Bruchpunkt und Bias zur Beurteilung multivariater Ausreißeridentifizierung

Dissertation

zur Erlangung des Grades
eines Doktors der Naturwissenschaften
der Universität Dortmund

Dem Fachbereich Statistik der Universität Dortmund
vorgelegt von

Claudia Becker

aus Haan

Dortmund 1996

Gutachter: Prof. Dr. U. Gather

Prof. Dr. W. Krämer

Tag der mündlichen Prüfung: 18.12.1996

Inhaltsverzeichnis

1 Einleitung

Die Auswirkung von Ausreißern auf die statistische Analyse von Datensätzen ist schon häufig beschrieben worden. Ausreißer, das heißt, Beobachtungen, die „weit entfernt" von der Hauptmasse der Daten liegen und unter Umständen nicht dem für die Daten unterstellten Modell gehorchen, während der größte Teil der Beobachtungen dies tut, können zu beträchtlichen Verfälschungen von Auswertungen führen. Insbesondere Teststatistiken und Schätzer der klassischen parametrischen Statistik wie zum Beispiel das arithmetische Mittel als Lageschätzer sind für derartige Verfälschungen anfällig. Diese Tatsache motiviert die Verwendung robuster statistischer Prozeduren, die auch bei Modellabweichungen und dem Auftreten von Ausreißern noch zufriedenstellende Ergebnisse liefern. Umfangreiche Beiträge zu diesem Aspekt des Umgangs mit Ausreißern findet man bei Huber (1972), Bickel (1976), Launer, Wilkinson (1979), Huber (1981), Hampel et al. (1986), Rousseeuw, Leroy (1987), Staudte, Sheater (1990), Stahel, Weisberg (1991 a,b) sowie Morgenthaler et al. (1993). Ein weiterer, in diesem Zusammenhang häufig verwendeter Ansatz besteht darin, die untersuchten Daten von Ausreißern zu bereinigen und anschließend klassische Verfahren anzuwenden.

Mit der Identifizierung von Ausreißern wird daher oft die Vorstellung verbunden, man suche diese Beobachtungen nur, um sie aus einem Datensatz zu entfernen und damit einer Beeinträchtigung der Auswertung zuvorzukommen. Dies muß jedoch keineswegs der Fall sein. Gerade die Ausreißer können von grundlegender Bedeutung sein, wie ein in der Zeitschrift DER SPIEGEL erschienener Artikel verdeutlicht (Neffe (1993)). Dieser Artikel beschäftigt sich mit HIV–infizierten Personen, bei denen die Krankheit AIDS noch nicht zum Ausbruch gekommen ist, obwohl die Infektion schon sehr lange (bis zu 14 Jahren) besteht. In diesem Fall sind die Überlebenden die Ausreißer, und gerade sie sind diejenigen, für die man sich interessiert. „Denn sollte es gelingen, in ihren Körpern ein physiologisches Pendant zu ihrer guten Konstitution zu finden, eine genetische, biochemische oder immunologische Erklärung für ihre Ausnahmeerscheinung – die bloßen Zahlen und Tabellen erhielten plötzlich einen anderen Sinn" (Neffe (1993), S. 209).

Die Identifizierung von Ausreißern ist daher nicht nur als Hilfsmittel zur „Säuberung" von Daten zu betrachten, sondern auch als ein Werkzeug zum Aufspüren von Besonderheiten, also durchaus als Selbstzweck.

Im gerade betrachteten Beispiel ist die Entdeckung der Ausreißer relativ leicht, weil zunächst nur eine Variable (Zeitraum seit der Infektion, in dem AIDS nicht ausgebrochen ist) für die Abweichung der Beobachtungen von den übrigen verantwortlich ist. In multivariaten Problemstellungen ist die Situation allerdings oft nicht so eindeutig. Beobachtungen, die in keiner der einzelnen Variablen als auffällig anzusehen sind, können trotzdem in der Kombination mehrerer Variablen beträchlich von allen anderen Beobachtungen abweichen. Ein Beispiel hierfür findet man bei Rousseeuw, Leroy (1987, S. 57). Dort sind das Hirngewicht (in Gramm) und das Körpergewicht (in Kilogramm) von 28 ausgewählten Tierarten aufgeführt, zusammengestellt aus größeren Datensätzen in Jerison (1973) und Allison, Cicchetti (1976). Trägt man die logarithmierten Daten gegeneinander ab, so sieht man recht deutlich (vgl. Abbildung 1.1), daß bei gemeinsamer Betrachtung der beiden Variablen einige Beobachtungen vom Rest der Daten separiert sind, obwohl sie in keiner der beiden einzelnen Komponenten wesentlich herausragen.

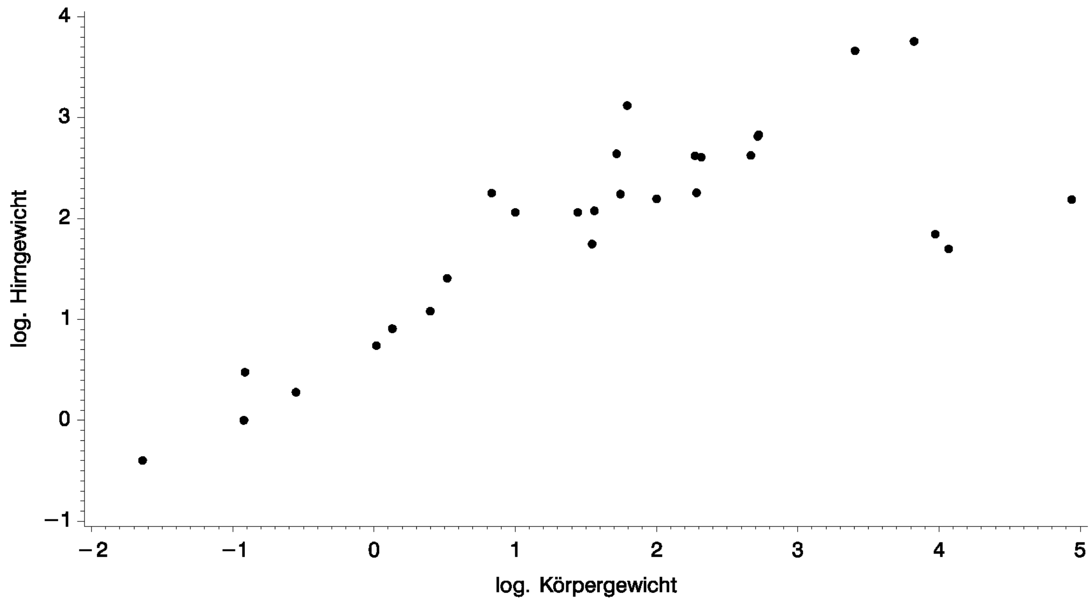
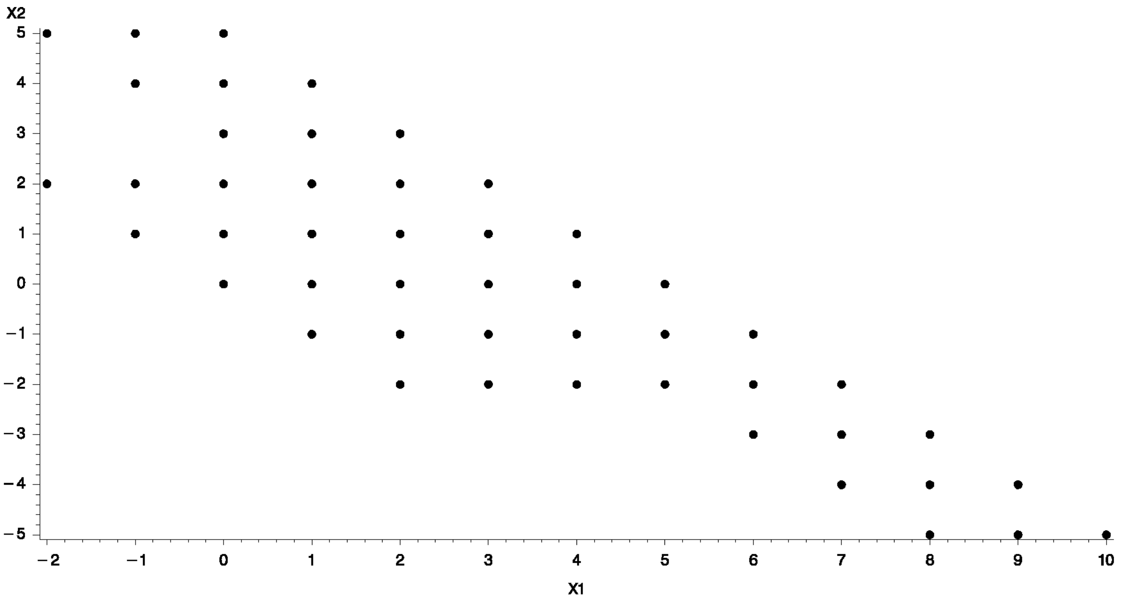


Abbildung 1.1: Logarithmiertes Körper- und Hirngewicht von 28 Spezies

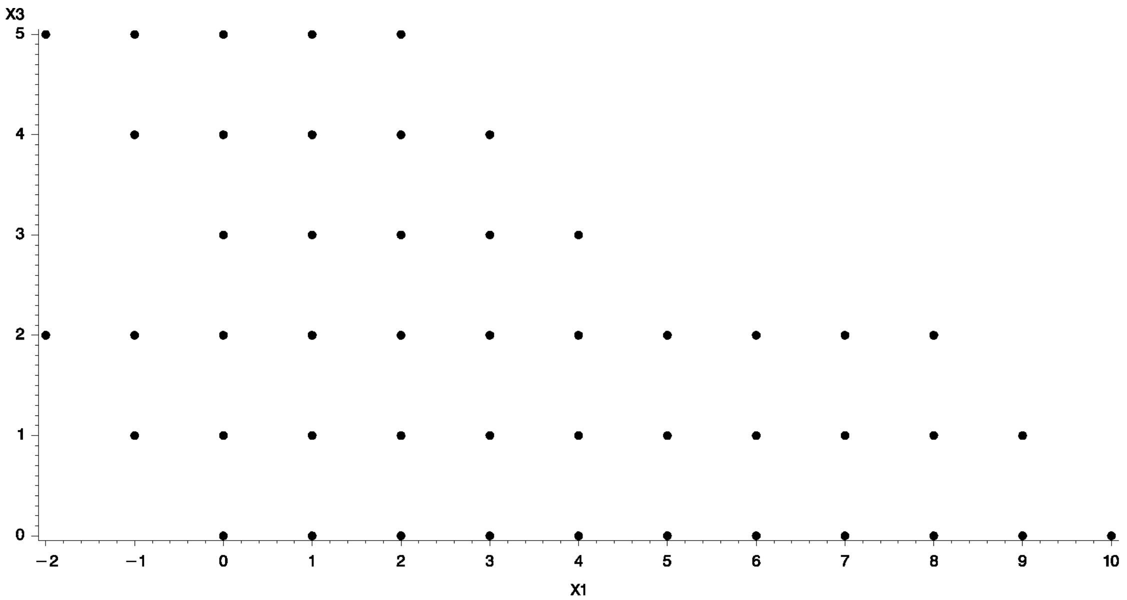
Tabelle 1.1: Simulierter Datensatz mit 49 Beobachtungen in den Variablen X1, X2, X3

X1	X2	X3	X1	X2	X3	X1	X2	X3	X1	X2	X3	X1	X2	X3
5	0	0	4	1	0	8	-5	2	2	2	1	0	0	5
9	-5	1	-1	1	5	0	3	2	2	-2	5	10	-5	0
2	0	3	1	2	2	6	-2	1	3	-1	3	0	1	4
-2	5	2	2	3	0	1	-1	5	6	-1	0	1	3	1
0	5	0	8	-3	0	-1	5	1	4	-2	3	-2	2	5
7	-4	2	4	0	1	2	1	2	5	-1	1	0	2	3
9	-4	0	1	0	4	6	-3	2	1	1	3	8	-4	1
-1	4	2	3	2	0	4	-1	2	0	4	1	5	-2	2
3	1	1	1	4	0	2	-1	4	3	0	2	3	-2	4
7	-3	1	-1	2	4	7	-2	0	0	0	0			

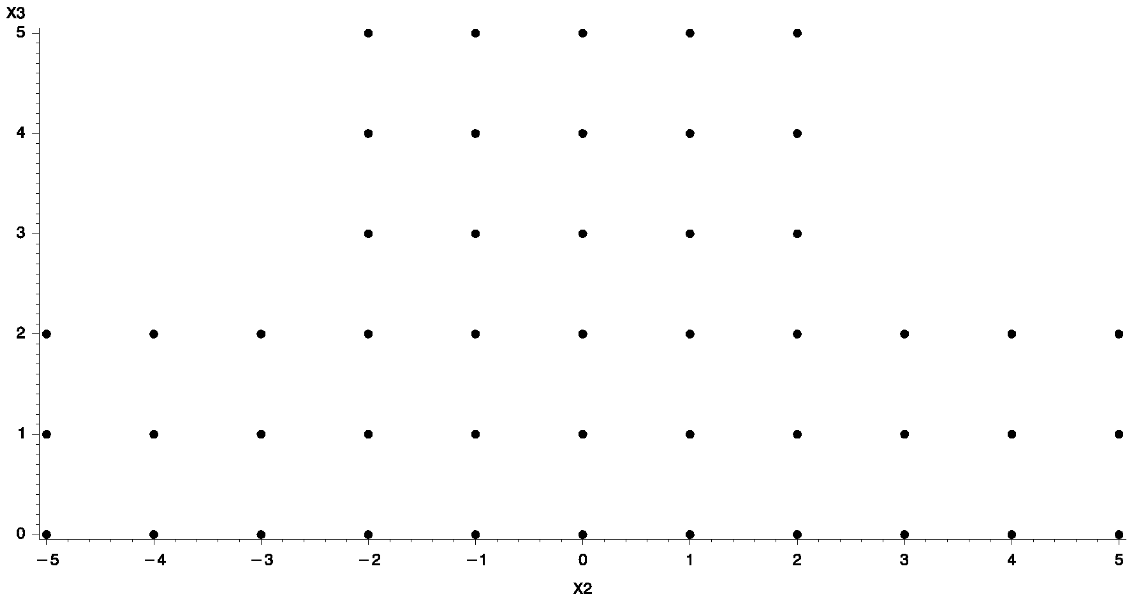
Dieses Problem pflanzt sich in höheren Dimensionen entsprechend fort. Schon bei drei Variablen, wo eine Scatterplotdarstellung noch möglich ist, ist es nicht mehr einfach, eine Beobachtung aufzuspüren, die nur in allen Variablen gemeinsam auffällig ist. Tabelle 1.1 enthält einen künstlich erzeugten Datensatz mit 49 Beobachtungen von drei Merkmalen. Er besteht aus 48 Punkten, die auf der Ebene durch die Koordinatenpunkte (5,0,0), (0,5,0), (0,0,5) liegen, und zusätzlich dem Punkt (0,0,0), der nicht auf dieser Ebene und mit klarem Abstand zu dieser liegt. Weder bei Betrachtung der Werte in den einzelnen Variablen noch beim Blick auf die zweidimensionalen Scatterplots (siehe Abbildung 1.2 a)-c)) entdeckt man die Auffälligkeit des zusätzlichen Punktes. Erst bei gemeinsamer Darstellung aller drei Variablen in einem dreidimensionalen Scatterplot fällt die Beobachtung (0,0,0) deutlich auf (vgl. Abbildung 1.3). Allerdings hat sich der Aufwand, diese Beobachtung zu entdecken, im Vergleich zum Aufwand bei einem zweidimensionalen Datensatz beträchtlich erhöht, da in der dreidimensionalen Darstellung zusätzlich ein geeigneter Blickwinkel auf die Punktwolke gefunden werden muß.


X2
5
4
3
2
1
0
−1
−2
−3
−4
−5
−2
−1
0
1
2
3
4
5
6
7
8
9
10
X1

1.2 a)


X3
5
4
3
2
1
0
−2
−1
0
1
2
3
4
5
6
7
8
9
10
X1

1.2 b)



1.2 c)

Abbildung 1.2: Zweidimensionale Scatterplots für die Daten aus Tabelle 1.1

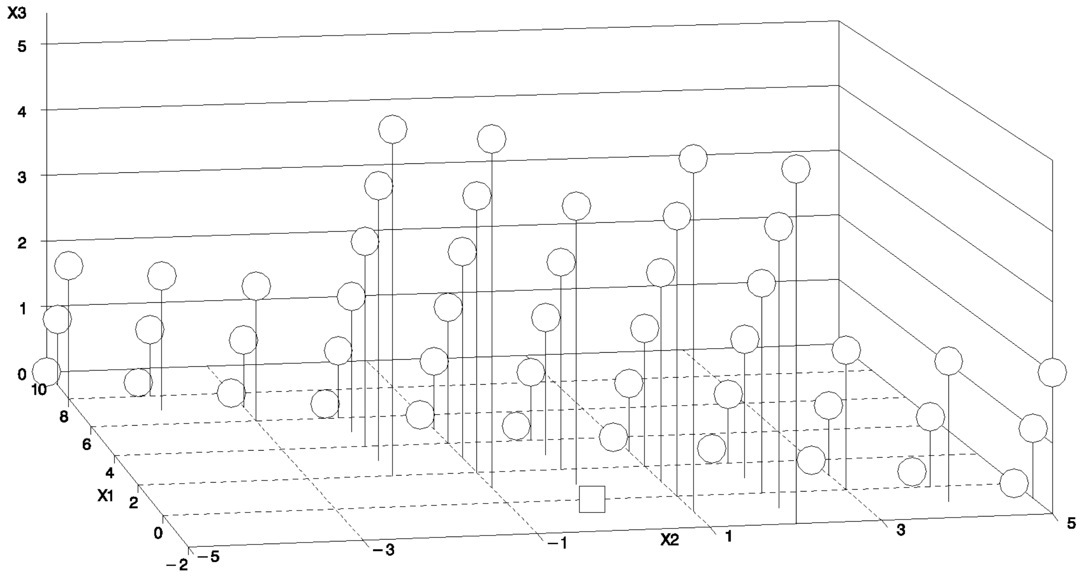


Abbildung 1.3: Dreidimensionaler Scatterplot der Beobachtungen aus Tabelle 1.1; der Punkt (0,0,0) ist durch ein Quadrat markiert

Ähnliche Situationen wie in den beiden hier behandelten Beispielen kann man sich ebenso für höhere Dimensionen vorstellen. Bei Rückzug auf die einfachen zweidimensionalen Projektionen ist es dann möglich, daß Auffälligkeiten nicht entdeckt werden. Beispiele dieser Art machen deutlich, daß es sinnvoll ist, Methoden zur Identifikation von Ausreißern in multivariaten Datensätzen zu erarbeiten. Da die Probleme der Ausreißer–Identifizierung schon in multivariaten Situationen relativ geringer Dimension nicht mehr einfach handhabbar sind, beschränkt sich die vorliegende Arbeit auf die Entwicklung von Strategien für solche Situationen. Für Probleme in höheren Dimensionen, das heißt, ab etwa 6 bis 7 betrachteten Variablen, muß zusätzlich der sogenannte Fluch der Dimension (engl. „curse–of–dimensionality") berücksichtigt werden, der insbesondere bei Regressionsverfahren und der Analyse von Strukturen in Daten bekannt ist (Bellman (1961), Friedman (1994)). In höheren Dimensionen ist es nämlich nicht mehr möglich, den Stichprobenraum selbst mit einer großen Stichprobe ausreichend dicht zu besetzen, um vernünftige Schätzer für die gesuchten Parameter zu finden. Dieser besonderen Schwierigkeit müßte mit anderen als den hier vorgestellten Methoden begegnet werden.

Es gibt bereits eine Reihe von Verfahren, mit deren Hilfe die Identifizierung von Ausreißern möglich ist. Sie lassen sich grob in zwei Kategorien einteilen, in die sogenannten Ausreißertests und in heuristische, häufig graphische Methoden. Das folgende Kapitel gibt einen Überblick über beide Methodentypen und stellt einige ausgewählte Verfahren kurz dar. In der Kategorie der Ausreißertests liegt der Schwerpunkt auf Methoden, die auf dem von Wilks (1963) definierten „scatter ratio" basieren. Im zweiten Teil werden drei ganz verschiedenartige graphische Techniken zur Identifizierung von Ausreißern näher vorgestellt, aufbauend auf Quantil–Quantil–Plots, Bäumen und Gesichter–Darstellungen von Beobachtungen. Als ein Verfahren, das Aspekte beider Kategorien beinhaltet, wird die Entdeckung von Ausreißern mittels einer Identifikationsregel, eines sogenannten Ausreißer–Identifizierers, vorgestellt. Dieser Ansatz, der eine Verallgemeinerung eines univariaten Verfahrens aus Davies, Gather (1993) darstellt, bildet die Grundlage für die weiteren Kapitel.

Für die Beurteilung von Identifizierungsverfahren werden geeignete Kriterien benötigt. Obwohl eine Reihe solcher Kriterien zum Vergleich von Identifizierungsregeln benutzt wird, gibt es einige Kritikpunkte, die zu bedenken sind. Viele Arbeiten, in denen Identifizierungsprozeduren beurteilt werden, gründen sich auf Simulationsstudien. Hier werden zwar Kriterien an solche Prozeduren festgelegt, die Ergebnisse der Beurteilungen hängen aber in erheblichem Maße vom jeweils verwendeten Simulationsdesign ab. In anderen Fällen werden oft die Identifizierungsregeln nicht als solche bewertet. So ist es bei der Konstruktion von Ausreißertests üblich, unter bestimmten Modellannahmen gleichmäßig beste oder lokal optimale Tests herzuleiten, die gewisse Niveau- bzw. Machtbedingungen erfüllen (z.B. Schwager, Margolin (1982)). Das Kriterium, das eine „gute“ Identifizierungsprozedur, die auf einem solchen Ausreißertest beruht, hier erfüllen muß, ist also eigentlich eine Forderung an die Qualität des Tests und nicht an das Identifizierungsverfahren. Ähnlich ist die Situation auch bei einer anderen Art des Vergleichs. Wenn in Identifizierungsverfahren Teststatistiken oder Schätzer verwendet werden, so wird oft das Verhalten dieser Statistiken miteinander verglichen, nicht jedoch das Verhalten der gesamten Identifizierungsregeln.

Die Ausreißer–Identifizierungsprozedur als solche wird also bisher häufig nicht bewertet. Zwar werden die zwei möglichen Arten von Fehlentscheidungen – Masking und Swamping – in der Regel erwähnt, ohne daß jedoch die Größe des jeweiligen Effekts in irgendeiner Weise quantifiziert wird.

Der sogenannte Swamping–Effekt bedeutet, daß durch den Einfluß von ungünstig gelegenen Ausreißern in einer Stichprobe Beobachtungen, die bezüglich des zugrundegelegten Modells keine Ausreißer sind, trotzdem als solche identifiziert werden. Beim sogenannten Masking–Effekt werden vorhandene Ausreißer nicht erkannt.

Erste Ansätze zur Quantifizierung des Masking–Effekts bei Ausreißertests findet man bei Bendre und Kale (1985, 1987). In der Arbeit von Davies und Gather (1993) werden Gütekriterien entwickelt, die die Größe von Masking- und Swamping–Effekt bei univariaten Identifizierern bewerten. Im dritten Kapitel der vorliegenden Arbeit werden die entsprechenden Kriterien, Masking- und Swamping–Bruchpunkt, für multivariate Ausreißer–Identifizierer vorgestellt. Diese Bruchpunkte spiegeln maximale Anteile von Ausreißern wider, die ein Identifizierer noch ohne Schwierigkeiten verkraften kann. In den Abschnitten 3.2 und 3.3 wird gezeigt, daß beide Bruchpunkte eng mit den Finite–sample Bruchpunkten der Schätzer zusammenhängen, auf denen die Identifizierungsprozedur beruht. Genauer wird durch die Bruchpunkte der Lokations- und Kovarianzschätzer, die einen Identifizierer festlegen, seine Anfälligkeit für Masking- und

Swamping–Effekt bestimmt. Damit wird beschrieben, wie sich das Verhalten verwendeter Statistiken auf das Verhalten von Identifizierungsprozeduren auswirkt. Auf diese Weise ist eine Bewertung solcher Prozeduren möglich. Verfolgt man bei der Konstruktion eines Ausreißer–Identifizierers das Ziel möglichst hoher Masking- und Swamping–Bruchpunkte, so führen die Resultate der Abschnitte 3.2 und 3.3 zu der Empfehlung, für einen Identifizierer Schätzer mit möglichst hohem Finite–sample Bruchpunkt zu verwenden. Dabei muß der eingesetzte Lokationsschätzer diesbezüglich mindestens so gut sein wie der Kovarianzschätzer. Es zeigt sich, daß sowohl die auf dem Minimum–Volume–Ellipsoid beruhenden Schätzer als auch S–Schätzer mit maximalem Bruchpunkt in diesem Sinne geeignet sind, wobei S–Schätzer wegen ihres besseren asymptotischen Verhaltens vorzuziehen sind.

Über das Bruchpunktverhalten hinaus interessiert man sich auch für das Verhalten von Ausreißer–Identifizierern bei Stichproben mit einem festen, eventuell kleineren, Anteil schlechter Beobachtungen. Diesem Interesse wird das Gütekriterium des maximalen asymptotischen Bias gerecht, mit dem sich das vierte Kapitel befaßt. Der Bias wird sowohl für einen Ausreißer–Identifizierer als auch für die ihn bestimmenden Schätzer definiert. Der Identifizierer wird dabei aufgefaßt als Schätzer für einen unbekannten Ausreißerbereich. Der maximale asymptotische Bias beschreibt die größtmögliche Abweichung eines Schätzers von seinem „Ziel“, die durch einen festen Anteil von Ausreißern in einer Stichprobe erreicht werden kann. Wie schon bei den Bruchpunkten, lassen sich auch in diesem Fall enge Beziehungen zwischen den Biasgrößen herstellen. So wird in den Sätzen 4.1 und 4.2 gezeigt, daß die Beschränktheit der Biaswerte der verwendeten Lokations- und Kovarianzschätzer stets notwendig dafür ist, daß der resultierende Identifizierer ebenfalls einen beschränkten Bias besitzt. Um zu sichern, daß diese Voraussetzung auch hinreichend für einen beschränkten Bias der Identifizierungsprozedur ist, muß eine zusätzliche Forderung an die eingehenden Schätzer gestellt werden (Satz 4.3). Es muß sich um konsistente Schätzer mit geeigneter Konvergenzgeschwindigkeit handeln ($\sqrt{N}$–Konsistenz). Als ein weiteres Ergebnis läßt sich festhalten, daß der maximale asymptotische Bias eines Ausreißer–Identifizierers ebenfalls eng mit den Bruchpunkten der in ihn eingehenden Schätzer verknüpft ist (Sätze 4.4 und

4.5). Daraus resultiert insgesamt die Empfehlung, für Identifizierer, wie sie in dieser Arbeit vorgestellt werden, Schätzer zu verwenden, die einerseits hochrobust sind, andererseits jedoch noch eine genügend hohe Konsistenzordnung besitzen. Die auf dem Minimum–Volume–Ellipsoid beruhenden Schätzer sind damit für solche Identifizierer nicht verwendbar, wenn sowohl gutes Masking- und Swamping–Verhalten als auch ein beschränkter Bias der Prozedur gefordert sind. Dagegen erweisen sich S–Schätzer als geeignet (Korollar 4.4).

