

$$s^2 = \frac{\sum \left(w_m^{(neu)}\right)^2 (y_m - \bar{y})^2}{\sum \left(w_m^{(neu)}\right)^2 - 1}$$

mit

$$w_m^{(neu)} = w_m^{(neu)}(d_m) = \begin{cases} 1 & \text{, falls} \quad d_m \leq d_0 \\ \frac{d_0}{d_m} \exp\left(\frac{-(d_m - d_0)^2}{2 \cdot 1{,}25^2}\right) & \text{, falls} \quad d_m > d_0 \end{cases}$$

und

$$d_m^2 = \frac{(y_m - \bar{y})^2}{s^2}, \qquad d_0 = \sqrt{v} + \sqrt{2}$$

Als Startwerte werden der Median und $(0,74 \cdot \text{IQR})^2$ empfohlen.

(5) Setze $w_m = \min\{w_m, w_m^{(neu)}\}$ und berechne $\bar{x}$ und V mit dem gerade bestimmten w_m

$$\bar{x} = \frac{\sum w_m x_m}{\sum w_m}$$

$$V = \frac{\sum w_m^2 (x - \bar{x})(x - \bar{x})^\top}{\sum w_m^2 - 1} \quad .$$

(6) Berechne den ersten Eingenwert und Eigenvektor u_1 von V.

(7) Wiederhole die Schritte (3) - (6) so lange, bis Konvergenz vorliegt.

Zur Bestimmung der weiteren Hauptkomponenten $u_i, 2 \leq i \leq p$, muss man die Daten auf den Raum orthogonal zu den bisherigen Eigenvektoren $u_1, \ldots, u_{i-1}$ projizieren und anschließend die Schritte (2) bis (7) wiederholen. Als Algorithmus ergibt sich hierfür:

(8) Bilde $x_{i,m} = (I - U_{i-1} U_{i-1}^\top) x_m$, wobei $U_{i-1} = (u_1, \ldots, u_{i-1})$.
Wähle als Startwert für den ersten Eigenvektor den zweiten Eigenvektor der letzten Iteration für den vorherigen Eigenvektor.

(9) Wiederhole Schritt (2) bis (7) mit $x_{i,m}$ anstelle von x_m und bestimme den ersten Eigenvektor u, der dann u_i ist.

Die Schritte (7) bis (9) werden so lange wiederholt, bis alle p Eigenwerte e_i und Eigenvektoren u_i mit entsprechenden Gewichten bestimmt sind.
Schließlich kann eine robuste Schätzung der Kovarianz- bzw. Korrelationsmatrix durch $UEU^\top$ alternativ zu V gefunden werden, die positiv semi-definit ist. Dabei ist U die Matrix, die sich spaltenweise aus den Eigenvektoren u_i zusammensetz und E eine Diagonalmatrix mit den Eigenwerten e_i als Einträge auf der Hauptdiagonalen.

Beispiel 7.3.1. Für den Datensatz A mit 109 Beobachtungen (das Datenpaar $(1,0;105,0)$ wird in der Grafik nicht berücksichtigt) ergibt sich nach Berechnung der robusten Hauptkomponentenanalyse folgendes Bild:

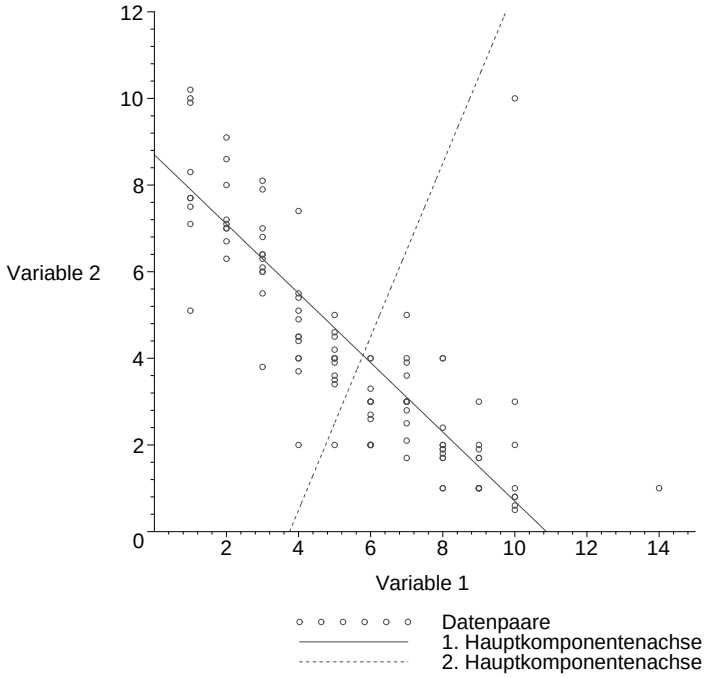

Abbildung 7.2: 108 von 109 Beobachtungen zweier Variablen und beide Hauptkomponentenachsen

Kapitel 8

Zusammenfassung und Ausblick

Im Rahmen dieser Arbeit wurden zu Beginn die Probleme aufgezeigt, die sich ergeben, wenn mit Ausreißern behaftete Daten mit nicht-robusten Verfahren in der Statistik untersucht werden. Verlässliche Aussagen sind dann in der Regel nicht möglich oder man ist gezwungen, vor der Untersuchung die Ausreißer zu eliminieren, um dann die nicht-robusten Verfahren anwenden zu können. Dieses Vorgehen wird insbesondere in der Arbeit von T. Dickhaus ([2]) ausführlich vorgestellt.
Es hat sich ebenfalls gezeigt, dass im univariaten Fall die Benutzung robuster Verfahren keinesfalls zu einem sehr viel größeren Rechenaufwand führt. Darüber hinaus wurden mit Sensitivitätskurve, Einflusskurve und dem Bruchpunkt Kriterien an die Hand gegeben, die es erlauben, die Robustheitseigenschaften eines Schätzers beurteilen zu können. In Bezug auf die L-Schätzer und die M-Schätzer wurden diese Kriterien an den Beispielschätzern gezeigt und Unterschiede deutlich gemacht. Im multivariaten Fall wurden Schätzer für die Lage und für Kovarianz- bzw. Korrelationsmatrizen angegeben und das Programmpaket ROBETH kurz vorgestellt, das auch in dem parallel zu dieser Arbeit entstandenen Computerprogramm seine Anwendung findet. Insbesondere wurden der Huber-Schätzer sowie Andrew‘s Wave-Schätzer und Tukeys Biweight-Schätzer in diesem Programm verwendet. Die Diskussion dieser Schätzmethoden diente damit auch der Vorbereitung zur robusten Hauptkomponentenanalyse. In ihr wird ausgehend von robusten Schätzern für Eigenwerte und Eigenvektoren eine gegenüber Ausreißern unempfindlichere Methode zur Berechnung der Hauptkomponenten im multivariaten Fall entwickelt und ein Algorithmus angegeben, der bereits in Aufgabenstellungen aus der Praxis Anwendung findet.
Im Rahmen der multivariaten Analyse hat sich gezeigt, dass bei separaten Schätzungen der einzelnen Einträge in der Kovarianz- bzw. Korrelationsmatrix diese in der Regel nicht positiv semi-definit sind. Dies kann zu Problemen führen bei weiteren Berechnungen, die auf diesen Matrizen beruhen. Um die positive Definitheit der Matrizen dennoch garantieren zu können, wird in [4] eine Möglichkeit aufgezeigt, mit Hilfe einer Shrinkage-Schätzung positiv semi-definite Matrizen zu erzeugen.
Außerdem gibt es neben den Familien der L- und M-Schätzer noch die Familie der S-Schätzer (s. [15]). Im Gegensatz zu den M-Schätzern, für die bei wachsender Dimension p die Robustheit „nachlässt“, da für den Bruchpunkt $\varepsilon^* = 1/p$ gilt, liegt der Bruchpunkt der S-Schätzer für alle p bei konstant 0,5. Ziel der S-Schätzer ist es, eine Ellipse minimalen Volumens zu finden, so dass 50% der Datenpunkte in ihr liegen. Der Nachteil dieser Vorgehensweise liegt in der schweren numerischen Berechenbarkeit und es bleibt abzuwarten, ob in Zukunft numerische Verfahren entwickelt werden, die eine effiziente Berechnung ermöglichen. Ebenfalls interessant ist in diesem Zusammenhang die Klasse der P-Schätzer (s. [13]). Sie basieren auf der Überlegung, die Daten so zu transformieren, dass die Streuung in alle Richtungen gleich ist. Eine solche Transformation existiert jedoch nicht, so dass man dazu übergeht, die Streuung so klein wie möglich zu bestimmen. Ebenfalls ist hierbei die numerische Berechnung problematisch und erfordert zur Zeit noch einen zu hohen Aufwand.
Insgesamt kann man daher feststellen, dass die Klasse der M-Schätzer die momentan geeignete Methode ist, Lage-und Streuungsparameter im univariaten und multivariaten Fall robust zu schätzen,