Das fünfte Kapitel beinhaltet die praktische Anwendung der in dieser Arbeit behandelten Identifizierer. Anhand verschiedener Datensätze, die im Zusammenhang mit der Entdeckung von Ausreißern bereits untersucht worden sind, wird überprüft, inwieweit sich die mit anderen Methoden erhaltenen Ergebnisse mit den hier vorgestellten Verfahren reproduzieren und verbessern lassen.
Es werden zwei Identifizierer betrachtet, die auf S–Schätzern beruhen und damit bezüglich der in dieser Arbeit erhaltenen Ergebnisse als geeignete Verfahren gelten. Es handelt sich dabei um ein Verfahren, das auf Tukeys Biweight–Funktion beruht, und um eine Modifikation desselben. Für das Verhalten der Identifizierer spielt die Normierung der Methoden durch ein toleriertes Fehlerniveau eine wichtige Rolle. Eine mögliche Normierung besteht in der Vorgabe der Wahrscheinlichkeit, mit der in einem Datensatz, der ausschließlich aus normalverteilten Beobachtungen besteht, fälschlicherweise eine Beobachtung als Ausreißer identifiziert wird. Wählt man diese Wahrscheinlichkeit eher klein (5%), so können robuste Identifizierer nicht alle Ausreißer erkennen. Bei etwas großzügigerer Normierung (Fehlerwahrscheinlichkeit 10%) werden die Ergebnisse anderer Methoden reproduziert und in einigen Fällen verbessert.
Mit der Anwendung eines Ausreißer–Identifizierers wird in der Regel das Ziel verfolgt, Auffälligkeiten in Datensätzen ausfindig zu machen und nicht nur extrem abweichende Beobachtungen zu entdecken. Diese Überlegung rechtfertigt die Normierung auf 10%, da eine Normierung auf 5% für dieses Ziel als zu restriktiv erkannt wird.
Bei den im fünften Kapitel betrachteten Beispielen zeigt sich der auf Tukeys Biweight–Funktion basierende Identifizierer seiner modifizierten Version überlegen. Die Ergebnisse der betrachteten Beispiele legen die Empfehlung nahe, den auf der modifizierten

Biweight–Funktion basierenden Identifizierer nicht zu verwenden, wenn im Verhältnis zur Dimension der Daten ein relativ geringer Stichprobenumfang vorliegt. Unterstützt wird dies durch die Resultate einer Simulation, in der für beide Regeln mittlere Größen des jeweils größten nicht identifizierten Ausreißers bestimmt werden. Auch hier schneidet der auf Tukeys Biweight beruhende Identifizierer besser ab als seine Modifikation, wenn der Stichprobenumfang im Verhältnis zur Dimension eher gering ist.

Ein kurzer Ausblick auf weitere interessierende Aspekte bildet den Abschluß der Arbeit.

2 Methoden zur Identifizierung von Ausreißern

Die Identifizierung von Ausreißern in einem Datensatz ist nur sinnvoll bei Zugrundelegung eines Modells, nach dem sich „gute" Beobachtungen verhalten sollen. Eine Beobachtung kann immer nur relativ zu einem betrachteten Modell als Ausreißer angesehen werden. Dabei kann ein solches Modell genau spezifiziert oder eher vage formuliert sein, etwa „die Beobachtungen entstammen einer Normalverteilung mit Erwartungswertvektor $\underline{0}$ und Kovarianzmatrix $\mathcal{I}$" oder aber „die Beobachtungen bilden eine homogene Punktwolke, sie sind alle ähnlich" (vgl. Choudhury, Das (1992), S. 92, „An outlier can be defined to be an abnormal item among a group of otherwise similar items."). Die aus der genauen Modellspezifikation resultierende Methodik zur Entdeckung von Ausreißern besteht häufig in der Anwendung sogenannter Ausreißertests. Die eher vage Modellvorstellung führt in der Regel zur Benutzung heuristischer Methoden, die oft mit graphischen Veranschaulichungen verbunden sind („The study of outliers in a data set is often inevitably an informal screening process preceeding fuller more formal analysis of the data. As such we must expect to encounter [...] a plethora of ad hoc approaches. Amongst these, there are many graphical methods of highlighting outliers.", Barnett, Lewis (1994), S. 41). Ein Verfahren, das Aspekte beider Arten von Modellspezifikationen vereint, ist die Anwendung von Ausreißer–Identifizierern. Drei verschiedene Ansätze zur Entdeckung von Ausreißern werden im folgenden vorgestellt.

2.1 Ausreißertests

Unter den Begriff „Ausreißertest" fallen zwei verschiedene Arten von Prozeduren. Bei der einen Art wird lediglich die Fragestellung untersucht, ob sich die Beobachtungen eines Datensatzes gemäß eines spezifizierten Nullmodells verhalten oder ob überhaupt Ausreißer in den Daten vorhanden sind. Beispiele hierzu findet man etwa in der grundlegenden Schlüsselarbeit von Schwager und Margolin (1982), im darauf aufbauenden Artikel von Sinha (1984) oder bei Naik (1990). Solche Tests machen keine Aussage darüber, welche der Beobachtungen als Ausreißer erkannt worden sind, wenn das Vorhandensein von Ausreißern festgestellt wird. Sie sind daher im Sinne der Identifizierung von Ausreißern nicht direkt anwendbar und bleiben im folgenden unberücksichtigt.

Prozeduren, die sich zur Identifizierung von Ausreißern eignen, benötigen im wesentlichen

- die Festlegung eines Nullmodells, dem die Beobachtungen gehorchen, falls keine Ausreißer vorliegen,
- die Spezifizierung eines Alternativmodells, das das Auftreten eines oder mehrerer Ausreißer beschreibt, und
- die Entscheidung für oder gegen das Alternativmodell, wobei mit einer Entscheidung für die Alternative die Benennung einer oder mehrerer Beobachtungen einhergeht, die als Ausreißer erkannt werden.

Dabei muß das Alternativmodell strenggenommen als System verschiedener Alternativen angesehen werden, so daß die hier betrachteten Ausreißertests tatsächlich Mehrentscheidungsverfahren sind (vgl. das folgende Beispiel). Von der ursprünglichen Zielsetzung her handelt es sich bei derartigen Ausreißertests nicht um Identifizierungsverfahren für Ausreißer. Vielmehr wird untersucht, ob alle Beobachtungen in einer Stichprobe Realisationen aus einem Nullmodell sind oder ob eine oder mehrere aus bestimmten anderen Verteilungen stammen. Im strengen Sinn wird also untersucht, ob Beobachtungen aus sogenannten kontaminierenden Verteilungen vorliegen, die im Alternativmodell spezifiziert werden. Dabei werden aber in der Regel diese Verteilungen so gewählt, daß unter dem Alternativmodell das Auftreten „extremer" Beobachtungen wahrscheinlicher ist als unter dem Nullmodell. Wird bei Durchführung des Tests entschieden, daß eine oder mehrere Beobachtungen aus einer der angegebenen kontaminierenden Verteilungen stammen, so werden diese Beobachtungen zugleich auch als Ausreißer identifiziert. Zur Illustration dient das folgende Beispiel (Barnett, Lewis (1994), S. 284 ff.).
Betrachtet wird das Nullmodell unabhängiger, identisch normalverteilter Zufallsvektoren $\underline{X}_1, \ldots, \underline{X}_N$, $\underline{X}_i \in I\!R^p, i = 1, \ldots, N$:

$$H_0 : \underline{X}_1, \ldots, \underline{X}_N \text{ i.i.d., } \underline{X}_i \sim N(\underline{\mu}, \Sigma), \; i = 1, \ldots, N.$$

Die Daten werden untersucht auf das Vorhandensein eines Ausreißers, der aus einer Normalverteilung mit derselben Kovarianzmatrix, jedoch mit verschobenem Mittel-wert, stammt. Das Alternativmodell (sogenannte mean slippage–Alternative gemäß Ferguson (1961)) hat die Form

$$\begin{aligned} H_1:\ & \underline{X}_1,\ldots,\underline{X}_N \text{ stochastisch unabhängig und normalverteilt mit} \\ & \mathrm{E}(\underline{X}_i) = \underline{\mu} + \underline{a},\ \underline{a} \neq \underline{0}, \quad \text{für ein } i, \\ & \mathrm{E}(\underline{X}_j) = \underline{\mu} \quad \forall j \neq i, \\ & \mathrm{Cov}(\underline{X}_j) = \Sigma \quad \forall j = 1,\ldots,N. \end{aligned}$$

Welche der Beobachtungen eines Datensatzes darauf untersucht wird, ob sie ein Ausreißer ist, wird mit Hilfe des sogenannten „one outlier scatter ratio" R_j festgelegt (Wilks (1963), S. 409). Die Größe R_j ist für eine Beobachtung $\underline{x}_j$ definiert durch

$$R_j = \frac{\det(A^{(j)})}{\det(A)}, \text{ wobei } A = \sum_{k=1}^{N}(\underline{x}_k - \overline{\underline{x}})(\underline{x}_k - \overline{\underline{x}})^T, \quad \overline{\underline{x}} = \frac{1}{N}\sum_{k=1}^{N}\underline{x}_k,$$

$$A^{(j)} = \sum_{\substack{k=1\\k\neq j}}^{N}(\underline{x}_k - \overline{\underline{x}}^{(j)})(\underline{x}_k - \overline{\underline{x}}^{(j)})^T \text{ und } \overline{\underline{x}}^{(j)} = \frac{1}{N-1}\sum_{\substack{k=1\\k\neq j}}^{N}\underline{x}_k,\ j = 1,\ldots,N.$$

Für diejenige Beobachtung $\underline{x}_i$, für die $R_i = \min\limits_{j=1,\ldots,N} R_j$ ist, wird der Ausreißertest dann durchgeführt. Es handelt sich dabei um die Beobachtung, deren Entfernung aus dem Datensatz grob gesagt die stärkste Reduzierung der Streuung innerhalb der Daten bewirkt. Formal leitet man die Teststatistik $\Lambda = \min\limits_{j=1,\ldots,N} R_j$ als Likelihood–Ratio–Statistik für das oben angegebene Testproblem her. Die Hypothese der identischen Verteilung aller $\underline{X}_j$ wird verworfen, falls Λ kleiner ist als ein geeigneter kritischer Wert. In diesem Fall wird die Beobachtung $\underline{x}_i$ als Ausreißer identifiziert.

Betrachtet man das oben spezifizierte Alternativmodell, so kann man strenggenommen nicht mehr von einer einzigen Alternative sprechen. Die Problemstellung hat sich vielmehr von einem einfachen Testproblem zu einem Mehrentscheidungsproblem geändert. Man verwendet wiederum die Hypothese

$$H_0: \underline{X}_1,\ldots,\underline{X}_N \text{ i.i.d., } \underline{X}_i \sim N(\underline{\mu},\Sigma),\ i = 1,\ldots,N.$$

Dagegen stellt man die N Alternativen

$$H_k:\ \underline{X}_1,\ldots,\underline{X}_N \text{ stochastisch unabhängig und normalverteilt mit}$$

$$\begin{aligned}
&\mathrm{E}(\underline{X}_k) = \underline{\mu} + \underline{a},\ \underline{a} \neq \underline{0}, \\
&\mathrm{E}(\underline{X}_i) = \underline{\mu} \quad \forall i \neq k, \\
&\mathrm{Cov}(\underline{X}_i) = \Sigma \quad \forall i = 1, \ldots, N,
\end{aligned}$$

$k = 1, \ldots, N$.
Das heißt, die Alternative H_k besagt, daß die Realisation der Zufallsvariablen $\underline{X}_k$ ein Ausreißer ist.
Für diese Situation entwickelten Karlin und Truax (1960) ein Mehrentscheidungsverfahren, kurz auch Test genannt, das auf der Statistik $D^2 := \max_{j=1,\ldots,N} d_j^2$ beruht, $d_j^2 := (\underline{x}_j - \overline{\underline{x}})^T A^{-1} (\underline{x}_j - \overline{\underline{x}})$, $j = 1, \ldots, N$. Man entscheidet sich für H_k, falls gerade $D^2 = d_k^2$ und D^2 größer ist als ein geeigneter kritischer Wert. Falls D^2 kleiner oder gleich dem kritischen Wert ist, kann keine Beobachtung als Ausreißer erkannt werden. Diese Prozedur ist äquivalent zur oben beschriebenen, da $R_j = 1 - \frac{N}{N-1} d_j^2$ (Barnett, Lewis (1994), S. 287). Das von Karlin und Truax konstruierte Verfahren ist optimal im folgenden Sinn:

1. Es ist invariant unter affin linearen Transformationen $A\underline{x} + \underline{b}$ der Beobachtungen, $A \in I\!R^{p \times p}$ regulär, $\underline{b} \in I\!R^p$.

2. Die Wahrscheinlichkeit der Entscheidung für H_k, falls H_k wahr ist, ist unabhängig von k, $1 \leq k \leq N$.

3. Unter allen Niveau-α-Tests mit diesen Eigenschaften ist der oben angegebene Test optimal in dem Sinn, daß die Wahrscheinlichkeit der Entscheidung für H_k, falls H_k wahr ist, stets größer oder gleich der Wahrscheinlichkeit ist, sich in diesem Fall für eine der anderen Hypothesen zu entscheiden, $k = 1, \ldots, N$ (Karlin, Truax (1960), Mathar (1981), S. 70 ff.).

Damit ist auch der Test von Wilks optimal im obigen Sinn.

Entsprechende Tests existieren auch für Alternativen, die mehr als einen Ausreißer festlegen. Bei der Verallgemeinerung des Ansatzes von Wilks benutzt man als Teststatistik den minimalen „m–outlier scatter ratio“ (Wilks (1963), S. 413 ff.). Dieser wird nach demselben Prinzip gebildet wie die Statistik Λ, nur daß statt einer Beobachtung jeweils m Beobachtungen gleichzeitig aus dem Datensatz herausgenommen werden. Man berechnet also

$$R_{\underline{v}} = \frac{\det(A^{(\underline{v})})}{\det(A)} \text{ mit } A \text{ wie oben,}$$

$$\underline{v} = (v_1, \ldots, v_m), v_i \in \{1, \ldots, N\}, v_i \neq v_j \; \forall \; i \neq j,$$

$$A^{(\underline{v})} = \sum_{\substack{k=1 \\ k \neq v_i, i=1,\ldots,m}}^{N} (\underline{x}_k - \overline{\underline{x}}^{(\underline{v})})(\underline{x}_k - \overline{\underline{x}}^{(\underline{v})})^T \text{ und}$$

$$\overline{\underline{x}}^{(\underline{v})} = \frac{1}{N-m} \sum_{\substack{k=1 \\ k \neq v_i, i=1,\ldots,m}}^{N} \underline{x}_k$$

für alle möglichen Indexvektoren $\underline{v}$. Daraus erhält man als Teststatistik $\Lambda^{(m)} := \min_{\underline{v}} R_{\underline{v}}$, wobei das Minimum über alle Vektoren $\underline{v} = (v_1, \ldots, v_m)$ mit $v_i \in \{1, \ldots, N\}$, $v_i \neq v_j \; \forall \; i \neq j$, gebildet wird. Ist die Teststatistik $\Lambda^{(m)}$ kleiner als ein geeigneter kritischer Wert, so gibt der zu $\Lambda^{(m)}$ gehörende Vektor $\underline{v}$ die Indizes der als Ausreißer identifizierten Beobachtungen an.

Die Verteilungen aller dieser Teststatistiken sind nicht geschlossen darstellbar. Kritische Werte für verschiedene Niveaus und Anzahlen von Ausreißern findet man zum Beispiel bei Wilks (1963), Jennings, Young (1988) oder Fung (1988).

In ähnlicher Weise gibt es Ausreißertests für eine große Menge anderer Alternativen wie auch für andere Nullmodelle, etwa für multivariate Pareto-, Exponential- oder Gleichverteilungen (Barnett (1979), Barnett, Lewis (1994), S. 293 ff.) und für allgemeinere Formen elliptisch symmetrischer Verteilungen (Hara (1988)). Ausführlichere Darstellungen verschiedener Ausreißertests sind beispielsweise bei Barnett, Lewis (1994, Kap. 7.3) oder Hawkins (1980, Kap. 8) zu finden.

Die Tatsache, daß solche Mehrentscheidungsverfahren verlangen, zumindest die Anzahl der Ausreißer, auf die getestet werden soll, vorab festzulegen, bedeutet eine starke Einschränkung. Um diesem Problem zu begegnen, kann man zu sogenannten konsekutiven Testverfahren übergehen. Dabei wird nicht von vornherein eine Aussage über die Anzahl möglicher Ausreißer getroffen. Das Verfahren entscheidet vielmehr „selbst“, wieviele der Beobachtungen als Ausreißer anzusehen sind. Man unterscheidet zwei

Arten von konsekutiven Testverfahren, die sogenannten Inward- und die Outward–Prozeduren (auch Testen „von außen nach innen“ bzw. „von innen nach außen“). Das prinzipielle Vorgehen bei konsekutiven Testverfahren wird in Hawkins (1980, S. 63 ff.) erläutert.

Bei Inward–Testprozeduren wird sukzessive getestet, ob die in einer gewissen Ordnung größte Beobachtung ein Ausreißer bezüglich einer gegebenen Stichprobe ist. Falls ja, wird die Beobachtung aus den Daten entfernt und die nächstgrößte anhand der reduzierten Stichprobe untersucht. Beispiele für Inward–Prozeduren findet man etwa bei Hawkins (1973) und Hawkins (1980, S. 63).

Im Gegensatz dazu gehen Outward–Prozeduren umgekehrt vor. Man reduziert eine vorliegende Stichprobe zunächst auf eine kleinere Teilstichprobe. Die herauszunehmenden Daten werden durch ein geeignetes Verfahren ausgewählt. Anschließend wird für die aus der Stichprobe entfernten Beobachtungen in der umgekehrten Reihenfolge ihres Herausnehmens nacheinander getestet, ob sie Ausreißer bezüglich der verbliebenen Stichprobe sind. Falls eine Beobachtung nicht als Ausreißer eingeordnet wird, wird sie der Stichprobe wieder zugefügt, und der nächste Test basiert auf der so erweiterten Teilstichprobe. Wird eine Beobachtung als Ausreißer erkannt, gelten gleichzeitig alle noch nicht in der Teilstichprobe enthaltenen Beobachtungen als Ausreißer. Eine grundlegende Arbeit zum Thema der Outward–Prozeduren ist der Artikel von Rosner (1975). Weitere Beiträge sind zu finden bei Kimber (1982), Prescott (1979), Simonoff (1984) sowie Sweeting (1983).

Die Vorgehensweise bei solchen konsekutiven Testverfahren erinnert stark an mehrschrittige multiple Testprozeduren. Tatsächlich läßt sich die Methodik der multiplen Testtheorie auf die Ausreißerentdeckung anwenden. Jeder Test, der prüft, ob eine Beobachtung ein Ausreißer ist oder nicht, wird dabei als einzelne Komponente eines multiplen Tests betrachtet. Die kritischen Werte für die Einzeltests werden dann nach Kriterien des multiplen Testens bestimmt, so daß die multiple Prozedur ein vorgegebenes multiples oder globales Niveau hält. Eine ausführliche Darstellung der Zusammenhänge von Ausreißeridentifizierung und multipler Testtheorie findet man bei Pigeot (1993) und Gather, Pigeot (1994), wo derartige Zusammenhänge insbesondere für Outward–Prozeduren besprochen werden.

Ein Beispiel für eine Outward–Prozedur, die auch als multipler Test verstanden werden kann, ist ein Verfahren, das von Caroni und Prescott (1992) als Erweiterung des Tests von Wilks vorgestellt wird. Sie greifen dabei auf die Methodik eines von Rosner (1975) eingeführten Verfahrens für univariate Datensätze zurück. Das konsekutive Outward–Verfahren läuft folgendermaßen ab.

Zunächst wird die oben definierte Teststatistik $\Lambda =: \Lambda_1$ von Wilks für die gesamte Stichprobe berechnet. Diejenige Beobachtung $\underline{x}_i$, die zu Λ_1 gehört ($\Lambda_1 = \min_{j=1,\dots,N} R_j = R_i$), wird aus den Daten entfernt. Aus der reduzierten Stichprobe berechnet man erneut die Wilks–Statistik, bezeichnet mit Λ_2. Wiederum wird die zugehörige Beobachtung herausgenommen. Dieses Verfahren wiederholt man, bis eine vorgegebene Zahl k^* von entfernten Beobachtungen erreicht ist. Diese Zahl k^* ist die maximale Anzahl von Ausreißern, auf die zu testen ist. Auf diese Weise erhält man eine Folge $\Lambda_{k^*}, \Lambda_{k^*-1}, \dots, \Lambda_1$ von Wilks–Statistiken. Beginnend mit Λ_{k^*}, werden diese der Reihe nach mit kritischen Werten $\lambda_{k^*}, \lambda_{k^*-1}, \dots, \lambda_1$ verglichen. Das heißt, im r–ten Schritt wird die Hypothese getestet, daß alle in diesem Schritt betrachteten Beobachtungen Realisationen von unabhängigen, identisch $N(\underline{\mu}, \Sigma)$–verteilten Zufallsvektoren sind. Dagegen steht die Alternative, daß die im r–ten Schritt hinzugekommene, zu Λ_r gehörende Beobachtung aus einer $N(\underline{\mu} + \underline{a}, \Sigma)$–Verteilung mit $\underline{a} \neq \underline{0}$ stammt. Dazu wird geprüft, ob $\Lambda_r < \lambda_r$ gilt. Ist dies nicht der Fall, wird die Hypothese des r–ten Schritts nicht verworfen, die zu Λ_r gehörende Beobachtung gilt nicht als Ausreißer, und es wird der nächste Vergleich (Λ_{r-1} mit λ_{r-1}) durchgeführt. Falls $\Lambda_r < \lambda_r$ ist, werden die Hypothesen des r–ten und aller folgenden Schritte zugunsten der entsprechenden Alternativen verworfen und damit die zu den Wilks–Statistiken $\Lambda_r, \Lambda_{r-1}, \dots, \Lambda_1$ gehörenden Beobachtungen als Ausreißer identifiziert. Die kritischen Werte für die einzelnen Stufen des Verfahrens werden von Caroni und Prescott approximativ bestimmt, wobei sie ein von Rosner (1983) vorgeschlagenes Verfahren für univariate Stichproben auf den multivariaten Fall verallgemeinern. Die auf diese Weise erhaltenen kritischen Werte entsprechen denen für den Wilks–Test auf einen Ausreißer für die jeweiligen Stichprobenumfänge. Ob die auf diese Weise konstruierte Outward–Prozedur, als multipler Test betrachtet, ein gegebenes multiples Niveau hält, ist noch nicht geklärt (Pigeot (1993), S. 87).

Für die Wahl einer geeigneten Zahl k^* als maximale Anzahl möglicher Ausreißer empfehlen Caroni und Prescott (1992, S. 356), Erfahrungswerte zu benutzen: „If sets of data of similar kinds are screened frequently, then it is possible, with the experience obtained, to establish a reasonable upper limit for the maximum number of outliers to use in the test procedure.“

Selbst wenn man auf solche Erfahrungen zurückgreifen kann, bleibt die Wahl von k^* ein nichttriviales Problem. Eine ungeeignete Wahl kann nämlich zwei Effekte begünstigen, die im Zusammenhang mit der Identifizierung von Ausreißern häufig auftauchen und zu falschen Entscheidungen führen. Es handelt sich hierbei um den Masking- und den Swamping–Effekt (vgl. auch Kap. 3). Beim Masking–Effekt werden vorhandene Ausreißer nicht erkannt, weil sie sich gegenseitig „maskieren“, beim Swamping–Effekt werden zu viele Beobachtungen als Ausreißer deklariert. Die klassischen Ausreißertests, bei denen die genaue Anzahl der Ausreißer, auf die zu testen ist, vorher festgelegt werden muß, unterliegen beiden Effekten in der Regel sehr leicht. Aber auch die konsekutiven Testverfahren, die zum Teil mit dem Ziel konstruiert werden, Masking- und Swamping–Effekten zu widerstehen, erweisen sich häufig als anfällig (Hawkins (1980), Kap. 5.2 und 5.3).
Ein weiteres Problem im Kontext der Ausreißertests besteht in der Festlegung eines Alternativmodells. Eine Alternative wie etwa in dem oben angegebenen Test von Wilks bedeutet eine starke Restriktion auf eine ganz bestimmte Art von Ausreißern. Es wird nicht nur die Verteilung der Ausreißer auf eine bestimmte Klasse eingeschränkt, sondern darüber hinaus wird sogar angenommen, daß auch unter der Alternative alle Beobachtungen Realisationen stochastisch unabhängiger Zufallsvariablen sind. Eine solche Annahme ist aber gerade für die Ausreißer nicht unbedingt angebracht, da diese durchaus voneinander ebenso wie von den regulären Beobachtungen abhängen können. Zwar existieren auch Tests für weniger restriktive Alternativen, für die Entwicklung optimaler Prozeduren sind jedoch die Einschränkungen notwendig.
Will man sich nicht darauf einlassen, daß mögliche Ausreißer in einem Datensatz auf eine bestimmte Weise, das heißt, nach einem bestimmten erzeugenden Mechanismus zustande gekommen sind, so kann man zur Ausreißer–Identifizierung andere Verfahren benutzen, die einen mehr explorativen Charakter haben.