auch wenn der Bruchpunkt als ein Kriterium der Robustheit mit wachsender Dimension abnimmt und man beste Ergebnisse nur für den Fall $p = 2$ erzielen kann.

Anhang A

Verwendete Symbole

Symbole	Bedeutung
$[\cdot]$	Gaußklammer
$\lvert M \rvert$	Mächtigkeit der Menge M
$\text{sgn}(\cdot)$	Vorzeichenfunktion
x^T	transponierter Vektor bzw. Matrix
$\mathbb{E}(\cdot)$	Erwartungswertoperator
$\text{Var}(\cdot)$	Varianzoperator
$\mathbb{P}(\cdot)$	Wahrscheinlichkeitsoperator
$\Phi(\cdot)$	Verteilungsfunktion der Standardnormalverteilung $\mathcal{N}(0,1)$
$\phi(\cdot)$	Dichte der Standardnormalverteilung $\mathcal{N}(0,1)$
I_p	p-dimensionale Einheitsmatrix
$1_A(\cdot)$	Indikatorfunktion auf der Menge A

Tabelle A.1: Übersicht der verwendeten Symbole

Anhang B

Verwendeter Datensatz

B.1 Datensatz A

Nr.	Variable 1	Variable 2	Nr.	Variable 1	Variable 2	Nr.	Variable 1	Variable 2
1	6,0	2,0	38	4,0	5,1	74	4,0	4,4
2	7,0	3,0	39	7,0	3,0	75	9,0	3,0
3	6,0	3,0	40	2,0	8,6	76	8,0	2,0
4	3,0	6,0	41	3,0	6,3	77	7,0	2,5
5	4,0	4,9	42	3,0	6,4	78	3,0	3,8
6	6,0	4,0	43	3,0	5,5	79	8,0	1,7
7	5,0	5,0	44	1,0	7,1	80	9,0	1,0
8	8,0	1,9	45	9,0	1,7	81	6,0	3,0
9	8,0	1,9	46	5,0	2,0	82	5,0	4,0
10	7,0	2,1	47	3,0	6,8	83	8,0	4,0
11	5,0	4,2	48	9,0	1,0	84	7,0	4,0
12	4,0	4,0	49	8,0	4,0	85	14,0	1,0
13	2,0	7,2	50	9,0	1,0	86	10,0	2,0
14	1,0	8,3	51	6,0	2,6	87	10,0	3,0
15	1,0	10,0	52	5,0	3,5	88	9,0	2,0
16	2,0	7,0	53	6,0	2,7	89	4,0	5,5
17	2,0	6,3	54	7,0	3,6	90	8,0	1,8
18	3,0	7,0	55	5,0	3,9	91	2,0	8,0
19	5,0	4,0	56	7,0	3,9	92	8,0	1,7
20	1,0	9,9	57	10,0	0,8	93	7,0	5,0
21	8,0	1,0	58	4,0	4,5	94	3,0	7,9
22	8,0	2,0	59	4,0	4,5	95	2,0	9,1
23	8,0	1,0	60	9,0	1,7	96	2,0	7,1
24	4,0	4,0	61	2,0	7,0	97	6,0	3,3
25	7,0	3,0	62	1,0	7,7	98	5,0	4,5
26	3,0	6,1	63	3,0	6,0	99	2,0	6,7
27	8,0	2,4	64	4,0	2,0	100	5,0	4,6
28	10,0	0,5	65	4,0	5,4	101	10,0	0,6
29	4,0	7,4	66	7,0	2,8	102	9,0	1,9
30	9,0	1,0	67	4,0	3,7	103	3,0	8,1
31	6,0	2,0	68	5,0	3,6	104	5,0	3,4
32	6,0	2,0	69	6,0	2,0	105	1,0	10,2
33	6,0	4,0	70	10,0	10,0	106	1,0	7,7
34	10,0	1,0	71	5,0	4,0	107	1,0	5,1
35	1,0	105,0	72	7,0	3,0	108	1,0	7,5
36	7,0	1,7	73	10,0	0,8	109	3,0	6,4
37	6,0	3,0						