2.2 Heuristische Verfahren zur Ausreißerentdeckung

Im Gegensatz zu den im vorigen Abschnitt behandelten Ausreißertests gehen die meisten heuristischen Verfahren zur Ausreißerentdeckung von einer anschaulichen Definition eines Ausreißers als einer Beobachtung aus, die nicht zum Rest der Daten paßt („an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data", Barnett, Lewis (1994), S. 7). In Verbindung mit dieser Definition trifft man häufig auf Methoden, die eine visuelle Veranschaulichung der beobachteten Stichprobe oder geeigneter statistischer Kenngrößen der Daten beinhalten. Die Idee dabei ist, daß sich Ausreißer durch ein Erscheinungsbild manifestieren, das sich sichtbar von dem der übrigen Beobachtungen unterscheidet.
Trotz der Anschaulichkeit des hier benutzten Ausreißerbegriffs wird in der Regel ein Modell für die guten Beobachtungen benötigt. Man bevorzugt dafür oft ein Normalverteilungsmodell, das auch die meisten der hier vorgestellten Vorgehensweisen unterstellen.

Ein Verfahren, das die graphische Darstellung statistischer Kenngrößen einer zu untersuchenden Stichprobe verwendet, wurde von Bacon–Shone und Fung (1987) vorgestellt. Es basiert auf der $\Lambda^{(m)}$–Statistik für den Test auf m Ausreißer von Wilks, die im vorigen Abschnitt eingeführt wurde. Bacon–Shone und Fung legen für die Beobachtungen, die keine Ausreißer sind, eine p–variate Normalverteilung zugrunde. Sie nutzen die Tatsache, daß die Verteilung der Größen $R_{\underline{v}}$, die zur Bestimmung von $\Lambda^{(m)}$ herangezogen werden, unter diesem Nullmodell approximiert werden kann. Für das von ihnen vorgeschlagene graphische Verfahren gehen sie von den Werten $R_{\underline{v}}$ über zu den transformierten Größen $W_{\underline{v}} = -[N - \frac{1}{2}(p + m + 3)]\ln(R_{\underline{v}})$, die approximativ χ^2_{pm}–verteilt sind. Dabei ist N der Umfang der betrachteten Stichprobe, p die Dimension der Daten und m die vermutete Anzahl von Ausreißern in den Daten. Durch einige Vereinfachungen gelangen Bacon–Shone und Fung zu folgendem graphischen Verfahren. Die berechneten $W_{\underline{v}}$ werden der Größe nach geordnet, und man betrachtet die $n - m + 1$ größten, $W_{\underline{v}}^{(1)} \geq \ldots \geq W_{\underline{v}}^{(n-m+1)}$. Diese trägt man gegen geeignete Quantile der χ^2_p–Verteilung auf, genauer notiert man die Punktepaare $(W_0^{(k)}, W_{\underline{v}}^{(k)})$ mit

$W_0^{(k)} = \chi^2_{p;1-\frac{k}{n-m+2}}$, $k = 1, \ldots, n-m+1$. Liegen keine Ausreißer vor, sollten diese Punkte annähernd auf einer Geraden liegen. Derartige Graphiken werden für verschiedene Werte von m, der Anzahl der vermuteten Ausreißer, erstellt. Für diejenige Zahl m, die der wahren Anzahl von Ausreißern entspricht, erwartet man, daß der Punkt $(W_0^{(1)}, W_{\underline{v}}^{(1)})$ deutlich abseits der von den restlichen Punkten gebildeten Geraden liegt („[...] if we consider the plot [...] for $m = m^*$ where there are actually m^* outliers in the sample, we would expect that the highest point will be distinct [...] and that the remaining points will lie roughly on a straight line [...]", Bacon–Shone, Fung (1987), S. 156). Auf diese Weise wird zugleich mit der Entdeckung von Ausreißern bei der beschriebenen Methode auch deren Anzahl entdeckt.

Der Ansatz des Verfahrens von Bacon–Shone und Fung besteht im wesentlichen darin, die Ausprägungen einer statistischen Kenngröße für die Daten gegen die unter dem Nullmodell erwarteten Ausprägungen abzutragen. Falls keine Ausreißer vorliegen, ergibt sich ein annähernd linearer Graph, und Ausreißer sind durch die deutliche Abweichung von der von den guten Beobachtungen gebildeten Geraden zu erkennen. Hier wird die Philosophie des Quantil–Quantil–Plots, der ursprünglich zur Anpassung einer geeigneten Verteilung dient, für den Kontext der Identifizierung von Ausreißern nutzbar gemacht. Dieses Prinzip findet man häufig bei heuristischen Verfahren zur Ausreißerentdeckung. So benutzt zum Beispiel Healy (1968) einen Plot von Mahalanobisdistanzen gegen die Quantile der χ^2–Verteilung. Gnanadesikan und Kettenring (1972, S. 111 ff.) verwenden zwei Typen verallgemeinerter Distanzmaße, wobei in einem Fall Quantile von Gamma–Verteilungen, im anderen Fall von Beta- oder F–Verteilungen für die Erstellung der Graphik herangezogen werden. An der gleichen Stelle findet sich ein weiteres solches Verfahren, das mit einem speziellen Korrelationskoeffizienten und Quantilen der Normalverteilung arbeitet. Eine neuere Arbeit von Easton und McCulloch (1990) stellt eine multivariate Verallgemeinerung von Quantil–Quantil–Plots vor, die ebenfalls für die Identifizierung von Ausreißern geeignet ist. Beachtenswert ist hier vor allem, daß eine Methodik entwickelt wird, die es bei der Einbettung in die Problematik der Ausreißerentdeckung erlaubt, für die guten Daten eine Reihe verschiedener Wahrscheinlichkeitsverteilungen zu unterstellen.

Viele Techniken, die für die graphische Darstellung multivariater Stichproben zur Verfügung stehen, lassen sich ebenfalls gut für die Bestimmung von Ausreißern nutzen. Kleiner und Hartigan (1981) stellen zwei solche Techniken vor, die Repräsentation mehrdimensionaler Daten durch Bäume und Burgen. Abbildung 2.1 zeigt die Darstellung der Stimmenanteile der Republikaner bei den Präsidentschaftswahlen von 1964–1976 in 48 amerikanischen Bundesstaaten. Dabei wurden die Wahljahre als Beobachtungen, die Stimmenanteile in den einzelnen Bundesstaaten als Variablen interpretiert.

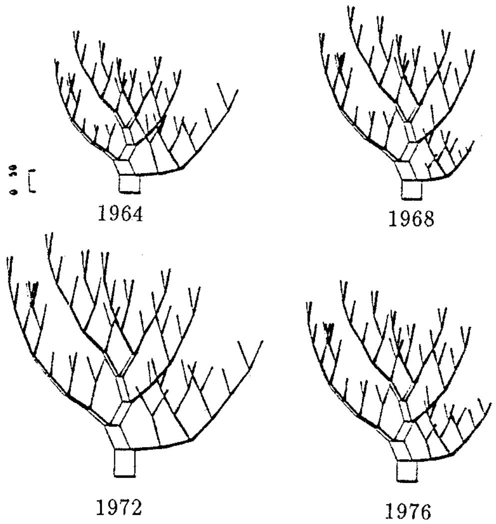


Abbildung 2.1: Stimmenanteile der Republikaner bei amerikanischen Präsidentschaftswahlen 1964–1976 in 48 Bundesstaaten, dargestellt als Bäume (Kleiner, Hartigan (1981), S. 267)

Jede Spitze eines Baumzweigs stellt einen Staat dar, und die Ordnung, nach der die Staaten mit den Zweigen assoziiert werden, kommt durch eine hierarchische Clusterung der Variablen zustande. Der zum Jahr 1972 gehörende Baum fällt auf den ersten Blick auf. Tatsächlich gab es 1972 einen deutlichen Wahlsieg der Republikaner in den USA. Ein weiteres eindrucksvolles Beispiel für die Entdeckung eines Ausreißers durch die graphische Darstellung einer Stichprobe findet man in der Diskussion über die Arbeit von Kleiner und Hartigan. Wainer (1981, S. 274) betrachtet die Stimmenanteile der Republikaner bei amerikanischen Präsidentschaftswahlen von 1932–1940 und 1960–1968 in 6 Bundesstaaten. Er präsentiert zwei Versionen der Darstellung dieser Daten als Gesichter, zum einen nach Chernoff (1973), zum anderen gemäß Flury und Riedwyl (1981). Diese zweite Version wird in Abbildung 2.2 dargestellt. Der Bundesstaat Mississippi fällt deutlich als möglicher Ausreißer ins Auge.

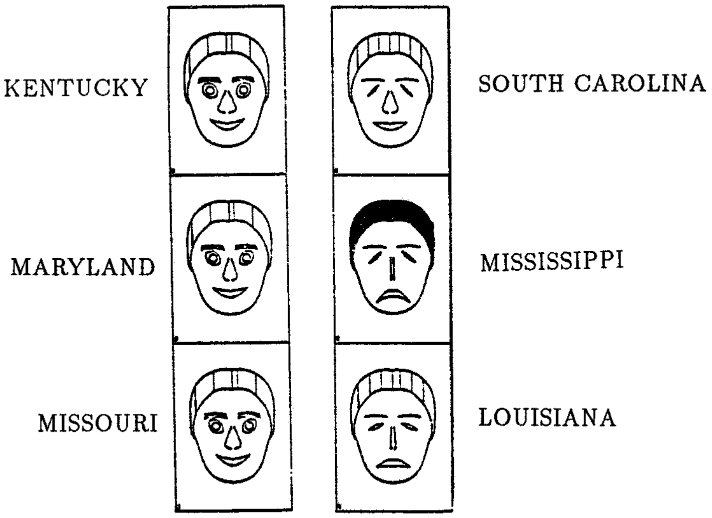


Abbildung 2.2: Stimmenanteile der Republikaner bei amerikanischen Präsidentschaftswahlen 1932–1940 und 1960–1968 in 6 Bundesstaaten, dargestellt als Gesichter (Wainer (1981), S. 274)

Es gibt viele Möglichkeiten für die graphische, also zweidimensionale Darstellung multivariater Daten. Dazu gehören unter anderem Profile (z.B. Bertin (1967)), Sterne (z.B. Goldwyn et al. (1971)), Glyphs (Anderson (1960)), Andrews–Plots (Andrews (1972)) oder Polyeder (Seaman et al. (1987)). Weitere Methoden finden sich bei Gnanadesikan (1977, Kap. 6.4). Alle diese Darstellungsweisen können auch dazu benutzt werden, Auffälligkeiten und damit auch potentielle Ausreißer zu erkennen.

Über die behandelten Prozeduren hinaus gibt es eine Fülle weiterer heuristischer Methoden, um Ausreißer in multivariaten Datensätzen zu identifizieren. Dazu zählen Verfahren, die auf der Hauptkomponentenanalyse beruhen (Rao (1964), S. 334, Barnett, Lewis (1994), S. 303 ff., Hawkins (1974), Hawkins (1980), S. 110 ff.), ebenso wie solche, die Ausreißer in jeder einzelnen Koordinate der Beobachtungen suchen (z.B. Barnett (1983)). Eine ausführliche Übersicht über diverse Methoden der Identifizierung von Ausreißern bietet die Arbeit von Gnanadesikan und Kettenring (1972).

Weitere Möglichkeiten findet man beispielsweise bei Bhandary (1992) oder Atkinson und Mulira (1993).

2.3 Ausreißer–Identifizierer

Die in den beiden vorangegangenen Abschnitten beschriebenen Ansätze bieten zwei sehr unterschiedliche Arten der Erkennung von Ausreißern. Die Ausreißertests werden zur Identifizierung von Ausreißern herangezogen, sind allerdings von ihrer Konzeption her eigentlich Tests auf das Vorliegen von Beobachtungen aus Verteilungen, die vom Nullmodell abweichen. Dabei herrscht eine starke Spezialisierung auf bestimmte Modelle vor. Die heuristischen Verfahren dagegen verzichten fast vollständig auf eine Modellbildung und gehen intuitiver an den Begriff des Ausreißers heran. Dabei wird in der Regel auf eine Formalisierung dieses Begriffs verzichtet. Die Benutzung sogenannter Ausreißer–Identifizierer, deren Definition auf Davies und Gather (1993) zurückgeht, ist eine weitere Methode zur Ausreißerentdeckung. Sie berücksichtigt einerseits die Modellbildung, wobei sie jedoch im Gegensatz zu den Ausreißertests unmittelbar zum Zweck der Identifizierung von Ausreißern konzipiert ist. Andererseits trägt diese Methode dem intuitiven Verständnis der Bezeichnung „Ausreißer“ Rechnung, wobei im Gegensatz zu den heuristischen Verfahren dieser Begriff formalisiert wird.
Ausreißer werden als solche Beobachtungen definiert, deren Position bezüglich eines zugrundegelegten Modells für die beobachteten Daten hinreichend unwahrscheinlich ist. Unter der Modellannahme einer multivariaten Normalverteilung läßt sich dann ein Bereich angeben, so daß alle Punkte in diesem Bereich als Ausreißer definiert werden.

Definition 2.1 (Becker (1992), S. 15)

Sei $\alpha \in]0,1[$, sei $\chi^2_{p;1-\alpha}$ das $(1-\alpha)$–Quantil der χ^2_p–Verteilung.
Die Menge

$$\text{out}(\alpha, \underline{\mu}, \Sigma) := \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) > \chi^2_{p;1-\alpha}\}$$

heißt *α–Ausreißer–Bereich der $N(\underline{\mu}, \Sigma)$–Verteilung.*

Mit dieser Formalisierung ist $\mathrm{P}(\underline{X} \in \mathrm{out}(\alpha, \underline{\mu}, \Sigma)) = \alpha$ für $\underline{X} \sim N(\underline{\mu}, \Sigma)$, das heißt, die Definition spiegelt gerade die anschauliche Vorstellung wider, daß Ausreißer Beobachtungen sind, die unter einem betrachteten Modell nur mit geringer Wahrscheinlichkeit vorkommen können.

In der Situation, daß eine Stichprobe aus $N(\underline{\mu}, \Sigma)$–verteilten Beobachtungen vorliegt, in der Ausreißer vermutet werden, könnten bei Kenntnis von $\underline{\mu}$ und Σ alle α–Ausreißer identifiziert werden. Da in der Regel aber $\underline{\mu}$ und Σ unbekannt sind, muß zur Identifizierung von Ausreißern der Bereich $\mathrm{out}(\alpha, \underline{\mu}, \Sigma)$ geschätzt werden. Dabei ist zu beachten, daß die Stichprobe, die als Grundlage der Schätzung dient, selbst Ausreißer enthalten kann. Dies kann in folgender Weise modelliert werden. Die Stichprobe $\underset{\sim}{x}_N = (\underline{x}_1, \ldots, \underline{x}_N)$ mit $\underline{x}_i \in \mathbb{R}^p, i = 1, \ldots, N$, enthält n reguläre Beobachtungen, die Realisationen unabhängiger und identisch $N(\underline{\mu}, \Sigma)$–verteilter Zufallsvariablen sind. Die restlichen $k = N - n$ Beobachtungen liegen im δ_N–Ausreißer–Bereich der $N(\underline{\mu}, \Sigma)$–Verteilung, wobei $\delta_N \in]0, 1[$ nicht bekannt ist. Ebenso ist die Anzahl k von δ_N–Ausreißern in der Stichprobe unbekannt. An die δ_N–Ausreißer werden keine weiteren Anforderungen gestellt. Es ist daher auch möglich, daß sich unter ihnen Beobachtungen aus der $N(\underline{\mu}, \Sigma)$–Verteilung befinden, die aber dennoch aufgrund ihrer Lage als δ_N–Ausreißer bewertet werden. Unter diesen Modellvoraussetzungen läßt sich ein Schätzer für einen Ausreißerbereich definieren.

Definition 2.2 (Becker (1992), S. 18)

Gegeben sei eine Stichprobe $\underset{\sim}{x}_N$ vom Umfang N, sei $0 \leq k < N/2, N = n + k, k$ unbekannt. Sei $\underset{\sim}{x}_N = (\underline{x}_1, \ldots, \underline{x}_N)$ mit $\underline{x}_i \in \mathbb{R}^p, p \geq 2, i = 1, \ldots, N$. Die Stichprobe $\underset{\sim}{x}_N$ enthalte eine Anzahl k von δ_N–Ausreißern; die restlichen $n = N - k$ regulären Beobachtungen seinen Realisationen unabhängiger, identisch $N(\underline{\mu}, \Sigma)$–verteilter Zufallsvektoren. Gegeben sei weiterhin $\alpha_N \in]0, 1[$.

Ein *multivariater (α_N–) Ausreißer–Identifizierer* ist definiert durch

eine nichtsinguläre, symmetrische Matrix $S = S(\underset{\sim}{x}_N) \in \mathbb{R}^{p \times p}$,

einen Vektor $\underline{m} = \underline{m}(\underset{\sim}{x}_N) \in \mathbb{R}^p$

und eine Zahl $c = c(p, N, \alpha_N) \in \mathbb{R}, c \geq 0$,

die nicht von der Anordnung der δ_N–Ausreißer in der Stichprobe abhängen.

Durch diese Größen wird ein Bereich

$$\underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N) := \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{m})^T S^{-1} (\underline{x} - \underline{m}) \geq c\}$$

festgelegt.
Alle $\underline{x} \in \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N)$ werden als α_N–Ausreißer bezüglich $N(\underline{\mu}, \Sigma)$ identifiziert.

Dabei ist c eine Normierungskonstante, die die Vergleichbarkeit verschiedener Identifizierer ermöglicht. Beispielsweise kann c so gewählt werden, daß

$$\mathrm{P}(\underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N) \subseteq \mathrm{out}(\alpha_N, \underline{\mu}, \Sigma)) = 1 - \alpha$$

für eine gegebene Zahl $\alpha \in]0, 1[$, falls $\underset{\sim}{x}_N$ nur aus Beobachtungen $N(\underline{\mu}, \Sigma)$–verteilter Zufallsvariablen besteht.
Die Bezeichnung $\underline{\mathrm{OR}}$ wird sowohl für den oben definierten Bereich als auch für den Identifizierer selbst verwendet.

Die Regel, nach der Beobachtungen als Ausreißer identifiziert werden, lautet also: betrachte zu jeder Beobachtung einen Abstand vom Typ einer Mahalanobisdistanz. Falls dieser Abstand zu groß ist, wird die Beobachtung als Ausreißer erkannt. Die Idee, eine Distanz vom Mahalanobis–Typ mit geeignet gewählten Schätzern zur Identifizierung von Ausreißern zu benutzen, wird auch von Rocke und Woodruff (1993) gewählt. Allerdings wird dort der Identifizierer nicht formal definiert. Auch Rousseeuw und van Zomeren (1990) verfolgen einen derartigen Ansatz, ebenfalls ohne formale Definition der Identifizierungsprozedur. Das dort verwendete Verfahren wird in Kapitel 5 im Beispiel 5.3.1 im Kontext der hier definierten Ausreißer–Identifizierer näher untersucht.

Die auf die oben angegebene Weise definierten Ausreißer–Bereiche und Identifizierer sind multivariate Verallgemeinerungen der von Davies und Gather (1993) eingeführten univariaten Konzepte. Betrachtet man den Ausreißerbereich

$$\mathrm{out}(\alpha, \underline{\mu}, \Sigma) = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) > \chi^2_{p;1-\alpha}\}$$

und setzt $p = 1$, so werden die Vektoren $\underline{x}$ und $\underline{\mu}$ zu reellen Zahlen x und μ. Die Kovarianzmatrix Σ reduziert sich auf die univariate Varianz σ^2 der $N(\mu, \sigma^2)$–Verteilung, und der Ausdruck $(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})$ wird gerade zu $\frac{(x-\mu)^2}{\sigma^2} = \left(\frac{|x-\mu|}{\sigma}\right)^2$. Weiter ist $\chi^2_{1;1-\alpha} = (z_{1-\alpha/2})^2$, so daß

$$\frac{(x-\mu)^2}{\sigma^2} > \chi^2_{1;1-\alpha} \Leftrightarrow |x-\mu| > \sigma z_{1-\alpha/2},$$

wobei $z_{1-\alpha/2}$ das $(1-\alpha/2)$–Quantil der Standardnormalverteilung bezeichne. Damit geht der multivariate Ausreißerbereich $\mathrm{out}(\alpha, \underline{\mu}, \Sigma)$ für $p = 1$ in den von Davies und Gather (1993, S. 782) definierten univariaten α–Ausreißer–Bereich $\mathrm{out}(\alpha, \mu, \sigma^2)$ von $N(\mu, \sigma^2)$ über,

$$\mathrm{out}(\alpha, \mu, \sigma^2) := \{x \in \mathbb{R} : |x-\mu| > \sigma z_{1-\alpha/2}\}.$$

In gleicher Weise läßt sich der multivariate α_N–Ausreißer–Identifizierer als Verallgemeinerung des entsprechenden univariaten Verfahrens auffassen. Dabei ist ein univariater Ausreißer–Identifizierer definiert durch zwei Zahlen $L = L(\underset{\sim}{x}_N, \alpha_N)$ und $R = R(\underset{\sim}{x}_N, \alpha_N)$ mit $L \leq R$, die einen Bereich

$$\mathrm{OR}(\underset{\sim}{x}_N, \alpha_N) :=]-\infty, L(\underset{\sim}{x}_N, \alpha_N)] \cup [R(\underset{\sim}{x}_N, \alpha_N), \infty[$$

bestimmen. Alle $x \in \mathrm{OR}(\underset{\sim}{x}_N, \alpha_N)$ werden als α_N–Ausreißer bezüglich $N(\mu, \sigma^2)$ identifiziert (Davies, Gather (1993), S. 783). Der Bereich OR ist als Komplement eines Intervalls die univariate Entsprechung des multivariaten Bereichs $\underline{\mathrm{OR}}$ (Komplement eines Ellipsoids). Dies läßt sich durch Umparametrisierung von OR erkennen. Mit $m(\underset{\sim}{x}_N) = \frac{1}{2}(L+R)$ und $c(\alpha_N, \underset{\sim}{x}_N)s(\underset{\sim}{x}_N) = \frac{1}{2}(R-L)$ läßt sich OR schreiben als

$$\mathrm{OR}(\underset{\sim}{x}_N, \alpha_N) := \{x \in \mathbb{R} : \left|\frac{x-m}{s}\right| \geq c\},$$

woraus sich die Analogie zum multivariaten Identifizierer $\underline{\mathrm{OR}}$ unmittelbar ergibt.

Wie man in Definition 2.2 sieht, werden hier Ausreißer–Identifizierer betrachtet, die auf einem Lokationsschätzer $\underline{m}$ und einem Kovarianzschätzer S basieren.
Für die beiden Schätzer sei affine Äquivarianz vorausgesetzt, das heißt, für jede affine Transformation $\underset{\sim}{x}_N \mapsto A\underset{\sim}{x}_N + \underline{b}$, $A \in \mathbb{R}^{p\times p}$, A regulär, $\underline{b} \in \mathbb{R}^p$, gelte

1. $\underline{m}(A\underset{\sim}{x}_N + \underline{b}) = A\underline{m}(\underset{\sim}{x}_N) + \underline{b}$,

2. $S(A\underset{\sim}{x}_N + \underline{b}) = A^T S(\underset{\sim}{x}_N) A$.

Dabei sei $\underset{\sim}{x}_N = (\underline{x}_1, \ldots, \underline{x}_N), \underline{x}_i \in \mathbb{R}^p$, und $A\underset{\sim}{x}_N + \underline{b} := (A\underline{x}_1 + \underline{b}, \ldots, A\underline{x}_N + \underline{b})$. Unter diesen Voraussetzungen ist auch der Identifizierer $\underline{\mathrm{OR}}$ affin äquivariant, d. h.,

$$\underline{x} \in \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N) \Leftrightarrow A\underline{x} + \underline{b} \in \underline{\mathrm{OR}}(A\underset{\sim}{x}_N + \underline{b}, \alpha_N).$$

Damit ist gesichert, daß eine (affine) Transformation der Stichprobe keinen Einfluß darauf hat, welche der Beobachtungen als Ausreißer identifiziert werden. Eine Änderung der Maßeinheit der Daten spielt also keine Rolle.

Das Konzept der Ausreißer–Identifizierer enthält Aspekte beider Arten von Methoden zur Ausreißerentdeckung, wie sie in den ersten beiden Abschnitten dieses Kapitels beschrieben werden. Die Form des Bereichs, der einen Identifizierer definiert, wird durch die Form des Ausreißerbereichs festgelegt. Der Ausreißerbereich wird bezüglich eines gewissen Wahrscheinlichkeitsmodells bestimmt, das dem Nullmodell für die „guten“ Daten entspricht. Hier geht der Aspekt der formalen Modellierung ein, der auch bei den im ersten Abschnitt dargestellten Ausreißertests zu finden ist. Die Definition des Ausreißerbereichs als eine Region, in der sich Beobachtungen unter dem gewählten Nullmodell kaum realisieren, lehnt an das intuitive Verständnis des Ausreißerbegriffs an, wie es bei den heuristischen Methoden im zweiten Abschnitt zum Tragen kommt. So entsteht ein Verfahren, das speziell für die Entdeckung von Ausreißern konzipiert ist und mit dem einerseits Ausreißer nach eindeutigen Kriterien formal identifiziert werden können und das andererseits leicht einsichtig ist. In den folgenden Kapiteln werden Eigenschaften dieses Verfahrens betrachtet.

3 Masking- und Swamping–Bruchpunkte von Ausreißer–Identifizierern

Um das Verhalten der im vorigen Abschnitt definierten Ausreißer–Identifizierer beurteilen zu können, benötigt man Bewertungs- bzw. Vergleichskriterien. Ein erster Schritt bei der Konstruktion solcher Kriterien ist die Charakterisierung des Worst–case–Verhaltens der zu beurteilenden Verfahren. Damit lassen sich Prozeduren finden, deren Leistung auch im schlechtesten Fall noch akzeptabel ist. Zwei mögliche Worst–case–Kriterien sind der Masking- und der Swamping–Bruchpunkt. Diese beiden Größen hängen mit den beiden im vorigen Kapitel kurz erwähnten Phänomenen des Masking und Swamping zusammen, die bei der Identifizierung von Ausreißern häufig auftreten. Die Bruchpunkte geben jeweils den kleinsten Anteil an Ausreißern in einer Stichprobe an, mit dem man bereits erreichen kann, daß ein Identifizierer dem Masking- bzw. Swamping–Effekt unterliegt. In den folgenden Abschnitten werden Masking- und Swamping–Bruchpunkt eines Identifizierers definiert und ihr Zusammenhang mit den Finite–sample Bruchpunkten der in dem Identifizierer verwendeten Schätzer untersucht.

3.1 Definitionen der Bruchpunkte

Die im vorigen Kapitel definierten Ausreißer–Identifizierer werden durch Lokations- und Kovarianzschätzer $\underline{m}$ und S festgelegt. Für die Beurteilung der Robustheit von Schätzern existiert unter anderem das Kriterium des Finite–sample Bruchpunkts. Diese Größe gibt an, wie groß der Anteil beliebig schlecht plazierter Beobachtungen in einer Stichprobe mindestens sein muß, damit der Schätzer beliebig unsinnige Ergebnisse liefert (Zusammenbruch).

Definition 3.1 (Donoho, Huber (1983), S. 160, Lopuhaä, Rousseeuw (1991), S. 231)

Sei $\underset{\sim}{x}_N = (\underline{x}_1^\tau, \ldots, \underline{x}_N^\tau)$ eine Stichprobe vom Umfang N von regulären Beobachtungen aus einer $N(\underline{\mu}, \Sigma)$–Verteilung, sei $\underset{\sim}{y}_{N,k} = (\underline{x}_{i_1}^\tau,\ \ldots,\ \underline{x}_{i_n}^\tau,\ \underline{y}_1,\ \ldots,\ \underline{y}_k)$, $\underline{y}_j \in \mathbb{R}^p$, $j = 1, \ldots, k$, $N = n + k$, eine durch Austausch von k Beobachtungen von $\underset{\sim}{x}_N$ durch beliebige Vektoren entstandene Stichprobe.

a) Sei $T := \{T(\underset{\sim}{x}_m)\}_{m \in \mathbb{N}}$ eine Folge von p–dimensionalen Lokationsschätzern für den Erwartungswert $\underline{\mu}$. Dann heißt

$$\varepsilon^*(\underset{\sim}{x}_N, T) := \min_{1 \leq k \leq N} \{ \frac{k}{N} : \sup_{\underset{\sim}{y}_{N,k}} \| T(\underset{\sim}{x}_N) - T(\underset{\sim}{y}_{N,k}) \| = \infty \}$$

der *Finite–sample Bruchpunkt von T.*

Dabei sei mit $\|.\|$ die euklidische Norm im $\mathbb{R}^p$ bezeichnet.

b) Sei $C := \{C(\underset{\sim}{x}_m)\}_{m \in \mathbb{N}}$ eine Folge von Schätzern für die Kovarianzmatrix Σ.
Zu einer symmetrischen Matrix $A \in \mathbb{R}^{p \times p}$ seien die Eigenwerte gegeben als $\lambda_1(A) \geq \ldots \geq \lambda_p(A)$,
und zu $A, B \in \mathbb{R}^{p \times p}, A, B$ positiv definit, sei die Abbildung D definiert durch $D(A, B) := \max\{|\lambda_1(A) - \lambda_1(B)|, |\frac{1}{\lambda_p(A)} - \frac{1}{\lambda_p(B)}|\}$.
Dann heißt

$$\varepsilon^*(\underset{\sim}{x}_N, C) := \min_{1 \leq k \leq N} \{ \frac{k}{N} : \sup_{\underset{\sim}{y}_{N,k}} D(C(\underset{\sim}{x}_N), C(\underset{\sim}{y}_{N,k})) = \infty \}$$

der *Finite–sample Bruchpunkt von C.*

Die hier betrachteten Schätzer $\underline{m}$ und S seien so beschaffen, daß ihre Finite–sample Bruchpunkte nicht von den in einer Stichprobe $\underset{\sim}{x}_N$ vorkommenden Werten $\underline{x}_1, \ldots, \underline{x}_N$ abhängen, das heißt, für zwei unterschiedliche Stichproben $\underset{\sim}{x}_N^1$ und $\underset{\sim}{x}_N^2$ des gleichen Umfangs N ergebe sich der gleiche Bruchpunkt $\varepsilon^*(\underset{\sim}{x}_N^1, T) = \varepsilon^*(\underset{\sim}{x}_N^2, T)$ für den Schätzer $T \in \{\underline{m}, S\}$. Diese Anforderung stellt keine starke Einschränkung dar; sie wird von den meisten Schätzern erfüllt (vgl. Donoho, Huber (1983), S. 161, Gordaliza (1991), S. 391).

Bei der obigen Definition von Finite–sample Bruchpunkten betrachtet man eine gegebene Stichprobe $\underset{\sim}{x}_N$ und führt den Zusammenbruch der Schätzer durch den Austausch (engl. replacement) von Beobachtungen herbei. Es gibt auch noch einen anderen Ansatz, bei dem zu einer gegebenen Stichprobe Beobachtungen hinzugefügt werden (engl. addition), so daß ein Zusammenbruch der Schätzer erreicht wird. Der Unterschied zwischen diesen beiden Ansätzen ist allerdings nur formal; das Ergebnis, der Wert des

Bruchpunkts, ändert sich dadurch nicht. Diese Tatsache läßt sich leicht einsehen, wenn man die beiden Versionen der Definition von Finite–sample Bruchpunkten betrachtet. Dies soll hier exemplarisch für Lokationsschätzer durchgeführt werden.

Der Finite–sample Bruchpunkt für den additiven Ansatz ist definiert als

$$\varepsilon_*(\underset{\sim}{x}_N, T) := \min_k \{ \frac{k}{n+k} : \sup_{\underset{\sim}{y}_{n+k}} \|T(\underset{\sim}{x}_n) - T(\underset{\sim}{y}_{n+k})\| = \infty \},$$

wobei $\underset{\sim}{x}_n = (\underline{x}_1^r, \ldots, \underline{x}_n^r)$ eine Stichprobe vom Umfang n von regulären Beobachtungen ist und $\underset{\sim}{y}_{n+k} = (\underline{x}_1^r, \ldots, \underline{x}_n^r, \underline{y}_1, \ldots, \underline{y}_k)$ eine durch Hinzufügen von k beliebigen Beobachtungen zu $\underset{\sim}{x}_n$ entstandene Stichprobe vom Umfang $N = n + k$ ist.

Sei nun $\varepsilon^*(\underset{\sim}{x}_N, T) = \frac{k}{N}$. Das bedeutet, daß sich zu jeder vorgegebenen Schranke $M \in \mathbb{R}$ eine Menge von k Punkten $\underline{y}_1, \ldots, \underline{y}_k$ finden läßt, so daß

$$\|T(\underset{\sim}{x}_N) - T(\underset{\sim}{y}_{N,k})\| > M.$$

Dabei ist $\underset{\sim}{x}_N = (\underline{x}_1^r, \ldots, \underline{x}_n^r, \underline{x}_{n+1}^r, \ldots, \underline{x}_N^r)$ und $\underset{\sim}{y}_{N,k} = (\underline{x}_1^r, \ldots, \underline{x}_n^r, \underline{y}_1, \ldots, \underline{y}_k)$.

Gleichzeitig ist $\underset{\sim}{y}_{N,k}$ die gleiche Stichprobe, die entsteht, wenn man zu den n Beobachtungen $\underline{x}_1^r, \ldots, \underline{x}_n^r$ die Punkte $\underline{y}_1, \ldots, \underline{y}_k$ hinzufügt, d. h., es ist $\underset{\sim}{y}_{N,k} = \underset{\sim}{y}_{n+k}$.

Betrachtet man die Stichproben $\underset{\sim}{x}_N$ und $\underset{\sim}{x}_n$ als gegeben, so läßt sich folgende Abschätzung durchführen:

$$\begin{aligned} M &< \|T(\underset{\sim}{x}_N) - T(\underset{\sim}{y}_{N,k})\| = \|T(\underset{\sim}{x}_N) - T(\underset{\sim}{y}_{n+k})\| \\ &= \|T(\underset{\sim}{x}_N) - T(\underset{\sim}{x}_n) + T(\underset{\sim}{x}_n) - T(\underset{\sim}{y}_{n+k})\| \\ &\leq \underbrace{\|T(\underset{\sim}{x}_N) - T(\underset{\sim}{x}_n)\|}_{=A \text{ (fest)}} + \|T(\underset{\sim}{x}_n) - T(\underset{\sim}{y}_{n+k})\|. \end{aligned}$$

Die Betrachtung von $M \to \infty$ liefert, daß mit $\varepsilon^*(\underset{\sim}{x}_N, T) = \frac{k}{N}$ für $\varepsilon_*(\underset{\sim}{x}_N, T)$ die Abschätzung $\varepsilon_*(\underset{\sim}{x}_N, T) \leq \frac{k}{N} = \frac{k}{n+k}$ gelten muß.

Umgekehrt kann man von $\varepsilon_*(\underset{\sim}{x}_N, T) = \frac{k}{N}$ mit der gleichen Argumentation darauf schließen, daß $\varepsilon^*(\underset{\sim}{x}_N, T) \leq \frac{k}{N}$ gilt.

Hier lautet die entsprechende Abschätzung:

$$\begin{aligned} M &< \|T(\underset{\sim}{x}_n) - T(\underset{\sim}{y}_{n+k})\| = \|T(\underset{\sim}{x}_n) - T(\underset{\sim}{y}_{N,k})\| \\ &= \|T(\underset{\sim}{x}_n) - T(\underset{\sim}{x}_N) + T(\underset{\sim}{x}_N) - T(\underset{\sim}{y}_{N,k})\| \end{aligned}$$

$$\leq \underbrace{\|T(\underset{\sim}{x}_n) - T(\underset{\sim}{x}_N)\|}_{=A \text{ (fest)}} + \|T(\underset{\sim}{x}_N) - T(\underset{\sim}{y}_{N,k})\|.$$

Insgesamt folgt also die Gleichheit der Bruchpunkte.

Damit spielt es keine Rolle, auf welchem der beiden Ansätze die Definition des Finite–sample Bruchpunkts beruht. Hier wird der Ansatz über den Austausch von Beobachtungen verwendet, da sich dieser in der Literatur stärker durchgesetzt hat.

Genauso wie Schätzer $\underline{m}$ und S zusammenbrechen können, kann man auch bei Identifizierern von einem Zusammenbruch sprechen. Man unterscheidet dabei zwei Möglichkeiten: den Zusammenbruch bezüglich des Masking–Effekts und den bezüglich des Swamping–Effekts. Der Masking–Effekt tritt auf, wenn das Vorhandensein eines oder mehrerer Ausreißer dazu führt, daß Ausreißer nicht identifiziert werden. Der Masking–Bruchpunkt eines Identifizierers $\underline{\text{OR}}$ gibt grob gesagt den kleinsten Anteil an Ausreißern an, der dazu führen kann, daß der Identifizierer dem Masking–Effekt unterliegt.

In Anlehnung an die entsprechenden Definitionen in Abschnitt 2.2 von Davies und Gather (1993) werden folgende Begriffsbildungen getroffen.

Definition 3.2 (Becker (1992), S. 23, S. 26 f.)

Gegeben seien eine Folge $\boldsymbol{\alpha} = (\alpha_N)_{N \in \mathbb{N}}, 0 < \alpha_N < 1,\ \delta \in]0,1[$ sowie reguläre Beobachtungen $\underset{\sim}{x}_n^r = (\underline{x}_1^r, \ldots, \underline{x}_n^r)$. Sei

$$\begin{aligned}
\beta^M \;&:=\; \beta^M(\underline{\text{OR}}, \alpha_N, \underset{\sim}{x}_n^r, k, \delta)\\
&:=\; \inf\{\, \beta > 0 : \text{es ex. } \delta\text{–Ausreißer } \underset{\sim}{x}_k^0 = (\underline{x}_1^0, \ldots, \underline{x}_k^0), \text{ so daß irgendein Punkt,}\\
&\qquad\quad \text{der ein } \beta\text{–Ausreißer ist, von } \underline{\text{OR}} \text{ auf Basis von } \underset{\sim}{x}_N = (\underset{\sim}{x}_n^r, \underset{\sim}{x}_k^0) \text{ nicht}\\
&\qquad\quad \text{als } \alpha_N\text{–Ausreißer identifiziert wird}\},\\
k^M \;&:=\; k^M(\underline{\text{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta)\\
&:=\; \min\{k : \beta^M(\underline{\text{OR}}, \alpha_{n+k}, \underset{\sim}{x}_n^r, k, \delta) = 0\}.
\end{aligned}$$

Die Zahl β^M heißt *Masking–Punkt*, und

$$\varepsilon^M(\underline{\text{OR}}) \;:=\; \varepsilon^M(\underline{\text{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta) \;:=\; \frac{k^M(\underline{\text{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta)}{n + k^M(\underline{\text{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta)}$$

heißt *Masking–Bruchpunkt des Identifizierers* $\underline{\text{OR}}$.

Der andere Fall des Zusammenbruchs, Swamping, bedeutet, daß durch das Auftreten von δ–Ausreißern Punkte, die keine α_N–Ausreißer sind, vom Identifizierer als beliebige Ausreißer eingestuft werden. Analog zum Masking–Bruchpunkt gibt der Swamping–Bruchpunkt den kleinsten Anteil von δ–Ausreißern an, der Swamping verursachen kann.

Definition 3.3 (Becker (1992), S. 24, S. 27)

Unter den Voraussetzungen von Definition 3.2 seien

$$\begin{aligned}
\beta^S &:= \beta^S(\underline{\mathrm{OR}}, \alpha_N, \underset{\sim}{x}_n^r, k, \delta) \\
&:= \inf\{\beta > 0 : \text{es ex. } \delta\text{–Ausreißer } \underset{\sim}{x}_k^0, \text{so daß irgendein Punkt, der} \\
&\qquad\quad \text{kein } \alpha_N\text{–Ausreißer ist, als } \beta\text{–Ausreißer identifiziert wird}\}, \\
k^S &:= k^S(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta) \\
&:= \min\{k : \beta^S(\underline{\mathrm{OR}}, \alpha_{n+k}, \underset{\sim}{x}_n^r, k, \delta) = 0\}.
\end{aligned}$$

Die Zahl β^S heißt *Swamping–Punkt*, und

$$\varepsilon^S(\underline{\mathrm{OR}}) := \varepsilon^S(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta) := \frac{k^S(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta)}{n + k^S(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta)}$$

heißt *Swamping–Bruchpunkt des Identifizierers $\underline{OR}$.*

Will man Finite–sample Bruchpunkte einerseits und Masking- und Swamping–Bruchpunkte andererseits in Beziehung zueinander setzen, so muß man beachten, daß diese beiden Typen von Bruchpunkten auf unterschiedlichen Ansätzen beruhen. Während die Finite–sample Bruchpunkte von Schätzern das Austauschen von Beobachtungen heranziehen, liegt den für Identifizierer definierten Versionen das Hinzufügen von Beobachtungen zugrunde. Mit den Überlegungen nach Definition 3.1 spielt es für den Finite–sample Bruchpunkt jedoch keine Rolle, welcher der beiden Ansätze zugrunde gelegt wird, sein Wert ändert sich dadurch nicht. Daher ist es möglich, die beiden Arten von Bruchpunkten trotz der unterschiedlichen Ansätze in Beziehung zueinander zu setzen.

3.2 Beziehungen zwischen Bruchpunkten von Schätzern und dem Masking–Bruchpunkt von Ausreißer–Identifizierern

Da Ausreißer–Identifizierer $\underline{\mathrm{OR}}$ über Lokations- und Kovarianzschätzer $\underline{m}$ und S definiert werden, ist es intuitiv einleuchtend, daß „gutes" oder „schlechtes" Verhalten von Identifizierern unmittelbar mit „gutem" oder „schlechtem" Verhalten der Schätzer zusammenhängt. Für Identifizierer im linearen Regressionsmodell wurde bereits nachgewiesen, daß deren Masking- und Swamping–Bruchpunkte entscheidend von den Finite–sample Bruchpunkten der verwendeten Schätzer bestimmt werden (Boscher (1992), Kap. 3). Auch für multivariate Ausreißer–Identifizierer lassen sich bezüglich des Gütekriteriums „Bruchpunkt" solche Zusammenhänge zeigen. Dazu werden zunächst Zusammenhänge der Schätzer mit dem Masking–Effekt betrachtet.

Wie bereits erwähnt, bedeutet Masking, daß beliebig „große" Ausreißer nicht als solche erkannt werden. Das Bruchpunktverhalten der in einem Identifizierer verwendeten Schätzer hat einen Einfluß auf den Masking–Bruchpunkt des Identifizierers, wie der folgende Satz zeigt.

Satz 3.1

Seien $\underline{\mathrm{OR}}$, $\underline{m}$ und S wie in Definition 2.2. Seien $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) =: \frac{k_1}{N}$ und $\varepsilon^*(\underset{\sim}{x}_N, S) =: \frac{k_2}{N}$ die Finite–sample Bruchpunkte der beiden Schätzer, wobei $k_i < \frac{N}{2}$ gelte, $i = 1, 2$. Seien weiter $k := \min\{k_1, k_2\}$, $\boldsymbol{\alpha} = (\alpha_N)_{N \in \mathbb{N}}$, $0 < \alpha_N < 1$, und $\delta \in]0, 1[$. Sei $\underset{\sim}{x}_n^r = (\underline{x}_1^r, \ldots, \underline{x}_n^r)$ eine Stichprobe regulärer Beobachtungen aus $N(\underline{\mu}, \Sigma)$–verteilten Zufallsvariablen.

Dann gilt für den Masking–Bruchpunkt $\varepsilon^M(\underline{\mathrm{OR}})$:

$$\varepsilon^M(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta) \geq \frac{k}{N}$$

mit $N = n + k$.

Beweis

Beachte die folgenden Definitionen:

$$\begin{aligned} \beta^M \quad = \quad & \inf\{\beta > 0 : \text{es ex. } \delta\text{–Ausreißer } \underset{\sim}{x}_q^0, \text{so daß irgendein Punkt in } \mathrm{out}(\beta, \underline{\mu}, \Sigma) \\ & \qquad \text{nicht als } \alpha_{n+q}\text{–Ausreißer identifiziert wird}\}, \end{aligned}$$

$$k^M = \min\{q : \beta^M(\alpha_{n+q}, \underset{\sim}{x}_n^r, q, \delta) = 0\},$$

$$\varepsilon^M = \frac{k^M}{n + k^M}.$$

Betrachte eine Situation mit $k-1$ Ausreißern.

Sei $\underset{\sim}{x}_{N-1} = (\underset{\sim}{x}_n^r, \underset{\sim}{x}_{k-1}^0)$ und $\underset{\sim}{x}_{k-1}^0 = (\underline{x}_1^0, \ldots, \underline{x}_{k-1}^0)$ ein Tupel von δ–Ausreißern.

Für $k < N$ gilt $\frac{k-1}{N-1} < \frac{k}{N}$. Da hier $k < \frac{N}{2}$, ist dieser Zusammenhang erfüllt, und daher kann durch diese Konstellation der Stichprobe weder ein Zusammenbruch des Lokationsschätzers $\underline{m}$ noch ein Zusammenbruch des Kovarianzschätzers S erfolgen. Das bedeutet:

1. $\exists\, a \in \mathbb{N} : \|\underline{m}(\underset{\sim}{x}_{N-1})\| \leq a$,

2. $\exists\, b \in \mathbb{N} : 0 < \lambda_p(S) \leq \ldots \leq \lambda_1(S) \leq b$, wobei $\lambda_i(S)$ die Eigenwerte von $S(\underset{\sim}{x}_{N-1})$ sind, $i = 1, \ldots, p$.

(Die Aussagen 1. und 2. sind unmittelbare Folgerungen aus den Definitionen der Finite–sample Bruchpunkte.)

Damit gilt für $\underline{\mathrm{OR}}$ und das Volumen von $\mathbb{R}^p \backslash \underline{\mathrm{OR}}$:

$$\underline{\mathrm{OR}}(\underset{\sim}{x}_{N-1}, \alpha_{N-1}) \neq \emptyset, \qquad 0 < \mathrm{vol}(\mathbb{R}^p \backslash \underline{\mathrm{OR}}) < \infty,$$

und das Ellipsoid $\mathbb{R}^p \backslash \underline{\mathrm{OR}}$ liegt in einer abgeschlossenen Teilmenge des $\mathbb{R}^p$, d. h., es existiert eine Kugel Ku mit Radius r, $0 < r < \infty$, $Ku = \{\underline{x} \in \mathbb{R}^p : \|\underline{x}\| \leq r\}$, so daß $\mathbb{R}^p \backslash \underline{\mathrm{OR}} \subseteq Ku$, zum Beispiel $r = \|\underline{m}(\underset{\sim}{x}_{N-1})\| + d\sqrt{\lambda_1(S)} \leq a + d\sqrt{b}$. (Die Zahl d ist ein Proportionalitätsfaktor: das Volumen des Ellipsoids ist proportional zum Produkt der Wurzeln der Eigenwerte von S.)

Dann folgt:

$$\mathbb{R}^p \backslash Ku \subseteq \underline{\mathrm{OR}}(\underset{\sim}{x}_{N-1}, \alpha_{N-1}), \tag{3.1}$$

d. h., alle Punkte außerhalb der Kugel Ku werden von $\underline{\mathrm{OR}}$ als α_{N-1}–Ausreißer identifiziert.

Nun gibt es eine Zahl $\beta \in \,]0,1[$, so daß die Kugel Ku im Ellipsoid $\mathbb{R}^p \backslash \mathrm{out}(\beta, \underline{\mu}, \Sigma) = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \leq \chi^2_{p;1-\beta}\}$ enthalten ist, d. h.,

$$Ku \subseteq \mathbb{R}^p \backslash \mathrm{out}(\beta, \underline{\mu}, \Sigma). \tag{3.2}$$

Die größte Zahl β, für die der Zusammenhang (3.2) erfüllt ist, sei mit β^* bezeichnet. Dann ist

$$\text{out}(\beta^*, \underline{\mu}, \Sigma) \quad \subseteq \quad I\!R^p \backslash Ku \quad \subseteq \quad \underline{\text{OR}}(\underset{\sim}{x}_{N-1}, \alpha_{N-1})$$

(unter Ausnutzung von (3.1) und (3.2)), und damit wird jeder β^*–Ausreißer als α_{N-1}–Ausreißer identifiziert. Gleiches gilt für alle Zahlen $\beta < \beta^*$.
Damit folgt sofort, daß $\beta^M(\alpha_{N-1}, \underset{\sim}{x}_n^r, k-1, \delta) \geq \beta^* > 0$ ist (nach Definition von β^M).
Dieselben Schritte sind für alle Zahlen l mit $0 \leq l < k$ anstelle von $k-1$ möglich.
Das bedeutet für den Masking-Bruchpunkt:

$$\varepsilon^M(\underline{\text{OR}}(\underset{\sim}{x}_{N-1}, \alpha_{N-1})) \ \geq \ \frac{k-1}{n+k-1},$$

und damit gilt (nach Definition von ε^M):

$$\varepsilon^M(\underline{\text{OR}}(\underset{\sim}{x}_N, \alpha_N)) \ \geq \ \frac{k}{n+k} = \frac{k}{N}.$$

□

Mit Satz 3.1 ist es gelungen, den Masking–Bruchpunkt eines Identifizierers nach unten abzuschätzen. Auf ähnliche Weise läßt sich eine Abschätzung nach oben gewinnen, die statt auf dem Minimum auf dem Maximum der beiden Finite–sample Bruchpunkte der Schätzer $\underline{m}$ und S beruht.

Satz 3.2

Seien $\underline{\text{OR}}$, $\underline{m}$ und S wie in Satz 3.1. Seien $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) =: \frac{k_1}{N}$ und $\varepsilon^*(\underset{\sim}{x}_N, S) =: \frac{k_2}{N}$ die Finite–sample Bruchpunkte der beiden Schätzer, wobei $k_i < \frac{N}{2}$ gelte, $i = 1, 2$. Seien weiter $K := \max\{k_1, k_2\}, \boldsymbol{\alpha} = (\alpha_N)_{N \in I\!N}, 0 < \alpha_N < 1$, und $\delta \in]0, 1[$. Sei $\underset{\sim}{x}_n^r = (\underline{x}_1^r, \ldots, \underline{x}_n^r)$ eine Stichprobe regulärer Beobachtungen aus $N(\underline{\mu}, \Sigma)$–verteilten Zufallsvariablen.
Dann gilt für den Masking–Bruchpunkt $\varepsilon^M(\underline{\text{OR}})$:

$$\varepsilon^M(\underline{\text{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta) \leq \frac{K}{N}$$

mit $N = n + K$.

Beweis

Annahme: Es gilt $\varepsilon^M(\underline{\mathrm{OR}}) > \frac{K}{N}$, das heißt, es gilt $\varepsilon^M(\underline{\mathrm{OR}}) \geq \frac{K+1}{N+1} = \frac{K+1}{n+K+1}$.
Da nach Definition von ε^M außerdem

$$\varepsilon^M = \frac{k^M}{n + k^M}$$

mit

$$k^M = \min\{q : \beta^M(\alpha_{n+q}, \underset{\sim}{x}_n^r, q, \delta) = 0\},$$

muß für $\varepsilon^M(\underline{\mathrm{OR}}) \geq \frac{K+1}{n+K+1}$ die Abschätzung $k^M \geq K + 1$ gelten.
Damit folgt, daß eine Zahl $\beta^* > 0$ existiert, so daß

$$\beta^M(\alpha_{n+K}, \underset{\sim}{x}_n^r, K, \delta) > \beta^*$$

für jede beliebige Kombination von K Beobachtungen, die als δ–Ausreißer an einer für den Identifizierer $\underline{\mathrm{OR}}$ möglichst ungünstigen Position plaziert werden.
Nach Definition ist

$$\begin{aligned}\beta^M(\alpha_{n+K}, \underset{\sim}{x}_n^r, K, \delta) \;=\; & \inf\{\beta > 0 : \text{es ex. } \delta\text{–Ausreißer } \underset{\sim}{x}_K^0, \text{so daß irgendein Punkt} \\ & \text{in } \mathrm{out}(\beta, \underline{\mu}, \Sigma) \text{ nicht als } \alpha_{n+K}\text{–Ausreißer identifiziert} \\ & \text{wird}\}.\end{aligned}$$

Mit $\beta^M(\alpha_{n+K}, \underset{\sim}{x}_n^r, K, \delta) > \beta^* > 0$ muß daher für die Zahl β^* gelten: bei jeder beliebigen Kombination $\underset{\sim}{x}_K^0$ von δ–Ausreißern werden alle Punkte im Bereich $\mathrm{out}(\beta^*, \underline{\mu}, \Sigma)$ als α_{n+K}–Ausreißer identifiziert. Das bedeutet, daß

$$\mathrm{out}(\beta^*, \underline{\mu}, \Sigma) \subseteq \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N)$$

mit $N = n + K$ gelten muß. Anders ausgedrückt, muß

$$I\!\!R^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N) \subseteq I\!\!R^p \backslash \mathrm{out}(\beta^*, \underline{\mu}, \Sigma)$$

gelten für beliebige $\underset{\sim}{x}_K^0$.
Das Ellipsoid $I\!\!R^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N)$ liegt also innerhalb einer abgeschlossenen Teilmenge des $I\!\!R^p$, denn es läßt sich eine abgeschlossene Kugel finden, so daß $I\!\!R^p \backslash \mathrm{out}(\beta^*, \underline{\mu}, \Sigma)$ innerhalb dieser Kugel liegt.