Tabelle B.1: Datensatz aus Industrieprojekt

Literaturverzeichnis

[1] Campbell, N.A. (1980)
Robust Procedures in Multivariate Analysis - I: Robust Covariance Estimation
Journal of Applied Statistics **29**, No. 3, Seite 231-237

[2] Dickhaus, T. (2003)
Statistische Verfahren für das Data Mining in einem Industrieprojekt
Interner Bericht FZJ-ZAM-IB-2003-08, ZAM Forschungszentrum Jülich

[3] Dikta, G. (2002)
Operation Research
Manuskript zur Vorlesung
FH Aachen, Abteilung Jülich

[4] Gnanadesikan, R. (1997)
Methods for Statistical Data Analysis of Multivariate Observations
Wiley, New York

[5] Hampel, F.R. (1974)
The Influence Curve and Its Role in Robust Estimation
Journal of the American Statistical Association **69**, Nr. 346, Seite 383-393

[6] Hampel, F.R., Ronchetti, E.M., Rousseeuw P.J., Stahel W.A. (1986)
Robust statistics: The approach based on influence functions
Wiley, New York

[7] Hartung, J., Elpelt, B. (1995)
Statistik - Lehr- und Handbuch der angewandten Statistik
R. Oldenbourg Verlag, München

[8] Hoaglin, D.C., Mosteller, F., Tukey, J.W. (1983)
Understanding Robust and Exploratory Data Analysis
Wiley, New York

[9] Huber, P.J. (1981)
Robust statistics
Wiley, New York

[10] Jackson, J.E. (1991)
A User‘s Guide To Principal Components
Wiley, New York

[11] Jolliffe, I.T. (1986)
Principal Component Analysis
Springer-Verlag, New York

[12] Lehn J., Wegmann H. (1992)
Einführung in die Statistik
B.G.Teubner Verlag, Stuttgart

[13] Maronna, R.A., Stahel, W.A., Yohai, V.J. (1992)
Bias-robust estimators of multivariate scatter based on projections
Journal of Multivariate Analysis, **42**, Seiten 141-161

[14] Marazzi, A. (1993)
Algorithms, Routines, and S Functions for Robust Statistics
The FORTRAN Library ROBETH with an Interface to S-Plus
Wadsworth & Brooks/Cole Publishing Company

[15] Rousseeuw, P.J., Leroy, A.M. (1987)
Robust regression and outlier detection
Wiley, New York

[16] Serfling, R.J. (1980)
Approximation Theorems of Mathematical Statistics
Wiley, New York