Dann folgt insbesondere, daß der Mittelpunkt $\underline{m}$ des Ellipsoids $I\!R^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N)$ innerhalb einer abgeschlossenen Teilmenge des $I\!R^p$ liegt. Damit kann kein Zusammenbruch des Schätzers $\underline{m} = \underline{m}(\underset{\sim}{x}_N)$ erfolgen, es muß

$$\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) > \frac{K}{N}$$

gelten.

Nach Voraussetzung ist aber $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) \leq \frac{K}{N}$, es ergibt sich also ein Widerspruch. Damit gilt die Behauptung, das heißt, es ist

$$\varepsilon^M(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta) \leq \frac{K}{N}$$

mit $N = n + K$.

□

Faßt man die Ergebnisse der beiden vorangegangenen Sätze zusammen, so erhält man:

$$\min(\varepsilon^*(\underset{\sim}{x}_N, \underline{m}), \varepsilon^*(\underset{\sim}{x}_N, S)) \leq \varepsilon^M(\underline{\mathrm{OR}}) \leq \max(\varepsilon^*(\underset{\sim}{x}_N, \underline{m}), \varepsilon^*(\underset{\sim}{x}_N, S)).$$

Die Finite–sample Bruchpunkte der Schätzer begrenzen also den Masking–Bruchpunkt des Identifizierers.

Korollar 3.1

Falls unter den Voraussetzungen der Sätze 3.1 und 3.2 die Finite–sample Bruchpunkte der Schätzer $\underline{m}$ und S übereinstimmen, dann folgt für den Masking–Bruchpunkt des Identifizierers $\underline{\mathrm{OR}}$:

$$\varepsilon^M(\underline{\mathrm{OR}}) = \varepsilon^*(\underset{\sim}{x}_N, \underline{m}) = \varepsilon^*(\underset{\sim}{x}_N, S).$$

Mit Hilfe der Sätze 3.1 und 3.2 läßt sich eine Abschätzung über die Größe des maximal erreichbaren Masking–Bruchpunkts angeben. Diese gilt für solche Stichproben, deren reguläre Beobachtungen in sogenannter *allgemeiner Lage* sind:

Definition 3.4 (Rousseeuw (1985), S. 288)

Sei $\underset{\sim}{x}_n = (\underline{x}_1, \ldots, \underline{x}_n)$ eine Stichprobe vom Umfang $n, n \in \mathbb{N}$, mit $\underline{x}_i \in \mathbb{R}^p, i = 1, \ldots, n$, sei $n > p$. Die Stichprobe $\underset{\sim}{x}_n$ heißt *in allgemeiner Lage*

$:\Leftrightarrow$ in jedem $(p-1)$–dimensionalen affinen Unterraum des $\mathbb{R}^p$

liegen höchstens p Punkte von $\underset{\sim}{x}_n$.

Satz 3.3

Es gelten die gleichen Voraussetzungen wie in Satz 3.1. Weiterhin sei die Stichprobe $\underset{\sim}{x}_n^r$ der regulären Beobachtungen in allgemeiner Lage, und es sei $n \geq p+1$. Dann gilt für den maximal erreichbaren Masking–Bruchpunkt $\varepsilon_{\max}^M = \max_{\underline{\mathrm{OR}}} \varepsilon^M(\underline{\mathrm{OR}})$ eines affin äquivarianten multivariaten Ausreißer–Identifizierers $\underline{\mathrm{OR}}$:

$$\frac{\left[\frac{N-p+1}{2}\right]}{N} \leq \varepsilon_{\max}^M \leq \frac{1}{2}.$$

Dabei bezeichne $[x]$ den ganzzahligen Anteil von x, $x \in \mathbb{R}$.

Beweis

Nach Aussage von Satz 3.1 gilt:

$$\varepsilon^M(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta) \geq \frac{k}{N} \quad \text{mit} \quad N = n + k.$$

Also folgt:

$$\begin{aligned} \varepsilon^M(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta) &\geq \min\left\{ \frac{k_1}{N}, \frac{k_2}{N} \right\} \\ &= \min\{ \varepsilon^*(\underset{\sim}{x}_N, \underline{m}), \varepsilon^*(\underset{\sim}{x}_N, S) \}. \end{aligned}$$

Daraus ergibt sich für den maximal möglichen Masking–Bruchpunkt $\varepsilon_{\max}^M$:

$$\varepsilon_{\max}^M \geq \min\{ \max_{\underline{m}} \varepsilon^*(\underset{\sim}{x}_N, \underline{m}), \max_S \varepsilon^*(\underset{\sim}{x}_N, S) \}.$$

Für die Finite–sample–Bruchpunkte von affin äquivarianten Schätzern gilt:

$$\frac{\left[\frac{N-p+1}{2}\right]}{N} \leq \max_{\underline{m}} \varepsilon^*(\underset{\sim}{x}_N, \underline{m}) \leq \frac{\left[\frac{N+1}{2}\right]}{N}$$

(die untere Grenze wird z. B. von bestimmten S–Schätzern angenommen (vgl. Abschnitt 4), die obere Grenze ist bisher „nur" Abschätzung (vgl. Lopuhaä, Rousseeuw (1991)), sie ist scharf für translationsäquivariante Schätzer),

$$\max_S \varepsilon^*(\underset{\sim}{x}_N, S) \quad = \quad \frac{\left[\frac{N-p+1}{2}\right]}{N}$$

(dieses Maximum wird z. B. vom MVE–Kovarianzschätzer und von bestimmten S–Schätzern angenommen (vgl. Davies (1987), S. 1288)).
Für diese Abschätzungen der Finite–sample–Bruchpunkte werden die Voraussetzungen der allgemeinen Lage und von $n \geq p+1$ benötigt, siehe z. B. Davies (1987, S. 1289).
Damit folgt sofort, daß

$$\varepsilon^M_{\max} \quad \geq \quad \frac{\left[\frac{N-p+1}{2}\right]}{N}.$$

Andererseits wurde bereits gezeigt (Becker (1992), S. 28 f.), daß für affin äquivariante multivariate Identifizierer stets

$$\varepsilon^M(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}^r_n, \delta) \leq \frac{1}{2}$$

gelten muß. Diese Abschätzung läßt sich auch mit Satz 3.2 herleiten, da die Zahl 1/2 eine obere Grenze für den maximalen Finite–sample Bruchpunkt affin äquivarianter Lokationsschätzer ist (vgl. Lopuhaä, Rousseeuw (1991), S. 232).
Also gilt die Behauptung, und es ist

$$\frac{\left[\frac{N-p+1}{2}\right]}{N} \quad \leq \quad \varepsilon^M_{\max} \quad \leq \quad \frac{1}{2}.$$

□

Damit ist die Möglichkeit gegeben zu beurteilen, wie „gut" sich ein beliebiger Ausreißer-Identifizierer $\underline{\mathrm{OR}}$ bezüglich des Masking–Effekts verhält, verglichen mit dem bestmöglichen Ergebnis.
Es liegt nahe zu vermuten, daß die Zahl $\frac{\left[\frac{N-p+1}{2}\right]}{N}$, die eine obere Schranke für den Finite–sample Bruchpunkt von affin äquivarianten Kovarianzschätzern darstellt, auch eine obere Schranke für den Masking–Bruchpunkt affin äquivarianter Identifizierer ist. Man kann aber leicht sehen, daß diese Vermutung nicht uneingeschränkt bewiesen

werden kann. Der Grund dafür liegt im Verhalten von Kovarianzschätzern beim Zusammenbruch. Laut Definition kann der Zusammenbruch eines Kovarianzschätzers auf zwei verschiedene Arten erfolgen.

Einerseits besteht die Möglichkeit, daß der größte Eigenwert der Schätzmatrix S beliebig groß wird. Dieses Ereignis kann man einen Zusammenbruch durch „Explosion" nennen. Betrachtet man einen auf S beruhenden Identifizierer $\underline{\text{OR}}$, so bedeutet dies folgendes: die längste Achse des Ellipsoids $\mathbb{R}^p \setminus \underline{\text{OR}}$ wird beliebig lang, der Bereich, in dem keine Beobachtung als Ausreißer identifiziert wird, dehnt sich also in Richtung der ersten Hauptachse aus. Wie man sich leicht klarmachen kann, gelangen auf diese Weise irgendwann beliebig weit entfernte Ausreißer in diesen Bereich, so daß sie nicht identifiziert werden. Damit unterliegt der Identifizierer dem Masking–Effekt.

Andererseits kann aber auch der kleinste Eigenwert der Matrix S beliebig nahe an Null herankommen. In diesem Fall kann man von „Implosion" sprechen. Das Komplement des zugehörigen Identifizierers, das heißt, der Bereich, in dem keine Ausreißer identifiziert werden, schrumpft entlang der letzten Hauptachse (in Richtung des Eigenvektors zum kleinsten Eigenwert). Das p–dimensionale Ellipsoid $\mathbb{R}^p \setminus \underline{\text{OR}}$ verliert sozusagen eine Dimension, es degeneriert (im Grenzfall) zu einem $(p-1)$–dimensionalen Gebilde. Im zweidimensionalen Fall beispielsweise fällt die Ellipse zu einem Intervall zusammen. Dies führt dazu, daß Punkte, die keine Ausreißer sind, in den Bereich außerhalb des Ellipsoids fallen und somit als Ausreißer identifiziert werden. Das ist gleichbedeutend mit einem Swamping–Effekt.

Damit kann ein Zusammenbruch des Schätzers S durch Implosion nicht zur Bestimmung des Masking–Bruchpunkts des Identifizierers $\underline{\text{OR}}$ herangezogen werden.

Schließt man den Zusammenbruch durch Implosion aus, so kann man die oben angeführte Vermutung über den maximalen Masking–Bruchpunkt beweisen.

Satz 3.4

Sei $\underline{\text{OR}}$ ein multivariater Ausreißer–Identifizierer, basierend auf einem Lokationsschätzer $\underline{m}$ und einem Kovarianzschätzer S. Es existiere keine Kombination von Beobachtungen, die einen Zusammenbruch des Schätzers S durch seinen kleinsten Eigenwert verursacht. Dann gilt für den maximal möglichen Masking–Bruchpunkt von $\underline{\mathrm{OR}}$:

$$\varepsilon^M_{\max} \quad = \quad \frac{\left[\frac{N-p+1}{2}\right]}{N}.$$

Beweis

Nimmt man an, daß es ein Paar $(\underline{m}, S)$ von affin äquivarianten Schätzern derart gibt, daß der Masking–Bruchpunkt ε^M des auf ihnen basierenden Identifizierers $\underline{\mathrm{OR}}$ größer ist als $\frac{\left[\frac{N-p+1}{2}\right]}{N}$, dann muß nach Definition von ε^M für $k := \frac{\left[\frac{N-p+1}{2}\right]}{N}$ gelten, daß der Masking–Punkt β^M des Identifizierers größer als Null ist.
Wenn aber

$$\beta^M(\alpha_N, \underset{\sim}{x}^r_n, k, \delta) > 0,$$

dann gibt es für alle Stichproben $\underset{\sim}{x}_N = (\underset{\sim}{x}^r_n, \underset{\sim}{x}^0_k)$ eine (gemeinsame) Schranke $\beta_0 > 0$, so daß

$$\mathrm{out}(\beta_0, \underline{\mu}, \Sigma) \quad \subseteq \quad \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N),$$

also

$$\mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N) \quad \subseteq \quad \mathbb{R}^p \backslash \mathrm{out}(\beta_0, \underline{\mu}, \Sigma).$$

Damit muß für die Schätzer $\underline{m}$ und S, die das Ellipsoid $\mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N)$ definieren, gelten:

1. Der Schätzer $\underline{m}$ ist der Mittelpunkt des Ellipsoids, und damit ist $\underline{m} \in \mathbb{R}^p \backslash \mathrm{out}(\beta_0, \underline{\mu}, \Sigma)$. Dann kann man eine p–dimensionale Kugel finden, so daß $\underline{m}$ innerhalb dieser Kugel liegen muß, d. h., $\exists M \in \mathbb{R} : \|\underline{m}(\underset{\sim}{x}_N)\| \leq M$.
   Also bricht der Schätzer $\underline{m}$ nicht zusammen. Das bedeutet für den Finite–sample Bruchpunkt von $\underline{m}$, daß $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) > \frac{\left[\frac{N-p+1}{2}\right]}{N}$ sein muß.
   Daraus läßt sich noch kein Widerspruch ableiten, da für Lokationsschätzer diese Schranke überschritten werden kann (vgl. dazu Lophuhaä, Rousseeuw (1991), S. 234).

2. Für das Volumen des Ellipsoids gilt $\mathrm{vol}(\mathbb{R}^p \backslash \underline{\mathrm{OR}}) \leq \mathrm{vol}(\mathbb{R}^p \backslash \mathrm{out})$. Nun existieren Konstanten $const_1$, $const_2 \in \mathbb{R}$, so daß $\mathrm{vol}(\mathbb{R}^p \backslash \underline{\mathrm{OR}}) = const_1(\det(S))^{1/2}$ und

$\mathrm{vol}(\mathbb{R}^p \setminus \mathrm{out}) = const_2(\det(\Sigma))^{1/2}$ ist.

Damit ist $\det(S) \leq \left(\frac{const_2}{const_1}\right)^2 \cdot \det(\Sigma) =: d$.

Andererseits ist $\det(S) = \prod_{i=1}^{p} \lambda_i$, wobei λ_i die Eigenwerte von S sind mit $\lambda_p \leq \ldots \leq \lambda_1$.

Also muß es eine Zahl $K \in \mathbb{R}$ geben, so daß $\lambda_1 \leq K$ ist.

Nach Voraussetzung gibt es keine Kombination von Beobachtungen, die einen Zusammenbruch des Kovarianzschätzers S durch seinen kleinsten Eigenwert verursacht. Das bedeutet insgesamt, daß der Kovarianzschätzer S nicht zusammenbricht, und das ist ein Widerspruch zu der von Davies (1987) gezeigten gültigen Abschätzung.

□

Wie bereits erläutert, kann die Größe des maximal möglichen Masking–Bruchpunkts eines affin äquivarianten Identifizierers im allgemeinen Fall nicht schärfer abgeschätzt werden. In einem Spezialfall kann man jedoch den Masking–Bruchpunkt eines Identifizierers nach oben genauer abschätzen. Wenn nämlich der Bruchpunkt des Lageschätzers kleiner ist als der des Kovarianzschätzers, dann bestimmt der Lageschätzer das Verhalten des Identifizierers.

Satz 3.5

Sei $\underline{\mathrm{OR}}$ ein multivariater Ausreißer–Identifizierer, basierend auf den affin äquivarianten Schätzern $\underline{m}$ und S.

Für die Finite–sample Bruchpunkte von $\underline{m}$ und S gelte $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) = \frac{k_1}{N}, \varepsilon^*(\underset{\sim}{x}_N, S) = \frac{k_2}{N}$, und es sei $k_1 < k_2$.

Weiter sei $\underset{\sim}{x}_n^r = (\underline{x}_1^r, \ldots, \underline{x}_n^r)$ eine Stichprobe regulärer Beobachtungen von $N(\underline{\mu}, \Sigma)$–verteilten Zufallsvariablen.

Dann gilt für den Masking–Bruchpunkt des Identifizierers $\underline{\mathrm{OR}}$:

$$\varepsilon^M(\underline{\mathrm{OR}}) \quad \leq \quad \frac{k_1}{N}$$

mit $N = n + k_1$.

Beweis

Sei o.B.d.A. $\underline{\mu} = \underline{0}, \Sigma = \mathcal{I}$.

Betrachte eine Stichprobe $\underset{\sim}{x}_N = (\underset{\sim}{x}_n^r, \underset{\sim}{x}_k^0)$ mit n regulären Beobachtungen von $N(\underline{0}, \mathcal{I})$–verteilten Zufallsvariablen und einer Anzahl $k = k_1$ von δ–Ausreißern.

Da die Anzahl der δ–Ausreißer in der Stichprobe $\underset{\sim}{x}_N$ gerade k_1 mit $k_1 < k_2$ ist, bricht der Kovarianzschätzer S nicht zusammen. Das heißt, es gilt für die Eigenwerte $\lambda_i(S)$, $i = 1, \ldots, p$, der Matrix S:

$$\exists\, C \in \mathbb{N} : \; 0 < \lambda_p(S) \leq \ldots \leq \lambda_1(S) < C,$$

und die Inverse S^{-1} existiert.

Damit folgt für das Ellipsoid $\mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N)$:

$$\mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N) = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{m})^T S^{-1} (\underline{x} - \underline{m}) < c(p, N, \alpha_N)\} \neq \emptyset.$$

Also besitzt das Ellipsoid ein nichtleeres Volumen. Insbesondere existiert ein Punkt $\underline{x} \in \mathbb{R}^p$, so daß $(\underline{x} - \underline{m})^T S^{-1} (\underline{x} - \underline{m}) < c(p, N, \alpha_N)$ gilt, z. B. $\underline{x} = \underline{m}$.

Da aber der Finite–sample Bruchpunkt $\varepsilon^*(\underline{m}, \underset{\sim}{x}_N) = \frac{k_1}{N} = \frac{k}{N}$ ist, existiert zu jeder Zahl $D \in \mathbb{R}$ eine Konstellation $\underset{\sim}{x}_k^0 = \underset{\sim}{x}_k^0(D)$ von δ–Ausreißern, so daß für die Stichprobe $\underset{\sim}{x}_N = (\underset{\sim}{x}_n^r, \underset{\sim}{x}_k^0(D))$ gilt:

$$\underline{m}(\underset{\sim}{x}_N)^T \underline{m}(\underset{\sim}{x}_N) > D.$$

Dies entspricht gerade dem Zusammenbruch des Lokationsschätzers.

Dieser Zusammenhang läßt sich insbesondere herstellen für $D = \chi^2_{p;1-\beta}$ mit beliebigen Werten von β.

Damit gibt es zu jeder Zahl $\beta, 0 < \beta < 1$, eine Menge von δ–Ausreißern $\underset{\sim}{x}_k^0$ mit $k = k_1$, so daß für den auf der Stichprobe $\underset{\sim}{x}_N = (\underset{\sim}{x}_n^r, \underset{\sim}{x}_k^0)$ beruhenden Lokationsschätzer $\underline{m}$ gilt:

1. $\underline{m} \in \mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N)$,

2. $\underline{m}^T \underline{m} > \chi^2_{p;1-\beta}$, d. h., $\underline{m} \in \mathrm{out}(\beta, \underline{0}, \mathcal{I})$.

Diese beiden Aussagen beschreiben gerade die Tatsache, daß der Identifizierer $\underline{\mathrm{OR}}$ dem Masking–Effekt unterliegt.

Damit folgt die Behauptung, es ist also

$$\varepsilon^M(\underline{\mathrm{OR}}) \quad \leq \quad \frac{k_1}{N} = \frac{k_1}{n+k_1}.$$

□

Korollar 3.2

Aus Satz 3.5 folgt unter Berücksichtigung von Satz 3.1 unmittelbar, daß in dem Fall, in dem $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) < \varepsilon^*(\underset{\sim}{x}_N, S)$ ist, der Masking–Bruchpunkt des Identifizierers gleich dem Finite–sample Bruchpunkt des Lokationsschätzers sein muß:

$$\varepsilon^M(\underline{\mathrm{OR}}) = \varepsilon^*(\underset{\sim}{x}_N, \underline{m}) = \frac{k_1}{N}.$$

3.3 Beziehungen zwischen Bruchpunkten von Schätzern und dem Swamping–Bruchpunkt von Ausreißer–Identifizierern

Der Begriff des Swamping bedeutet ebenso wie der des Masking ein Fehlverhalten eines Identifizierers. Im Gegensatz zum Masking werden beim Swamping Datenpunkte, die keine Ausreißer sind, vom Identifizierer als beliebig „große“ (d. h., beliebig weit entfernte) Ausreißer eingestuft. Wie beim Masking, so läßt sich auch beim Swamping–Bruchpunkt eine unmittelbare Beziehung zu den Bruchpunkten der in den Identifizierer eingehenden Schätzer herleiten. Während für die Ergebnisse des vorigen Abschnitts die gewählte Normierung der Prozeduren nicht explizit benötigt wird, muß bei der Herleitung der folgenden Aussage die Normierung des Identifizierers berücksichtigt werden.

Satz 3.6

Seien $\underline{\mathrm{OR}}$, $\underline{m}$, S wie bisher, seien weiter $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) = \frac{k_1}{N}, \varepsilon^*(\underset{\sim}{x}_N, S) = \frac{k_2}{N}$, und sei $k := \min\{k_1, k_2\}$. Der Identifizierer $\underline{\mathrm{OR}}$ sei normiert gemäß einer der beiden Bedingungen

$$\mathrm{P}(\underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N) \subseteq \mathrm{out}(\alpha_N, \underline{\mu}, \Sigma)) = 1 - \alpha$$

für eine gegebene Zahl $\alpha \in]0,1[$ oder

$$\mathrm{P}(\underline{X}_i \in \mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha$$

für $\alpha \in]0,1[$ und $\alpha_N = 1 - (1-\alpha)^{1/N}$

mit $\underset{\sim}{X}_N = (\underline{X}_1, \ldots, \underline{X}_N)$, $\underline{X}_1, \ldots, \underline{X}_N$ stochastisch unabhängig und identisch $N(\underline{\mu}, \Sigma)$–verteilt.

Dann gilt für den Swamping–Bruchpunkt von $\underline{\mathrm{OR}}$:

$$\varepsilon^S(\underline{\mathrm{OR}}) \geq \frac{k}{N}.$$

Beweis

Betrachte zunächst die Normierungsbedingungen. Für die erste Bedingung gilt:

$$\begin{aligned} & \mathrm{P}(\underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N) \subseteq \mathrm{out}(\alpha_N, \underline{\mu}, \Sigma)) = 1 - \alpha \\ \Leftrightarrow\quad & \mathrm{P}(\mathbb{R}^p \backslash \mathrm{out}(\alpha_N, \underline{\mu}, \Sigma) \subseteq \mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N)) = 1 - \alpha. \end{aligned}$$

Nun ist $\mathbb{R}^p \backslash \mathrm{out}(\alpha_N, \underline{\mu}, \Sigma) = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \leq \chi^2_{p;1-\alpha_N}\}$
und $\mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N) = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{m})^T S^{-1} (\underline{x} - \underline{m}) < c(p, N, \alpha_N)\}$.
Für die zweite Bedingung folgt aus dem Zusammenhang $\alpha_N = 1 - (1-\alpha)^{1/N}$:

$$\begin{aligned} & \mathrm{P}(\underline{X}_i \in \mathbb{R}^p \backslash \mathrm{out}(\alpha_N, \underline{\mu}, \Sigma), i = 1, \ldots, N) \\ & = \mathrm{P}(\underline{X}_i \in \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{\mu})^T \Sigma^{-1} (\underline{x} - \underline{\mu}) \leq \chi^2_{p;1-\alpha_N}\}, i = 1, \ldots, N) = 1 - \alpha. \end{aligned}$$

Gleichzeitig soll

$$\begin{aligned} & \mathrm{P}(\underline{X}_i \in \mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) \\ & = \mathrm{P}(\underline{X}_i \in \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{m})^T S^{-1} (\underline{x} - \underline{m}) < c(p, N, \alpha_N)\}, i = 1, \ldots, N) \\ & = 1 - \alpha \end{aligned}$$

erfüllt sein.

Falls α_N sehr klein ist, kann jede dieser Normierungsbedingungen nur dann erfüllt werden, wenn $c(p, N, \alpha_N)$ groß genug ist. Genauer muß gelten: $c(p, N, \beta) \to \infty$ für $\beta \to 0$. Damit läßt sich zu jeder vorgegebenen Schranke $d \in \mathbb{R}$ eine Zahl $\beta = \beta(d) > 0$ finden, so daß $c(p, N, \beta) > d$ gilt.

Sei nun o.B.d.A. $\underline{\mu} = \underline{0}$ und $\Sigma = \mathcal{I}$. Betrachte eine Stichprobe $\underset{\sim}{x}_{N-1}$ mit n regulären

Beobachtungen und einer Anzahl $k-1$ von δ–Ausreißern, $N-1 = n+k-1$.

Wie im Beweis von Satz 3.1 kann man überlegen, daß für die Schätzer $\underline{m}$ und S gelten muß:

1. $\exists\, a \in \mathbb{N} : \|\underline{m}(\underset{\sim}{x}_{N-1})\| \leq a$,

2. $\exists\, b \in \mathbb{N} : 0 < \lambda_p(S) \leq \ldots \leq \lambda_1(S) \leq b$, wobei $\lambda_i(S)$ die Eigenwerte von $S(\underset{\sim}{x}_{N-1})$ sind, $i = 1, \ldots, p$.

Der Mittelpunkt des Ellipsoids $\mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_{N-1}, \beta)$ ist gerade $\underline{m}(\underset{\sim}{x}_{N-1})$. Nach Aussage 1. kann dieser Mittelpunkt höchstens den Abstand a vom Ursprung haben.

Der Bereich $\mathbb{R}^p \backslash \mathrm{out}(\alpha_{N-1}, \underline{0}, \mathcal{I})$ ist eine Kugel um den Ursprung mit Radius $(\chi^2_{p;1-\alpha_{N-1}})^{1/2}$. Dieser Bereich ist in einer Kugel K mit Mittelpunkt $\underline{m}$ und Radius $a + (\chi^2_{p;1-\alpha_{N-1}})^{1/2}$ enthalten, $K = \{\underline{x} \in \mathbb{R}^p : \|\underline{x} - \underline{m}\| \leq a + (\chi^2_{p;1-\alpha_{N-1}})^{1/2}\}$.

Durch geeignete Wahl von β kann nun das Ellipsoid $\mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_{N-1}, \beta)$ so weit „aufgebläht“ werden, daß es die Kugel K enthält. Dazu reicht es, die kürzeste Hauptachse von $\mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_{N-1}, \beta)$ auf eine Länge von mindestens $a + (\chi^2_{p;1-\alpha_{N-1}})^{1/2}$ zu strecken.

Die kürzeste Hauptachse von $\mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_{N-1}, \beta)$ hat die Länge $(\lambda_p(S))^{1/2}(c(p,N,\beta))^{1/2}$. Dabei ist nach Aussage 2. der Eigenwert $\lambda_p(S)$ größer als Null, da S nicht zusammenbricht. Um die Hauptachse auf die benötigte Länge zu bringen, muß

$$(c(p,N,\beta))^{1/2} \geq \frac{1}{(\lambda_p(S))^{1/2}}(a + (\chi^2_{p;1-\alpha_{N-1}})^{1/2})$$

beziehungsweise

$$c(p,N,\beta) \geq \frac{1}{\lambda_p(S)}(a^2 + 2a(\chi^2_{p;1-\alpha_{N-1}})^{1/2} + \chi^2_{p;1-\alpha_{N-1}})$$

gelten. Wegen der Normierungsbedingung an den Identifizierer läßt sich eine Zahl $\beta^* > 0$ finden, so daß die obige Bedingung erfüllt ist. Damit gilt:

$$\mathbb{R}^p \backslash \mathrm{out}(\alpha_{N-1}, \underline{0}, \mathcal{I}) \subseteq K \subseteq \mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_{N-1}, \beta^*),$$

insbesondere existiert also eine Zahl $\beta^* > 0$, so daß

$$\mathbb{R}^p \backslash \mathrm{out}(\alpha_{N-1}, \underline{0}, \mathcal{I}) \subseteq \mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_{N-1}, \beta^*).$$

Das bedeutet, daß keine Beobachtung, die kein α_{N-1}–Ausreißer ist, als β^*–Ausreißer identifiziert wird. Für alle $\beta \leq \beta^*$ findet also kein Swamping statt.

Damit ist $\beta^S(\underline{\text{OR}}, \alpha_{N-1}, \underset{\sim}{x}^r_n, k-1, \delta) > 0$ und $k^S(\underline{\text{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}^r_n, \delta) > k-1$. Eine entsprechende Argumentation ist für alle Zahlen l mit $0 \le l < k$ anstelle von $k-1$ möglich. Also folgt die Behauptung, und es ist

$$\varepsilon^S(\underline{\text{OR}}) \ge \frac{k}{N}.$$

□

Auch den Swamping–Bruchpunkt kann man nach oben abschätzen, wenn der Bruchpunkt des Lokationsschätzers den des Kovarianzschätzers unterschreitet.

Satz 3.7

Seien $\underline{\text{OR}}$, $\underline{m}$, S wie bisher, seien weiter $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) = \frac{k_1}{N}, \varepsilon^*(\underset{\sim}{x}_N, S) = \frac{k_2}{N}$, und es gelte $k_1 < k_2$. Dann gilt für den Swamping–Bruchpunkt des Identifizierers $\underline{\text{OR}}$:

$$\varepsilon^S(\underline{\text{OR}}) \le \frac{k_1}{N}.$$

Beweis

Analog zum Beweis von Satz 3.5 betrachte o.B.d.A. den Fall $\underline{\mu} = \underline{0}, \Sigma = \mathcal{I}$ und eine Stichprobe $\underset{\sim}{x}_N = (\underset{\sim}{x}^r_n, \underset{\sim}{x}^0_k)$ mit n regulären Beobachtungen und einer Anzahl $k = k_1$ von δ–Ausreißern.

Mit der gleichen Argumentation wie dort folgt, daß die Eigenwerte der Matrix S durch eine Zahl C nach oben beschränkt sind und die Inverse S^{-1} existiert.

Dann gibt es zu jeder Zahl $\beta, 0 < \beta < 1$, eine Zahl $D = D(\beta) \in \mathbb{R}$, so daß das Volumen des Ellipsoids $\mathbb{R}^p \backslash \underline{\text{OR}}(\underset{\sim}{x}_N, \beta)$ durch D beschränkt wird:

$$\text{vol}(\mathbb{R}^p \backslash \underline{\text{OR}}(\underset{\sim}{x}_N, \beta)) \le D.$$

Damit läßt sich $\mathbb{R}^p \backslash \underline{\text{OR}}(\underset{\sim}{x}_N, \beta)$ in eine Kugel K mit Mittelpunkt $\underline{m}$ und festem Radius $r = r(\beta), 0 < r < \infty$, einbetten, $K = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{m})^T (\underline{x} - \underline{m}) \le r^2\}$.

Da der Finite–sample Bruchpunkt $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) = \frac{k_1}{N} = \frac{k}{N}$ ist, läßt sich eine Konstellation $\underset{\sim}{x}^0_k = \underset{\sim}{x}^0_k(r)$ von δ–Ausreißern finden, so daß für $\underset{\sim}{x}_N = (\underset{\sim}{x}^r_n, \underset{\sim}{x}^0_k)$ gilt:

$$\underline{m}(\underset{\sim}{x}_N)^T \underline{m}(\underset{\sim}{x}_N) > r^2.$$

Damit folgt sofort, daß

$$(\underline{0} - \underline{m}(\underset{\sim}{x}_N))^T(\underline{0} - \underline{m}(\underset{\sim}{x}_N)) > r^2$$

ist und somit der Punkt $\underline{x} = \underline{0}$ nicht in der Kugel K enthalten ist.
Dann ist aber $\underline{0} \notin I\!R^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \beta)$ bzw.

$$\underline{0} \in \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \beta).$$

Andererseits ist $I\!R^p \backslash \mathrm{out}(\alpha_N, \underline{0}, \mathcal{I}) = \{\underline{x} \in I\!R^p : \underline{x}^T \underline{x} \leq \chi^2_{p;1-\alpha_N}\}$ und daher

$$\underline{0} \in I\!R^p \backslash \mathrm{out}(\alpha_N, \underline{0}, \mathcal{I}).$$

Damit findet man zu jeder Zahl $\beta, 0 < \beta < 1$, eine Menge $\underset{\sim}{x}_k^0$ von δ–Ausreißern mit $k = k_1$, so daß gilt:

1. $\underline{0} \in \underline{\mathrm{OR}}(\underset{\sim}{x}_N, \beta)$,

2. $\underline{0} \notin \mathrm{out}(\alpha_N, \underline{0}, \mathcal{I})$.

Dies bedeutet, daß der Identifizierer $\underline{\mathrm{OR}}$ dem Swamping–Effekt unterliegt.
Also gilt

$$\varepsilon^S(\underline{\mathrm{OR}}) \leq \frac{k_1}{N} = \frac{k_1}{n + k_1}.$$

□

Aus den beiden vorangegangenen Sätzen und Korollar 3.2 folgt unmittelbar der folgende Zusammenhang.

Korollar 3.3

Falls für die Finite–sample Bruchpunkte der Schätzer $\underline{m}$ und S gilt, daß $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) = \frac{k_1}{N} < \varepsilon^*(\underset{\sim}{x}_N, S) = \frac{k_2}{N}$ ist, ist der Swamping–Bruchpunkt des Identifizierers gleich dem Finite–sample Bruchpunkt des Lokationsschätzers. In diesem Fall stimmen zudem Masking- und Swamping–Bruchpunkt des Identifizierers überein:

$$\varepsilon^S(\underline{\mathrm{OR}}) = \varepsilon^*(\underset{\sim}{x}_N, \underline{m}) = \frac{k_1}{N} = \varepsilon^M(\underline{\mathrm{OR}}).$$

Betrachtet man Aussage und Beweis von Satz 3.7, so fällt die starke Ähnlichkeit zu Satz 3.5 auf. In der Tat wird in beiden Fällen dieselbe Argumentation zum Beweis der Behauptung verwendet: es wird eine Stichprobe derart konstruiert, daß sich die beiden Ellipsoide $\mathbb{R}^p \backslash \mathrm{out}$ und $\mathbb{R}^p \backslash \underline{\mathrm{OR}}$ nicht schneiden. In diesem Fall treten Masking- und Swamping–Effekt gleichzeitig auf, und der Identifizierer bricht bezüglich beider Kriterien zusammen. Beide Sätze behandeln die Situation, daß der Finite–sample Bruchpunkt des beteiligten Lokationsschätzers kleiner ist als der des Kovarianzschätzers. Wie man sieht, ist dann der Lokationsschätzer ausschlaggebend sowohl für das Masking- als auch für das Swamping–Verhalten des Identifizierers. Anders ausgedrückt, kann ein schlechter Finite–sample Bruchpunkt des Lokationsschätzers nicht durch einen noch so guten Bruchpunkt des Kovarianzschätzers aufgefangen werden. Ist man also daran interessiert, Identifizierer mit hohen Masking- und Swamping–Bruchpunkten zu konstruieren, so muß der verwendete Lokationsschätzer bezüglich des Kriteriums „Zusammenbruch“ mindestens so gut sein wie der Kovarianzschätzer, und beide sollten einen möglichst hohen Finite–sample Bruchpunkt aufweisen.

3.4 Bruchpunkteigenschaften von Identifizierern, die auf bestimmten Schätzern beruhen

Schätzer, die das im vorigen Abschnitt geforderte Kriterium eines hohen Bruchpunkts erfüllen, sind beispielsweise bestimmte S–Schätzer. S–Schätzer wurden von Rousseeuw und Yohai (1984) im Rahmen der robusten Regression eingeführt und von Davies (1987) auf den Fall eines multivariaten Lokations–Skalenmodells erweitert. Hier soll der Spezialfall einer zugrundeliegenden multivariaten Normalverteilung betrachtet werden.

Definition 3.5 (Davies (1987), S. 1270 f.)

Seien $\underline{X}_1, \ldots, \underline{X}_N$ unabhängige, identisch $N(\underline{\mu}, \Sigma)$–verteilte Zufallsvariablen.
Sei $\kappa : \mathbb{R}_+ \to [0,1]$ eine monoton fallende, linksseitig stetige Funktion, die die folgenden Voraussetzungen erfüllt:

$\kappa(0) = 1$,

κ ist stetig an der Stelle 0,

und es existiert eine Zahl $c > 0$, so daß

$\kappa(u) > 0, \quad 0 \le u \le c,$

$\kappa(u) = 0, \quad u > c.$

Alle Lösungen $\underline{m}_{\mathrm{S}}$ und S_{S} des Minimierungsproblems

$$\min_{S \in \mathrm{PDS}(p)} \det(S)$$

unter der Bedingung

$$\frac{1}{N}\sum_{i=1}^{N} \kappa((\underline{X}_i - \underline{m})^T S^{-1}(\underline{X}_i - \underline{m})) \;\ge\; 1-\varepsilon$$

heißen *S–Schätzer für* $\underline{\mu}$ *und* Σ.

Dabei sei PDS(p) die Menge aller positiv definiten, symmetrischen Matrizen des $I\!R^{p\times p}$, und es gelte

$$1-\varepsilon = E(\kappa((\underline{X}-\underline{\mu})^T\Sigma^{-1}(\underline{X}-\underline{\mu}))).$$

Die auf diese Weise definierten S–Schätzer sind affin äquivariant (Davies (1987), S. 1271). Wählt man speziell die Funktion κ so, daß für die Zahl ε die Beziehung $\varepsilon = \frac{1}{2} - \frac{p+1}{2N}$ gilt, so ergeben sich Schätzer $\underline{m}$ und S mit dem für S–Schätzer maximal möglichen Finite–sample Bruchpunkt von $\left[\frac{N-p+1}{2}\right]/N$ (Davies (1987), S. 1288).

Diese S–Schätzer seien im folgenden als $\mathrm{S_{MB}}$–Schätzer (für „S–Schätzer mit **m**aximalem **B**ruchpunkt") bezeichnet.

Für affin äquivariante Kovarianzschätzer ist dies auch allgemein der höchste erreichbare Bruchpunkt (Davies (1987), S. 1289), für Lokationsschätzer scheint dieser Wert noch nicht die obere Grenze zu sein (Lopuhaä, Rousseeuw (1991), S. 234).

Ein weiteres Paar von Schätzern, das dieselben Bruchpunkte erreicht, besteht aus den auf dem Minimum–Volume–Ellipsoid beruhenden Lokations- und Kovarianzschätzern. Die Definition dieses Ellipsoids stammt von Rousseeuw (1985) und wurde von Davies (1987) sowie von Lopuhaä und Rousseeuw (1991) leicht modifiziert, um die oben genannten Bruchpunkte der resultierenden Schätzer zu erreichen.

Definition 3.6 (Rousseeuw (1985), S. 289, Lopuhaä, Rousseeuw (1991), S. 235)

Sei $\underset{\sim}{x}_N$ eine Stichprobe vom Umfang N. Sei $h := \left[\frac{N+p+1}{2}\right]$.
Jedes Ellipsoid minimalen Volumens unter allen Ellipsoiden, die mindestens h Punkte von $\underset{\sim}{x}_N$ enthalten, heißt *Minimum–Volume–Ellipsoid (MVE)*.
Der auf dem MVE beruhende Lokationsschätzer ist definiert als das Zentrum des MVE, *der zugehörige Kovarianzschätzer* als die Stichprobenkovarianzmatrix der im MVE liegenden Beobachtungen von $\underset{\sim}{x}_N$, multipliziert mit einem Korrekturfaktor, um für den Fall von Beobachtungen multivariat normalverteilter Zufallsvariablen Fisher–Konsistenz zu erreichen.

Laut Davies (1987) können die MVE–Schätzer als Spezialfall der S–Schätzer betrachtet werden. Diese Aussage wird allerdings von Christmann et al. (1994) widerlegt, so daß hier MVE–Schätzer und S–Schätzer einzeln aufgeführt werden.
Die folgende Aussage folgt unmittelbar aus Korollar 3.1.

Korollar 3.4

Sei $\underline{\mathrm{OR}}_{\mathrm{S_{MB}}}$ ein Ausreißer–Identifizierer, der auf $\mathrm{S_{MB}}$–Schätzern, $\underline{\mathrm{OR}}_{\mathrm{MVE}}$ ein Identifizierer, der auf den MVE–Schätzern beruht. Sei $\underline{\mathrm{OR}} \in \{\underline{\mathrm{OR}}_{\mathrm{S_{MB}}}, \underline{\mathrm{OR}}_{\mathrm{MVE}}\}$. Sei $\underset{\sim}{x}_N = (\underset{\sim}{x}_n^r, \underset{\sim}{x}_k^0)$ eine Stichprobe von n regulären und k nichtregulären Beobachtungen, wobei die regulären Beobachtungen $\underset{\sim}{x}_n^r = (\underline{x}_1^r, \ldots, \underline{x}_n^r)$ in allgemeiner Lage seien. Sei außerdem $n \geq p+1$.
Dann gilt für den Masking–Bruchpunkt eines derartigen Identifizierers $\underline{\mathrm{OR}}$:

$$\varepsilon^M(\underline{\mathrm{OR}}) \quad = \quad \frac{\left[\frac{N-p+1}{2}\right]}{N}.$$

Da der Masking–Bruchpunkt der oben genannten Identifizierer bekannt ist, kann er zur Abschätzung des Swamping–Bruchpunkts herangezogen werden. Dazu wird ein Satz über den Zusammenhang von Masking- und Swamping–Bruchpunkt ausgenutzt.

Satz 3.8

Sei $\underline{\mathrm{OR}} \in \{\underline{\mathrm{OR}}_{\mathrm{S_{MB}}}, \underline{\mathrm{OR}}_{\mathrm{MVE}}\}$. Unter den gleichen Voraussetzungen wie in Korollar 3.4 gilt für den Swamping–Bruchpunkt des Identifizierers $\underline{\mathrm{OR}}$:

$$\varepsilon^S(\underline{\mathrm{OR}}) \quad \leq \quad \frac{\left[\frac{N+p}{2}\right]}{N-1}.$$

Beweis

Betrachte den Masking–Bruchpunkt:

$$\begin{aligned}\frac{\left[\frac{N-p+1}{2}\right]}{N} \quad &= \quad \frac{\left[\frac{N-p+1}{2}\right]}{N-\left[\frac{N-p+1}{2}\right]+\left[\frac{N-p+1}{2}\right]} \quad = \quad \frac{\left[\frac{N-p+1}{2}\right]-1+1}{m+\left[\frac{N-p+1}{2}\right]-1+1} \\ &= \quad \frac{n+1}{m+n+1}\end{aligned}$$

mit $m := N - \left[\frac{N-p+1}{2}\right]$.

Damit ist ein Satz über die Abschätzung des Swamping–Bruchpunkts anwendbar. Mit

$$m^* := \begin{cases} \inf\{m \,:\, \varepsilon^M(\underline{\mathrm{OR}}, \boldsymbol{\beta}, \underset{\sim}{x}^r_m, \delta) \;\geq\; \frac{n+1}{n+1+m} \\ \qquad \forall\, \underset{\sim}{x}^r_m,\ \forall\, \boldsymbol{\beta} = (\beta_j)_{j\in I\!N} \text{ mit } \beta_j = \beta\ \forall\, j \in I\!N\} \\ \\ \infty, \quad \text{falls kein solches } m \text{ existiert} \end{cases}$$

gilt nämlich für affin äquivariante Identifizierer:

$$\varepsilon^S(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}^r_n, \delta) \leq \frac{m^*}{n+m^*}$$

(vgl. Becker (1992), S. 34).

Hier ist $m^* = N - \left[\frac{N-p+1}{2}\right]$, und es gilt:

$$\begin{aligned}\varepsilon^S(\underline{\mathrm{OR}}) \quad &\leq \quad \frac{m^*}{n+m^*} \quad = \quad \frac{N-\left[\frac{N-p+1}{2}\right]}{\left[\frac{N-p+1}{2}\right]-1+N-\left[\frac{N-p+1}{2}\right]} \\ &= \quad \frac{N-\left[\frac{N-p+1}{2}\right]}{N-1}.\end{aligned}$$

Betrachte den Zähler:

$$\begin{aligned} & N - \left[\frac{N-p+1}{2}\right] \\ = & \left\{ \begin{array}{ll} N - \frac{N-p+1}{2} = \frac{N+p-1}{2}, & \text{falls } N-p+1 \text{ gerade} \\ N - \frac{N-p}{2} = \frac{N+p}{2}, & \text{falls } N-p+1 \text{ ungerade} \end{array} \right\}. \end{aligned}$$

Also ist $N - \left[\frac{N-p+1}{2}\right] = \left[\frac{N+p}{2}\right]$, und es folgt für den Swamping–Bruchpunkt:

$$\varepsilon^S(\underline{\mathrm{OR}}) \quad \leq \quad \frac{\left[\frac{N+p}{2}\right]}{N-1}.$$

$\square$

In bezug auf das Kriterium möglichst hoher Masking- und Swamping–Bruchpunkte lassen sich also Ausreißer–Identifizierer empfehlen, die auf $\mathrm{S_{MB}}$–Schätzern oder den MVE–Schätzern beruhen.

Im folgenden Kapitel wird ein weiteres Gütekriterium für multivariate Identifizierer vorgestellt.

4 Biasbetrachtungen

Bisher wurden für Ausreißer–Identifizierer Bruchpunkte (Masking- und Swamping–Bruchpunkt) als Gütekriterien betrachtet. Mit der Angabe eines Bruchpunkts wird das Verhalten eines Identifizierers natürlich nicht vollständig beschrieben. Bei der Betrachtung von Bruchpunkten versucht man herauszufinden, wie groß der Anteil von Ausreißern in einer Stichprobe sein muß, um den Zusammenbruch eines Identifizierers zu erreichen. Bei der Konstruktion von Verfahren mit hohem Bruchpunkt verfolgt man das Ziel, sich gegen ein Versagen bei großen Anzahlen von Ausreißern abzusichern. Andererseits werden nicht in jeder Stichprobe tatsächlich große Anzahlen von Ausreißern vorhanden sein. Es interessieren daher auch Aussagen über das Verhalten von Identifizierern bei Stichproben mit einem bestimmten, eventuell kleineren Anteil an Ausreißern.

Weiß man zum Beispiel aus Erfahrung, daß bei einer bestimmten Art von Daten in der Regel höchstens ein Anteil von etwa 10% der Beobachtungen als Ausreißer einzustufen ist, so möchte man eine Aussage darüber bekommen, wie sich ein Anteil von 10% an Ausreißern auf einen Identifizierer auswirkt. Im folgenden wird ein Gütekriterium entwickelt, anhand dessen ein Identifizierer in diesem Sinn beurteilt werden kann.

4.1 Der maximale asymptotische Bias

Das Ziel einer Schätzung besteht darin, die zu schätzende Größe möglichst genau zu „treffen“. Daher eignet sich die Bestimmung des Bias als Gütemaß. Die Definition des Bias eines Ausreißer–Identifizierers ist nicht ganz unproblematisch. Da es sich bei den hier betrachteten Identifizierern ebenso wie bei den Ausreißerbereichen um Komplemente von Ellipsoiden handelt, muß überlegt werden, wie ein Unterschied zwischen den Bereichen außerhalb zweier Ellipsoide bestimmt werden kann. Gleichwertig dazu kann der Unterschied zwischen den Ellipsoiden selbst festgelegt werden.

Da die hier betrachteten Ellipsoide unmittelbar auf Lokations- und Kovarianzschätzern basieren, kann man als Ausgangspunkt für die Biasdefinition ein entsprechendes Konzept für solche Schätzer heranziehen. Für Paare $(\underline{m}, S)$ von Lokations- und Kovarianzschätzern schlägt Tyler (1994) eine Definition des maximalen Bias vor, der durch eine „Verunreinigung“ von k Elementen einer Stichprobe vom Umfang N verursacht wird.

Definition 4.1 (Tyler (1994), S. 1027)

Sei $\underline{y}_N$ eine Stichprobe vom Umfang N und $\underline{y}_{N,k}$ durch Austausch von k Beobachtungen aus $\underline{y}_N$ durch beliebige Punkte entstanden. Für ein Paar $(\underline{m}, S)$ von Lokations- und Kovarianzschätzern ist der *maximale Bias*, der durch Kontamination eines Anteils von k/N der Stichprobenelemente entstehen kann, definiert als

$$\begin{aligned} b(\frac{k}{N}, \underline{y}_N; \underline{m}, S) &= \sup_{\underline{y}_{N,k}} [\max\{\|S(\underline{y}_N)^{-1/2}(\underline{m}(\underline{y}_N) - \underline{m}(\underline{y}_{N,k}))\|, \\ &\quad \operatorname{tr}(S(\underline{y}_N)S^{-1}(\underline{y}_{N,k}) + S^{-1}(\underline{y}_N)S(\underline{y}_{N,k}))\}]. \end{aligned}$$

Diese Definition stützt sich im wesentlichen auf den euklidischen Abstand der Lokationsschätzer und die Eigenwerte der Kovarianzschätzer.
Dabei handelt es sich um eine Definition des Bias an einer Stichprobe, d. h., Tyler bestimmt den Unterschied der Schätzer, wenn sie zum einen aus einer Stichprobe $\underline{y}_N$, zum anderen aus der durch den Austausch von k Beobachtungen entstandenen Stichprobe $\underline{y}_{N,k}$ berechnet werden. Dadurch wird nur derjenige Teil des gesamten Bias bestimmt, der durch den Austausch der k Beobachtungen durch beliebige Punkte entsteht. Um dies zu verdeutlichen, betrachte man als Beispiel die Differenz $\underline{m}(\underline{y}_N) - \underline{m}(\underline{y}_{N,k})$. Mit $\underline{m}(\underline{y}_N)$ wird der Lageparameter $\underline{\mu}$ geschätzt, ebenso mit $\underline{m}(\underline{y}_{N,k})$. Da an den Schätzer $\underline{m}$ keine weiteren Voraussetzungen gestellt werden, gilt für die Erwartungswerte: $E(\underline{m}(\underline{y}_N)) = \underline{\mu} + \text{Biasterm}_1$, $E(\underline{m}(\underline{y}_{N,k})) = \underline{\mu} + \text{Biasterm}_1 + \text{Biasterm}_2$, wobei der Biasterm_2 gerade zusätzlich zum Biasterm_1 durch den Austausch der k Beobachtungen zustande kommt. Betrachtet man die Differenz der Schätzer, so ist $E(\underline{m}(\underline{y}_N) - \underline{m}(\underline{y}_{N,k})) = \text{Biasterm}_2$, also gerade der „Austausch–Bias“. Für den Vergleich zweier Schätzer bezüglich ihres Bias ist aber nicht nur dieser „Austausch–Bias“ wichtig, sondern auch der im obigen Biasterm_1 auftretende sonstige Bias der Schätzer. Aus diesem Grund werden hier als Bezugsgrößen für die Schätzer an der „verunreinigten“ Stichprobe die zu schätzenden Parameter der zugrundeliegenden Verteilung gewählt. Auf diese Weise erhält man eine etwas allgemeinere Definition des Bias.

Eine weitere Modifikation des Ansatzes von Tyler ist notwendig, wenn man bedenkt, daß bei der Definition des maximalen Bias eines Ausreißer–Identifizierers beachtet werden muß, daß die Verzerrung durch die Ausreißer in der Stichprobe verursacht wird und nicht durch eine beliebige Modellabweichung. Eine Lösung dieses Problems findet man bei Davies und Gather (1993, S. 787), die eine Definition von Huber (1981, S. 12) auf die Situation von Datensätzen mit einem gewissen Anteil an Ausreißern zuschneiden.
Mit den bisherigen Überlegungen wird die Tyler'sche Biasdefinition für ein Paar $(\underline{m}, S)$ von Lokations- und Kovarianzschätzern also in folgender Weise geändert: $\underline{m}(\underset{\sim}{y}_N)$ und $S(\underset{\sim}{y}_N)$ werden ersetzt durch die zu schätzenden Größen $\underline{\mu}$ und Σ, und $b(\frac{k}{N}, \underset{\sim}{y}_N; \underline{m}, S)$ wird zu

$$b(\frac{k}{N}, \underset{\sim}{x}_N; \underline{m}, S) = \sup_{\underset{\sim}{x}_k^0}[\max\{\|\Sigma^{-1/2}(\underline{\mu} - \underline{m}(\underset{\sim}{x}_N))\|,\ \mathrm{tr}(\Sigma S^{-1}(\underset{\sim}{x}_N) + \Sigma^{-1}S(\underset{\sim}{x}_N))\}].$$

Dabei wird mit $\underset{\sim}{x}_N = (\underset{\sim}{x}_n^r, \underset{\sim}{x}_k^0)$ wie bisher eine Stichprobe mit n regulären Beobachtungen und einer Anzahl k von δ_N–Ausreißern bezeichnet. Das Supremum wird damit gebildet über alle möglichen Kombinationen $\underset{\sim}{x}_k^0$ von δ_N–Ausreißern, die der regulären Stichprobe hinzugefügt werden können.
Im weiteren soll für die Biasbetrachtung eines Identifizierers nur der Fall betrachtet werden, daß die Prozedur aufgrund von „Explosion“ zusammenbricht. Dadurch ist es möglich, einen Teil aus Tylers Definition aufzugeben. In der Biasdefinition für den Kovarianzschätzer bestimmt Tyler die Spur der Matrizen $S(\underset{\sim}{y}_N)S^{-1}(\underset{\sim}{y}_{N,k})$ und $S^{-1}(\underset{\sim}{y}_N)S(\underset{\sim}{y}_{N,k})$. Bricht der Schätzer S aufgrund von „Explosion“ zusammen, so wird $\mathrm{tr}(S^{-1}(\underset{\sim}{y}_N)S(\underset{\sim}{y}_{N,k}))$ beliebig groß, während für den Fall des Zusammenbruchs aufgrund von „Implosion“ die Spur von $S(\underset{\sim}{y}_N)S^{-1}(\underset{\sim}{y}_{N,k})$ betroffen ist. Läßt man diesen letzteren Fall außer Betracht, so kann man auf $\mathrm{tr}(S(\underset{\sim}{y}_N)S^{-1}(\underset{\sim}{y}_{N,k}))$ in Tylers Definition bzw. auf $\mathrm{tr}(\Sigma S^{-1}(\underset{\sim}{x}_N))$ in der modifizierten Definition verzichten. Damit wird die Biasdefinition für ein Paar $(\underline{m}, S)$ weiter verändert zu

$$b(\frac{k}{N}, \underset{\sim}{x}_N; \underline{m}, S) = \sup_{\underset{\sim}{x}_k^0}[\max\{\|\Sigma^{-1/2}(\underline{\mu} - \underline{m}(\underset{\sim}{x}_N))\|,\ \mathrm{tr}(\Sigma^{-1}S(\underset{\sim}{x}_N))\}].$$

Es handelt sich aber nach wie vor um eine Begriffsbildung für den Bias eines Paares $(\underline{m}, S)$ von Schätzern. Dies trifft die Situation für einen Identifizierer noch nicht angemessen. Stellt man die zu vergleichenden Ellipsoide $\mathbb{R}^p \backslash \text{out}$ und $\mathbb{R}^p \backslash \underline{\text{OR}}$ auf dieselbe Art dar, nämlich durch

$$\mathbb{R}^p \backslash \text{out} = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{\mu})^T (\chi^2_{p;1-\alpha_N} \Sigma)^{-1} (\underline{x} - \underline{\mu}) \leq 1\}$$

und

$$\mathbb{R}^p \backslash \underline{\text{OR}} = \{\underline{x} \in \mathbb{R}^p : (\underline{x} - \underline{m})^T (c(p, N, \alpha_N) S)^{-1} (\underline{x} - \underline{m}) < 1\},$$

so sieht man, daß anstelle eines Vergleichs von S mit der Kovarianzmatrix Σ tatsächlich ein Vergleich zwischen den Matrizen $c(p, N, \alpha_N)S$ und $\chi^2_{p;1-\alpha_N}\Sigma$ angestellt werden muß. Damit wird der Ansatz für die Biasdefinition erneut geändert, und zwar zu

$$b(\frac{k}{N}, \underset{\sim}{x}_N; \underline{m}, S) = \sup_{\underset{\sim}{x}^0_k} [\max\{\|\Sigma^{-1/2}(\underline{\mu} - \underline{m}(\underset{\sim}{x}_N))\|, \text{ tr}\left(\frac{c(p, N, \alpha_N)}{\chi^2_{p;1-\alpha_N}} \Sigma^{-1} S(\underset{\sim}{x}_N)\right)\}].$$

So, wie die Biasdefinition bis jetzt entwickelt wurde, ist insbesondere die von Tyler für seine Definition angestrebte Eigenschaft der Invarianz erhalten geblieben. Das heißt, daß sich der Wert des Bias unter einer linearen Transformation der Daten nicht ändert, wenn affin äquivariante Schätzer benutzt werden (Tyler (1994), S. 1027). Diese Eigenschaft ist zunächst verwirrend, da sie nicht mit der gängigen Vorstellung des Bias eines Schätzers übereinstimmt. Schränkt man sich aber auf affin äquivariante Schätzer ein, so wird eine solche Invarianzeigenschaft, also gewissermaßen eine Normierung des Bias, durchaus einsehbar. Die affine Äquivarianz erzwingt, daß ein Schätzer eine lineare Datentransformation in geeigneter Weise mitvollzieht. Ändert sich dann aber unter der entsprechenden Transformation der Wert des Bias, so wird der Schätzer für seine Äquivarianzeigenschaft gewissermaßen noch „bestraft“. Um dies auszuschließen, ist der Ansatz eines transformationsinvarianten Bias angemessen. Da in dieser Arbeit nur affin äquivariante Ausreißer–Identifizierer betrachtet werden, ist die Beibehaltung des Invarianzkonzepts für den Bias einer Identifizierungsprozedur sinnvoll.

Betrachtet man die bis jetzt motivierte Modifikation der Tyler'schen Biasdefinition, so läßt sich im ersten der beiden Ausdrücke, aus denen das Maximum gebildet wird, bereits eine Interpretation bezüglich der Ellipsoidstruktur des Identifizierers geben. Der Ausdruck $\|\Sigma^{-1/2}(\underline{\mu} - \underline{m}(\underset{\sim}{x}_N))\|$ beschreibt gerade den (normierten) Abstand des Mittelpunkts $\underline{m}$ des Schätzellipsoids $\mathbb{R}^p \backslash \underline{\text{OR}}$ zum Mittelpunkt des zu schätzenden Ellipsoids $\mathbb{R}^p \backslash \text{out}$.

Dagegen liefert der Ausdruck $\operatorname{tr}\left(\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}\Sigma^{-1}S(\underset{\sim}{x}_N)\right)$ in diesem Zusammenhang keine unmittelbare Interpretation, die auch die Ellipsoidstruktur des Ausreißer–Identifizierers berücksichtigt. Dazu müssen weitere Informationen in die Definition einbezogen werden.

In einer Arbeit von Harter (1964) findet sich ein Ansatz für den Vergleich von Intervallschätzern für einen eindimensionalen Parameter. Harter benutzt die Abweichungen der Intervallendpunkte vom zu schätzenden Parameter als Vergleichskriterium. Zwar liegt hier eine andere Situation zugrunde, da das zum Identifizierer $\underline{\mathrm{OR}}$ gehörende Ellipsoid kein Bereichsschätzer für einen Punkt, sondern für einen Bereich ist, jedoch kann man in Anlehnung an Harter die Abstände zwischen den Achsenendpunkten des zu schätzenden Ellipsoids $\mathbb{R}^p\backslash\mathrm{out}$ und des Schätzellipsoids $\mathbb{R}^p\backslash\underline{\mathrm{OR}}$ als Gütekriterium heranziehen. Läßt man denjenigen Teil des Unterschieds zwischen den beiden Ellipsoiden, der auf die Abweichung der Mittelpunkte zurückzuführen ist, außer Betracht und setzt voraus, daß die Mittelpunkte von $\mathbb{R}^p\backslash\underline{\mathrm{OR}}(\underset{\sim}{x}_N,\alpha_N)$ und $\mathbb{R}^p\backslash\mathrm{out}(\alpha_N,\underline{\mu},\Sigma)$ übereinstimmen und ohne Beschränkung der Allgemeinheit gleich Null sind, so betrachtet man als Unterschied zwischen den Ellipsoiden normierte Abstände ihrer Achsenendpunkte. Abbildung 4.1 verdeutlicht die Idee für den Fall zweidimensionaler Beobachtungen. Für

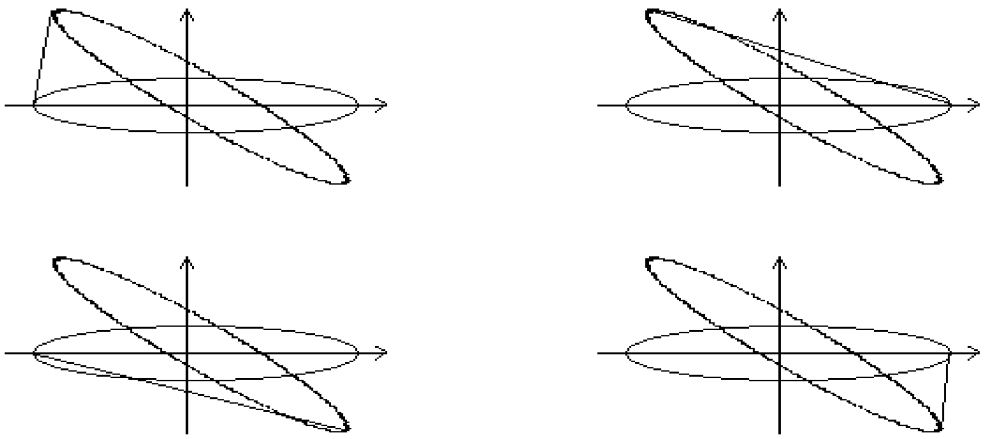


Abbildung 4.1: Abstände der Achsenendpunkte der längsten Hauptachsen zweier Ellipsen

jede Hauptachse existieren zwei Endpunkte, nämlich für $\mathbb{R}^p \backslash \text{out}(\alpha_N, \underline{\mu}, \Sigma)$ die Punkte $\tilde{\lambda}_i \underline{u}_i$ und $-\tilde{\lambda}_i \underline{u}_i$ mit $\tilde{\lambda}_i = \sqrt{\chi^2_{p;1-\alpha_N} \lambda_i}$, $i = 1, \ldots, p$, und für $\mathbb{R}^p \backslash \underline{\text{OR}}(\underset{\sim}{x}_N, \alpha_N)$ die Punkte $\tilde{\xi}_i \underline{v}_i$ und $-\tilde{\xi}_i \underline{v}_i$ mit $\tilde{\xi}_i = \sqrt{c(p, N, \alpha_N)\xi_i}$, $i = 1, \ldots, p$. Dabei bezeichnen λ_i bzw. ξ_i die Eigenwerte von Σ bzw. S mit jeweils zugehörigen normierten Eigenvektoren $\underline{u}_i$ bzw. $\underline{v}_i$, $i = 1, \ldots, p$. Damit ergeben sich die vier möglichen Abstände $\|\tilde{\lambda}_i \underline{u}_i - \tilde{\xi}_i \underline{v}_i\|$, $\|\tilde{\lambda}_i \underline{u}_i + \tilde{\xi}_i \underline{v}_i\|$, $\|-\tilde{\lambda}_i \underline{u}_i - \tilde{\xi}_i \underline{v}_i\|$ und $\|-\tilde{\lambda}_i \underline{u}_i + \tilde{\xi}_i \underline{v}_i\|$, $i = 1, \ldots, p$. Als Maß für den Abstand der Achsen voneinander bietet sich die mittlere Summe der Abstände der Endpunkte an, also

$$
\begin{aligned}
& \frac{1}{4}(\|\tilde{\lambda}_i \underline{u}_i - \tilde{\xi}_i \underline{v}_i\| + \|\tilde{\lambda}_i \underline{u}_i + \tilde{\xi}_i \underline{v}_i\| + \|-\tilde{\lambda}_i \underline{u}_i - \tilde{\xi}_i \underline{v}_i\| + \|-\tilde{\lambda}_i \underline{u}_i + \tilde{\xi}_i \underline{v}_i\|) \\
= \; & \frac{1}{4}(\|\tilde{\lambda}_i \underline{u}_i - \tilde{\xi}_i \underline{v}_i\| + \|\tilde{\lambda}_i \underline{u}_i + \tilde{\xi}_i \underline{v}_i\| + \|\tilde{\lambda}_i \underline{u}_i + \tilde{\xi}_i \underline{v}_i\| + \|\tilde{\lambda}_i \underline{u}_i - \tilde{\xi}_i \underline{v}_i\|) \\
= \; & \frac{1}{2}(\|\tilde{\lambda}_i \underline{u}_i - \tilde{\xi}_i \underline{v}_i\| + \|\tilde{\lambda}_i \underline{u}_i + \tilde{\xi}_i \underline{v}_i\|), \quad i = 1, \ldots, p.
\end{aligned}
$$

Dies ist als Maß für den durch die Abweichungen der Hauptachsen begründeten Unterschied der Ellipsoide noch nicht zufriedenstellend. Betrachtet man nämlich den Fall, daß bei differierenden Achsenrichtungen die Hauptachsenlängen der i–ten Achsen gleich sind, ergibt sich als Wert für den Unterschied der i–ten Achsen der Ausdruck $\frac{1}{2}\tilde{\lambda}_i(\|\underline{u}_i - \underline{v}_i\| + \|\underline{u}_i + \underline{v}_i\|)$. Bei zwei gleichlangen Achsen, die in einem festen Winkel zueinander stehen, erhielte man also für verschiedene Längen auch verschiedene Werte für den Unterschied. Um diesem Phänomen zu begegnen, wird der oben betrachtete mittlere Abstand durch Multiplikation mit $\frac{1}{\tilde{\lambda}_i}$ normiert. Hier spiegelt sich die Idee der Invarianzeigenschaft aus Tylers Definition. Auf diese Weise erhält man als den Teil des Bias, der auf die unterschiedliche Achsenlänge und -lage der beiden Ellipsoide $\mathbb{R}^p \backslash \underline{\text{OR}}$ und $\mathbb{R}^p \backslash \text{out}$ zurückgeht, die Größe

$$
\frac{1}{2} \sum_{i=1}^{p} \left(\|\underline{u}_i - \sqrt{\frac{c(p, N, \alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N} \lambda_i}} \, \underline{v}_i\| + \|\underline{u}_i + \sqrt{\frac{c(p, N, \alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N} \lambda_i}} \, \underline{v}_i\| \right). \tag{4.1}
$$

Bei vollständiger Übereinstimmung der beiden Ellipsoide berechnet sich diese Größe gerade zu p, so daß der Wert p abgezogen werden muß, um bei korrekter Schätzung des Ausreißerbereichs einen Biaswert von Null zu erhalten.

Aus der Biasdefinition von Tyler ergibt sich durch die zuvor durchgeführten Modifikationen die Größe

$$\mathrm{tr}\left(\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}\Sigma^{-1}S\right)$$

als Maß für den Unterschied zwischen dem Kovarianzschätzer S und der zu schätzenden Matrix Σ, wobei gleichzeitig die Normierungskonstanten des Identifizierers $\underline{\mathrm{OR}}$ und des zu schätzenden Bereichs out berücksichtigt werden. Aufgrund einer Abschätzung von Theobald (1975, S. 462), die besagt, daß

$$\mathrm{tr}\left(\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}\Sigma^{-1}S\right) \geq \sum_{i=1}^{p}\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}\frac{\xi_i}{\lambda_i}$$

(mit λ_i, ξ_i Eigenwerte von Σ bzw. S), ergeben sich daraus die Größen $\dfrac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}\dfrac{\xi_i}{\lambda_i}$, $i = 1,\dots,p$, als charakterisierend für den Bias von $c(p,N,\alpha_N)S$.

Die auf diese Weise aus der Definition von Tyler abgeleiteten Größen beschreiben einerseits den Bias für den Kovarianzschätzer, der in den Identifizierer $\underline{\mathrm{OR}}$ eingeht. Gleichzeitig stimmen die Terme $\sqrt{c(p,N,\alpha_N)\xi_i}$ und $\sqrt{\chi^2_{p;1-\alpha_N}\lambda_i}$ mit den Achsenlängen der Hauptachsen von $\mathbb{R}^p \backslash \underline{\mathrm{OR}}$ bzw. $\mathbb{R}^p \backslash \mathrm{out}$ überein, so daß diese Größen im Kontext der Ausreißer–Identifizierer auch eine geometrische Interpretation besitzen. Die in (4.1) darüber hinaus eingehenden Eigenvektoren der Matrizen S und Σ charakterisieren als Ergänzung dazu diejenige Komponente der Abweichung zwischen $\mathbb{R}^p \backslash \underline{\mathrm{OR}}$ und $\mathbb{R}^p \backslash \mathrm{out}$, die durch die Unterschiede in den Hauptachsenrichtungen der beiden Ellipsoide zustandekommt.

Nach den bisherigen Überlegungen sollten also in die Definition des Bias eines Ausreißer-Identifizierers die beiden folgenden Größen aufgenommen werden:

$$\|\Sigma^{-1/2}(\underline{\mu} - \underline{m}(\underset{\sim}{x}_N))\|$$

als Maß für die Abweichung der Ellipsoidmittelpunkte und

$$\frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i - \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\| + \|\underline{u}_i + \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right) - p$$

als Maß für die Abweichung der Achsenendpunkte der Hauptachsen.

Folgt man an dieser Stelle weiterhin der Idee von Tyler (1994), so könnte man aus diesen beiden Größen das Maximum bilden als Maß für die Gesamtabweichung der beiden Ellipsoide $I\!R^p\backslash\underline{\text{OR}}$ und $I\!R^p\backslash\text{out}$. Dies würde aber dazu führen, daß beispielsweise zwei Identifizierer, die auf dem gleichen Kovarianzschätzer, aber unterschiedlichen Schätzern für die Lage beruhen, denselben Bias besitzen, falls nur in beiden Fällen der Lokationsschätzer besser ist als der Kovarianzschätzer. Ob gleichzeitig einer der Lokationsschätzer dem anderen überlegen ist, würde dabei nicht berücksichtigt. Um auch in solchen Fällen zu unterschiedlichen Biaswerten zu gelangen, wird hier statt dessen ein additiver Ansatz gewählt, das heißt, die beiden oben erhaltenen Größen werden zur Bestimmung der Gesamtabweichung von $I\!R^p\backslash\underline{\text{OR}}$ und $I\!R^p\backslash\text{out}$ (bzw. äquivalent dazu von $\underline{\text{OR}}$ und out) addiert.

Die vorangegangenen Überlegungen motivieren die folgende Definition.

Definition 4.2

Sei $(\underline{X}_i^r)_{i\in I\!N}$ eine Folge von unabhängigen, identisch nach $N(\underline{\mu},\Sigma)$ verteilten Zufallsvektoren mit zugehörigen Beobachtungen $(\underline{x}_i^r)_{i\in I\!N}$, $\underline{x}_i^r \in I\!R^p$, sei $\underset{\sim}{x}_n^r := (\underline{x}_1^r,\ldots,\underline{x}_n^r)$.
Sei weiter $k := [\eta n]$, $0<\eta<1$, $N := n+k$ und $\boldsymbol{\delta} := (\delta_i)_{i\in I\!N}$ mit $\delta_i \in]0,1[$.
Sei $\underset{\sim}{x}_k^0 := (\underline{x}_1^0,\ldots,\underline{x}_k^0)$ ein Tupel aus δ_N–Ausreißern und $\underset{\sim}{x}_N := (\underset{\sim}{x}_n^r, \underset{\sim}{x}_k^0)$.
Sei $\underline{\text{OR}} = \underline{\text{OR}}(\underset{\sim}{x}_N, \alpha_N)$ ein Ausreißer–Identifizierer gemäß Definition 2.2, basierend auf den Schätzern $\underline{m}$ und S für $\underline{\mu}$ bzw. Σ sowie einer Normierungskonstante $c(p,N,\alpha_N)$.
Seien weiter λ_i bzw. ξ_i die Eigenwerte von Σ bzw. S mit jeweils zugehörigen normierten Eigenvektoren $\underline{u}_i$ bzw. $\underline{v}_i$, $i=1,\ldots,p$.
Der *maximale asymptotische Bias des Identifizierers* $\underline{\text{OR}}$ ist definiert durch

$$\begin{aligned} B(\underline{\text{OR}},\eta,\boldsymbol{\delta}) \;&:=\; \limsup_{N\to\infty}\Big(\sup_{\underset{\sim}{x}_k^0\in\,\text{out}(\delta_N,\underline{\mu},\Sigma)} \big(\|\Sigma^{-1/2}(\underline{\mu}-\underline{m})\| \\ &\quad+\frac{1}{2}\sum_{i=1}^{p}\left(\Big\|\underline{u}_i-\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\Big\|+\Big\|\underline{u}_i+\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\Big\|\right)-p\big)\Big). \end{aligned}$$

Dabei bezeichnet $\|.\|$ die euklidische Norm des $I\!R^p$.
Die Schreibweise $\underset{\sim}{x}_k^0 \in \text{out}(\delta_N,\underline{\mu},\Sigma)$ steht kurz für $\underline{x}_i^0 \in \text{out}(\delta_N,\underline{\mu},\Sigma)$, $i=1,\ldots,k$.

Bei der Betrachtung der Bruchpunkte hat sich gezeigt, daß das Verhalten eines Identifizieres zwingend vom Verhalten der in ihm verwendeten Schätzer bestimmt wird.

Zur Untersuchung entsprechender Zusammenhänge für den maximalen asymptotischen Bias müssen zunächst Biasdefinitionen für die Schätzer gefunden werden. Dabei werden hier die Biasterme für Lokations- und Kovarianzschätzer einzeln definiert. Damit liegen ähnlich wie bei der Begriffsbildung für die Finite–sample Bruchpunkte von Schätzern allgemeinere Definitionen auch für solche Fälle vor, in denen Schätzer für Lage und Kovarianz nicht simultan betrachtet werden.

Definition 4.3

Gegeben seien die Voraussetzungen von Definition 4.2.
Für einen Lokationsschätzer $\underline{m} = \underline{m}(\underset{\sim}{x}_N)$ für den Erwartungswert $\underline{\mu}$ ist *der maximale asymptotische Bias* definiert durch

$$b(\underline{m}, \eta, \boldsymbol{\delta}) := \limsup_{N \to \infty} \Big(\sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N, \underline{\mu}, \Sigma)} \| \underline{m}(\underset{\sim}{x}_N) - \underline{\mu} \| \Big).$$

Für die entsprechende Definition des maximalen asymptotischen Bias eines Kovarianzschätzers benötigt man eine geeignete Matrixnorm. Hier bietet sich die Spektralnorm an.

Definition 4.4 (Zurmühl, Falk (1986), S. 36)

Sei $A \in \mathbb{R}^{p \times p}$ eine beliebige Matrix. Sei ξ_1 der größte Eigenwert der Matrix $A^T A$. Dann ist *die Spektralnorm von A* definiert durch

$$\|A\|_2 \ := \ \sqrt{\xi_1}.$$

Falls die Matrix A symmetrisch ist, läßt sich die Spektralnorm auch darstellen als $\|A\|_2 = |\zeta|$, wobei ζ der betragsmäßig größte Eigenwert von A ist. In diesem Fall ist also die Spektralnorm von A gleich dem Spektralradius von A.

Nun läßt sich analog zu Definition 4.3 der maximale asymptotische Bias eines Kovarianzschätzers definieren.

Definition 4.5

Mit den Voraussetzungen und Bezeichnungen aus Definition 4.2 ist *der maximale asymptotische Bias* eines Schätzers $S = S(\underset{\sim}{x}_N)$ für die Kovarianzmatrix Σ definiert durch

$$b(S, \eta, \boldsymbol{\delta}) := \limsup_{N\to\infty} \Big(\sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N, \underline{\mu}, \Sigma)} \|S(\underset{\sim}{x}_N) - \Sigma\|_2 \Big).$$

Dabei bezeichnet $\|.\|_2$ die Spektralnorm des $I\!R^{p\times p}$.

Die Wahl des euklidischen Abstands als Maß für die Abweichung von $\underline{m}$ zu $\underline{\mu}$ in Definition 4.3 ergibt sich in natürlicher Weise aus dem Fall univariater Beobachtungen. Dort wird die Abweichung durch den Betrag der Differenz $m - \mu$ gemessen.
Für die Bestimmung des Abstands von S zu Σ wird gerade die Spektralnorm herangezogen, da sie die Eigenschaft der Verträglichkeit mit der euklidischen Vektornorm erfüllt. Die Verträglichkeit bedeutet, daß $\|A\underline{x}\| \leq \|A\|_2 \|\underline{x}\| \quad \forall A \in I\!R^{p\times p}, \forall \underline{x} \in I\!R^p$ gilt (vgl. Zurmühl, Falk (1986), S. 36). Die Definition des Abstands zweier Matrizen über die Spektralnorm folgt einem Vorschlag von Woodruff und Rocke (1993, S. 70).

Beide Definitionen des maximalen asymptotischen Bias gewährleisten, daß ein besonders kleiner Wert von b für eine besonders gute Schätzung spricht. Ein Wert von Null für b ist dabei das Optimum.
Wie man sieht, wird hier im Gegensatz zur Biasdefinition für einen Ausreißer–Identifizierer weder für einen Lokationsschätzer noch für einen Kovarianzschätzer eine Invarianz des Bias unter linearen Transformationen angestrebt. Da die Biasterme für Lokations- und Kovarianzschätzer einzeln definiert werden, um eine größere Allgemeinheit der Definition zu gewährleisten, schränkt man sich auch nicht von vornherein auf die Eigenschaft der affinen Äquivarianz für die Schätzer ein. Gleichzeitig wird aber für solche Schätzer durch die hier erstellten Definitionen eine Art von Äquivarianz unter Skalentransformationen erreicht, wie sie für einen Bias im allgemeinen auch wünschenswert ist.
Für den Spezialfall affin äquivarianter Schätzer gilt nämlich für die hier benutzten Biasdefinitionen

$$b(\underline{m}(A\underset{\sim}{x}_N), \eta, \boldsymbol{\delta}) \leq \|A\|_2 b(\underline{m}(\underset{\sim}{x}_N), \eta, \boldsymbol{\delta})$$

und

$$b(S(A\underset{\sim}{x}_N),\eta,\boldsymbol{\delta}) \leq \|A\|_2^2 b(S(\underset{\sim}{x}_N),\eta,\boldsymbol{\delta}).$$

Eine Gleichheit statt der Ungleichungen, wie sie aus dem univariaten Fall bekannt ist, ist im multivariaten aufgrund der Eigenschaften von Vektor- und Matrixnormen nicht zu erreichen, so daß die hier angegebenen Zusammenhänge die beste Angleichung an die Skalenäquivarianz des Bias im univariaten Fall darstellen.

Mit den so gewählten Biasdefinitionen können nun Beziehungen zwischen den Größen $b(\underline{m},\eta,\boldsymbol{\delta})$, $b(S,\eta,\boldsymbol{\delta})$ und $B(\underline{\mathrm{OR}},\eta,\boldsymbol{\delta})$ hergestellt werden.

Satz 4.1

Gegeben seien die Voraussetzungen von Definition 4.2. Dann gilt der folgende Zusammenhang.
Falls $b(\underline{m},\eta,\boldsymbol{\delta}) = \infty$ ist, so ist auch $B(\underline{\mathrm{OR}},\eta,\boldsymbol{\delta}) = \infty$.

Beweis
Sei $b(\underline{m},\eta,\boldsymbol{\delta}) = \infty$. Es ist

$$\begin{aligned} B(\underline{\mathrm{OR}},\eta,\boldsymbol{\delta}) &= \limsup_{N\to\infty}\Big(\sup_{\underset{\sim}{x}_k^0 \in \mathrm{out}(\delta_N,\underline{\mu},\Sigma)} \Big(\|\Sigma^{-1/2}(\underline{\mu}-\underline{m})\| \\ &\quad + \frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i - \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\| + \|\underline{u}_i + \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right) - p\Big)\Big). \end{aligned}$$

Nun gilt

$$\begin{aligned} &\|\Sigma^{-1/2}(\underline{\mu}-\underline{m})\| \\ &+ \frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i - \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\| + \|\underline{u}_i + \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right) - p \\ \geq\ & \|\Sigma^{-1/2}(\underline{\mu}-\underline{m})\| - p \quad = \quad \frac{\|\Sigma^{1/2}\|_2}{\|\Sigma^{1/2}\|_2}\|\Sigma^{-1/2}(\underline{\mu}-\underline{m})\| - p \\ \geq\ & \frac{1}{\|\Sigma^{1/2}\|_2}\|\underline{\mu}-\underline{m}\| - p \quad = \quad \frac{1}{\sqrt{\lambda_1}}\|\underline{\mu}-\underline{m}\| - p. \end{aligned}$$

Diese Ungleichung bleibt für jedes $N \in \mathbb{N}$ erhalten, wenn auf beiden Seiten das Supremum über alle $\underset{\sim}{x}_k^0 \in \text{out}(\delta_N, \underline{\mu}, \Sigma)$ gebildet wird, und damit ist

$$\begin{aligned}
&\sup_{\underset{\sim}{x}_k^0 \in \text{out}(\delta_N,\underline{\mu},\Sigma)} (\|\Sigma^{-1/2}(\underline{\mu} - \underline{m})\| \\
&\quad + \frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i - \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\| + \|\underline{u}_i + \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right) - p) \\
\geq\ &\sup_{\underset{\sim}{x}_k^0 \in \text{out}(\delta_N,\underline{\mu},\Sigma)} (\frac{1}{\sqrt{\lambda_1}}\|\underline{\mu} - \underline{m}\| - p). \qquad (4.2)
\end{aligned}$$

Wegen der Voraussetzung, daß $b(\underline{m}, \eta, \boldsymbol{\delta}) = \infty$ ist, gilt

$$\limsup_{N\to\infty}(\sup_{\underset{\sim}{x}_k^0 \in \text{out}(\delta_N,\underline{\mu},\Sigma)} (\|\underline{\mu} - \underline{m}\|)) = \infty.$$

Daraus folgt: für alle $R \in \mathbb{R}$ existiert eine unendliche Indexmenge $I_R \subseteq \mathbb{N}$, $|I_R| = \infty$, so daß

$$\sup_{\underset{\sim}{x}_k^0 \in \text{out}(\delta_N,\underline{\mu},\Sigma)} (\|\underline{\mu} - \underline{m}\|) > R \quad \forall\, N \in I_R.$$

Daher ist

$$\sup_{\underset{\sim}{x}_k^0 \in \text{out}(\delta_{N_0},\underline{\mu},\Sigma)} (\frac{1}{\sqrt{\lambda_1}}\|\underline{\mu} - \underline{m}\| - p) > \frac{1}{\sqrt{\lambda_1}}R - p \quad \forall\, N \in I_R$$

und somit

$$\limsup_{N\to\infty} \sup_{\underset{\sim}{x}_k^0 \in \text{out}(\delta_{N_0},\underline{\mu},\Sigma)} (\frac{1}{\sqrt{\lambda_1}}\|\underline{\mu} - \underline{m}\| - p) = \infty.$$

Mit (4.2) folgt unmittelbar die Behauptung. □

Ein Identifizierer besitzt also einen beliebig großen maximalen asymptotischen Bias, wenn der in ihm verwendete Lokationsschätzer beliebig weit von seinem „Ziel“ abweicht. Die Umkehrung des Satzes führt zu der Folgerung, daß ein Identifizierer mit beschränktem Bias nur auf einem Lokationsschätzer mit beschränktem Bias beruhen kann. Die Beschränktheit von $b(\underline{m}, \eta, \boldsymbol{\delta})$ ist damit eine notwendige Voraussetzung für die Beschränktheit von $B(\underline{\text{OR}}, \eta, \boldsymbol{\delta})$.

Ein entsprechender Zusammenhang besteht auch zwischen dem maximalen asymptotischen Bias des Identifizierers und dem des verwendeten Kovarianzschätzers.

Satz 4.2

Unter den Voraussetzungen von Definition 4.2 gilt:

falls $b(S,\eta,\boldsymbol{\delta})=\infty$ ist, so ist auch $B(\underline{\mathrm{OR}},\eta,\boldsymbol{\delta})=\infty$.

Beweis

Der Beweis verläuft ähnlich wie der von Satz 4.1. Man nimmt an, daß $b(S,\eta,\boldsymbol{\delta})=\infty$ ist. Analog zum Beweis von Satz 4.1 schätzt man zunächst ab:

$$\begin{aligned}
&\|\Sigma^{-1/2}(\underline{\mu}-\underline{m})\| \\
&\quad+\frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i-\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|+\|\underline{u}_i+\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right)-p \\
\geq\ &\frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i-\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|+\|\underline{u}_i+\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right)-p \\
\geq\ &\frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i+\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i-\underline{u}_i+\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right)-p \\
=\ &\frac{1}{2}\sum_{i=1}^{p}\|2\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|-p \quad=\quad \sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}}\sum_{i=1}^{p}\sqrt{\frac{\xi_i}{\lambda_i}}-p
\end{aligned}$$

(wegen $\|\underline{a}-\underline{b}\|\leq\|\underline{a}\|+\|\underline{b}\|\ \forall\,\underline{a},\underline{b}\in\mathbb{R}^p$ und $\|\underline{v}_i\|=1, i=1,\ldots,p$)

$$\geq\ \sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}\lambda_1}}\sqrt{\xi_1}-p \quad=\quad \sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}\lambda_1}}\sqrt{\|S\|_2}-p$$

(wegen $\sqrt{\frac{\xi_1}{\lambda_1}}\leq\sum_{i=1}^{p}\sqrt{\frac{\xi_i}{\lambda_i}}$).

Diese Ungleichung bleibt für jedes $N\in\mathbb{N}$ erhalten, wenn auf beiden Seiten das Supremum über alle $\underset{\sim}{x}_k^0\in\operatorname{out}(\delta_N,\underline{\mu},\Sigma)$ gebildet wird, so daß

$$\begin{aligned}
&\sup_{\underset{\sim}{x}_k^0\in\operatorname{out}(\delta_N,\underline{\mu},\Sigma)}(\|\Sigma^{-1/2}(\underline{\mu}-\underline{m})\| \\
&\quad+\frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i-\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|+\|\underline{u}_i+\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right)-p) \\
\geq\ &\sup_{\underset{\sim}{x}_k^0\in\operatorname{out}(\delta_N,\underline{\mu},\Sigma)}(\sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}\lambda_1}}\sqrt{\|S\|_2}-p).
\end{aligned}\tag{4.3}$$

Weiterhin ist nach Voraussetzung $b(S,\eta,\boldsymbol{\delta})=\infty$, das heißt, es ist

$$\limsup_{N\to\infty}(\sup_{\underset{\sim}{x}_k^0\in\operatorname{out}(\delta_N,\underline{\mu},\Sigma)}(\|S-\Sigma\|_2))=\infty,$$

und damit existiert zu jedem $R \in \mathbb{R}$ eine unendliche Indexmenge $I_R \subset \mathbb{N}, |I_R| = \infty$, so daß

$$\sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N, \underline{\mu}, \Sigma)} (\|S - \Sigma\|) > R \quad \forall\, N \in I_R.$$

Wegen $\|S - \Sigma\|_2 \le \|S\|_2 + \|\Sigma\|_2 = \|S\|_2 + \lambda_1$ muß daher

$$\sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N, \underline{\mu}, \Sigma)} \|S\|_2 > R - \lambda_1 \quad \forall\, N \in I_R$$

gelten, d.h.,

$$\limsup_{N \to \infty} \sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N, \underline{\mu}, \Sigma)} \|S\|_2 = \infty.$$

Zusammen mit (4.3) folgt die Behauptung analog zum Beweis von Satz 4.1.

□

Die Beschränktheit von $b(S, \eta, \boldsymbol{\delta})$ ist also ebenso wie die von $b(\underline{m}, \eta, \boldsymbol{\delta})$ eine notwendige Voraussetzung für die Beschränktheit von $B(\underline{\text{OR}}, \eta, \boldsymbol{\delta})$. Ein Identifizierer, der einen beschränkten maximalen asymptotischen Bias besitzen soll, kann nicht auf Schätzern mit beliebig großem asymptotischen Bias im Sinne der Definitionen 4.3 und 4.5 beruhen. Unter einer zusätzlichen Voraussetzung an die Normierungskonstante $c(p, N, \alpha_N)$ eines Identifizierers ist die Beschränktheit von $b(\underline{m})$ und $b(S)$ auch hinreichend für einen beschränkten Bias von $\underline{\text{OR}}$.

Satz 4.3

Seien die Voraussetzungen von Definition 4.2 gegeben. Sei $\underline{\text{OR}} = \underline{\text{OR}}(\underset{\sim}{x}_N, \alpha_N)$ ein Ausreißer–Identifizierer gemäß Definition 2.2, basierend auf den Schätzern $\underline{m}$ und S für $\underline{\mu}$ bzw. Σ sowie einer Normierungskonstante $c(p, N, \alpha_N)$. Für $c(p, N, \alpha_N)$ gelte zusätzlich die Bedingung $c(p, N, \alpha_N) = \mathcal{O}(\chi^2_{p;1-\alpha_N}) (N \to \infty)$, das heißt, es existiere eine Zahl $M \in \mathbb{R}$, so daß $\left| \frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}} \right| < M$ für $N \to \infty$. Dann gilt der Zusammenhang:

$$b(\underline{m}, \eta, \boldsymbol{\delta}) < \infty \text{ und } b(S, \eta, \boldsymbol{\delta}) < \infty \quad \Rightarrow \quad B(\underline{\text{OR}}, \eta, \boldsymbol{\delta}) < \infty.$$

Beweis

Nach Definition ist

$$\begin{aligned} B(\underline{\mathrm{OR}}) &= B(\underline{\mathrm{OR}}, \eta, \boldsymbol{\delta}) = \limsup_{N\to\infty}\Big(\sup_{\underset{\sim}{x}_k^0 \in \,\mathrm{out}(\delta_N,\underline{\mu},\Sigma)} (\|\Sigma^{-1/2}(\underline{\mu}-\underline{m})\| \\ &+ \frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i - \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\| + \|\underline{u}_i + \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right) - p)\Big), \end{aligned}$$

wobei λ_i bzw. ξ_i die Eigenwerte von Σ bzw. S mit jeweils zugehörigen normierten Eigenvektoren $\underline{u}_i$ bzw. $\underline{v}_i$ sind, $i = 1, \ldots, p$, und jeweils $\lambda_p \leq \ldots \leq \lambda_1$ und $\xi_p \leq \ldots \leq \xi_1$ gilt. Nun ist

$$\begin{aligned} &\|\Sigma^{-1/2}(\underline{\mu}-\underline{m})\| \\ &+\frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i - \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\| + \|\underline{u}_i + \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\underline{v}_i\|\right) - p \\ \leq\; &\|\Sigma^{-1/2}\|_2\|\underline{\mu}-\underline{m}\| \\ &+\frac{1}{2}\sum_{i=1}^{p}\left(\|\underline{u}_i\| + \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\|\underline{v}_i\| + \|\underline{u}_i\| + \sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}}\,\|\underline{v}_i\|\right) - p \\ =\; &\frac{1}{\sqrt{\lambda_p}}\,\|\underline{\mu}-\underline{m}\| + \sum_{i=1}^{p}\sqrt{\frac{c(p,N,\alpha_N)\xi_i}{\chi^2_{p;1-\alpha_N}\lambda_i}} \\ \leq\; &\frac{1}{\sqrt{\lambda_p}}\,\|\underline{\mu}-\underline{m}\| + \sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}}\sum_{i=1}^{p}\sqrt{\frac{\xi_i}{\lambda_p}} \\ \leq\; &\frac{1}{\sqrt{\lambda_p}}\,\|\underline{\mu}-\underline{m}\| + p\sqrt{\frac{c(p,N,\alpha_N)\xi_1}{\chi^2_{p;1-\alpha_N}\lambda_p}} \\ \leq\; &\frac{1}{\sqrt{\lambda_p}}\,\|\underline{\mu}-\underline{m}\| + p\sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}\lambda_p}}\max\{1,\xi_1\} \end{aligned}$$

(wegen $\sqrt{\xi_1} \leq \max\{1, \xi_1\}$)

$$\begin{aligned} =\; &\frac{1}{\sqrt{\lambda_p}}\,\|\underline{\mu}-\underline{m}\| + p\sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}\lambda_p}}\max\{1,\|S\|_2\} \\ \leq\; &\frac{1}{\sqrt{\lambda_p}}\,\|\underline{\mu}-\underline{m}\| + p\sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}\lambda_p}}\max\{1,\|S-\Sigma\|_2+\lambda_1\} \end{aligned}$$

(wegen $\|S\|_2 \leq \|S-\Sigma\|_2 + \|\Sigma\|_2 = \|S-\Sigma\|_2 + \lambda_1$).

Diese Ungleichung bleibt für jedes $N \in I\!N$ erhalten, wenn auf beiden Seiten das Supremum über alle $\underset{\sim}{x}_k^0 \in$ out$(\delta_N, \underline{\mu}, \Sigma)$ gebildet wird. Mit dem Übergang zum Limes

Superior folgt (Blatter (1979), S. 138):

$$\begin{aligned}
B(\underline{\mathrm{OR}}) &\leq \limsup_{N\to\infty}\Big(\sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N,\underline{\mu},\Sigma)} \left(\frac{1}{\sqrt{\lambda_p}} \|\underline{\mu}-\underline{m}\| \right.\\
&\quad \left. + \, p\sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}\lambda_p}} \max\{1, \|S-\Sigma\|_2 + \lambda_1\} \right)\Big)\\
&\leq \limsup_{N\to\infty}\Big(\sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N,\underline{\mu},\Sigma)} \left(\frac{1}{\sqrt{\lambda_p}} \|\underline{\mu}-\underline{m}\| \right)\Big)\\
&\quad + \limsup_{N\to\infty}\Big(\sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N,\underline{\mu},\Sigma)} \left(p\sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}\lambda_p}} \max\{1, \|S-\Sigma\|_2 + \lambda_1\} \right)\Big)\\
&= \frac{1}{\sqrt{\lambda_p}}\, b(\underline{m})\\
&\quad + \frac{p}{\sqrt{\lambda_p}} \limsup_{N\to\infty}\left(\sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}} \sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N,\underline{\mu},\Sigma)} (\max\{1, \|S-\Sigma\|_2 + \lambda_1\})\right).
\end{aligned}$$

Wegen der Voraussetzung an das Verhalten von $c(p,N,\alpha_N)$ mit $N \to \infty$ ist $\sqrt{\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}} < \max\{1, \sqrt{M}\}$, falls N groß genug ist, so daß weiter

$$\begin{aligned}
B(\underline{\mathrm{OR}}) &\leq \frac{1}{\sqrt{\lambda_p}}\, b(\underline{m})\\
&\quad + \frac{p}{\sqrt{\lambda_p}} \max\{1,\sqrt{M}\} \limsup_{N\to\infty}\Big(\sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N,\underline{\mu},\Sigma)} (\max\{1, \|S-\Sigma\|_2 + \lambda_1\})\Big)\\
&= \frac{1}{\sqrt{\lambda_p}}\, b(\underline{m})\\
&\quad + \frac{p}{\sqrt{\lambda_p}} \max\{1,\sqrt{M}\} \max\{1, \lambda_1 + \limsup_{N\to\infty}\Big(\sup_{\underset{\sim}{x}_k^0 \in \operatorname{out}(\delta_N,\underline{\mu},\Sigma)} (\|S-\Sigma\|_2)\Big)\}\\
&= \frac{1}{\sqrt{\lambda_p}}\, b(\underline{m}) + \frac{p}{\sqrt{\lambda_p}} \max\{1,\sqrt{M}\} \max\{1, \lambda_1 + b(S)\} < \infty.
\end{aligned}$$

Also folgt unter den gegebenen Voraussetzungen insgesamt:

$$b(\underline{m}) < \infty \text{ und } b(S) < \infty \quad \Rightarrow \quad B(\underline{\mathrm{OR}}) < \infty.$$

$\square$

Satz 4.3 liefert eine hinreichende Bedingung für die Beschränktheit von $B(\underline{\mathrm{OR}})$. Der maximale asymptotische Bias eines Ausreißer–Identifizierers bleibt beschränkt, falls die Biaswerte der verwendeten Schätzer $\underline{m}$ und S beschränkt sind und die Normierungskonstante $c(p, N, \alpha_N)$ mit wachsendem Stichprobenumfang N nicht zu schnell wächst. Genauer gesagt, darf $c(p, N, \alpha_N)$ höchstens so schnell wachsen wie das $(1-\alpha_N)$–Quantil der χ^2_p–Verteilung, wenn N beliebig groß wird. Zusätzlich ist durch die Beziehung

$$B(\underline{\mathrm{OR}}) \leq \frac{1}{\sqrt{\lambda_p}}\, b(\underline{m}) + \frac{p}{\sqrt{\lambda_p}} \max\{1, \sqrt{M}\} \max\{1, \lambda_1 + b(S)\},$$

die sich im Beweis des Satzes ergibt, eine Abschätzung für die Größe von $\mathrm{B}(\underline{\mathrm{OR}})$ gegeben, wenn $b(\underline{m})$ und $b(S)$ sowie die Größenordnung des Wachstums von $c(p, N, \alpha_N)$ bekannt sind. Die Größe des maximalen asymptotischen Bias von $\underline{\mathrm{OR}}$ wird somit unmittelbar durch die Biaswerte der $\underline{\mathrm{OR}}$ bestimmenden Schätzer beeinflußt.

4.2 Zusammenhänge zwischen maximalem asymptotischem Bias und dem Finite–sample Bruchpunkt

Für die Bestimmung des maximalen asymptotischen Bias eines Identifizierers kann man nach den Überlegungen des vorigen Abschnitts zunächst die Biaswerte der beiden beteiligten Schätzer berechnen. Diese stehen in enger Beziehung zu den Bruchpunkten der Schätzer. Ein solcher Zusammenhang ist intuitiv einleuchtend. Bei der Bestimmung des Bruchpunkts wird der kleinste Anteil an Ausreißern gesucht, der den Schätzer beliebig weit von seinem „Ziel“ entfernt. Bei der Biasbestimmung wird die maximale Entfernung vom Ziel bei einem vorgegebenen Anteil an Ausreißern gemessen. Wenn dieser Anteil den Bruchpunkt des Schätzers übersteigt, ist zu erwarten, daß die Entfernung beliebig groß wird, unabhängig von der Größe der Stichprobe.

Huber (1981, S. 13) stellt eine ähnliche Überlegung in umgekehrter Richtung an. Er definiert den asymptotischen Bruchpunkt eines Schätzers über seinen maximalen asymptotischen Bias. Es wird mit dem Zusammenbruch des Schätzers gleichgesetzt, wenn sein Bias den schlechtestmöglichen Wert annimmt.

Der Zusammenhang zwischen Bruchpunkten und Bias wird in den beiden folgenden Sätzen präzisiert.

Satz 4.4

Sei $\underline{m} = \underline{m}(\underset{\sim}{x}_N)$ ein Lokationsschätzer mit Finite–sample Bruchpunkt $\varepsilon^*(\underset{\sim}{x}_N, \underline{m}) = \frac{k_1}{N}$, $k_1 \geq 1$. Dann gilt für den maximalen asymptotischen Bias:

$$b(\underline{m}, \eta, \boldsymbol{\delta}) \quad = \quad \infty$$

für alle Zahlen η mit $k = [\eta n] \geq k_1$.

Beweis

Betrachte zunächst einen festen Wert $N \in I\!N$.

Es ist

$$\varepsilon^*(\underset{\sim}{x}_N^r, \underline{m}) := \min_{1 \leq k \leq N} \{\frac{k}{N} : \sup_{\underset{\sim}{y}_{N,k}} \|\underline{m}(\underset{\sim}{x}_N^r) - \underline{m}(\underset{\sim}{y}_{N,k})\| = \infty\} = \frac{k_1}{N}.$$

Dabei ist $\underset{\sim}{x}_N^r$ eine Stichprobe vom Umfang N, die nur reguläre Beobachtungen enthält, $\underset{\sim}{y}_{N,k}$ eine Stichprobe vom Umfang N mit n regulären und k beliebigen Beobachtungen, $N = n + k$.

Nun gilt:

$$\begin{aligned} &\|\underline{m}(\underset{\sim}{x}_N^r) - \underline{m}(\underset{\sim}{y}_{N,k_1})\| = \|\underline{m}(\underset{\sim}{x}_N^r) - \underline{\mu} - (\underline{m}(\underset{\sim}{y}_{N,k_1}) - \underline{\mu})\| \\ &\leq \|\underline{m}(\underset{\sim}{x}_N^r) - \underline{\mu}\| + \|\underline{m}(\underset{\sim}{y}_{N,k_1}) - \underline{\mu}\|. \end{aligned} \tag{4.4}$$

Da nach Voraussetzung $\varepsilon^*(\underset{\sim}{x}_N^r, \underline{m}) = \frac{k_1}{N}$ ist, muß

$$\sup_{\underset{\sim}{x}_{k_1}^0 \in \operatorname{out}(\delta_N, \underline{\mu}, \Sigma)} \|\underline{m}(\underset{\sim}{x}_N^r) - \underline{\mu}\| < \infty$$

gelten.

Außerdem folgt aus der Größe des Bruchpunkts von $\frac{k_1}{N}$, daß

$$\sup_{\underset{\sim}{x}_{k_1}^0 \in \operatorname{out}(\delta_N, \underline{\mu}, \Sigma)} \|\underline{m}(\underset{\sim}{x}_N^r) - \underline{m}(\underset{\sim}{x}_N)\| = \infty,$$

wobei $\underset{\sim}{x}_N^r = (\underline{x}_1^r, \ldots, \underline{x}_N^r)$ und $\underset{\sim}{x}_N = (\underline{x}_{i_1}^r, \ldots, \underline{x}_{i_n}^r, \underline{x}_1^0, \ldots, \underline{x}_{k_1}^0) = (\underset{\sim}{x}_n^r, \underset{\sim}{x}_{k_1}^0)$ ist.

Aufgrund von Ungleichung (4.4) folgt dann:

$$\sup_{\underset{\sim}{x}_{k_1}^0 \in \operatorname{out}(\delta_N, \underline{\mu}, \Sigma)} \|\underline{m}(\underset{\sim}{x}_N) - \underline{\mu}\| = \infty$$

für jeden Wert von N.

Das gleiche gilt für jede Zahl k mit $k > k_1$ mit derselben Argumentation.

Also folgt insgesamt die Behauptung.

$\square$

Für eine entsprechende Aussage über Kovarianzschätzer muß wie erwartet dieselbe Bedingung wie in Satz 3.4 an das Zusammenbruchverhalten gestellt werden.

Satz 4.5

Sei $S = S(\underset{\sim}{x}_N)$ ein Kovarianzschätzer mit Finite–sample Bruchpunkt $\varepsilon^*(\underset{\sim}{x}_N, S) = \frac{k_2}{N}$, $k_2 \geq 1$. Es existiere keine Kombination von Beobachtungen, die einen Zusammenbruch des Schätzers S durch seinen kleinsten Eigenwert verursacht. Dann gilt für den maximalen asymptotischen Bias:

$$b(S, \eta, \boldsymbol{\delta}) \quad = \quad \infty$$

für alle Zahlen η mit $k = [\eta n] \geq k_2$.

Beweis

Der Beweis verläuft analog zu dem von Satz 4.4. Zu jedem $N \in I\!N$ betrachtet man

$$\begin{aligned}
\varepsilon^*(\underset{\sim}{x}^r_N, S) \quad &= \quad \min_{1\leq k\leq N}\{\frac{k}{N} : \sup_{\underset{\sim}{y}_{N,k}} \max\{|\xi_1(S(\underset{\sim}{x}^r_N)) - \xi_1(S(\underset{\sim}{y}_{N,k}))|, \\
&\qquad |\frac{1}{\xi_p(S(\underset{\sim}{x}^r_N))} - \frac{1}{\xi_p(S(\underset{\sim}{y}_{N,k}))}|\} = \infty\} \\
&= \quad \frac{k_2}{N}.
\end{aligned}$$

Dabei sind $\underset{\sim}{x}^r_N, \underset{\sim}{y}_{N,k}$ wie vorher, und $\xi_1(.), \xi_p(.)$ ist jeweils der größte bzw. kleinste Eigenwert der entsprechenden Matrix.

Wegen der Voraussetzung an die Art des Zusammenbruchs für den Schätzer S wird das Supremum in $\varepsilon^*(\underset{\sim}{x}^r_N, S)$ von der Differenz der größten Eigenwerte angenommen. Wie im Beweis von Satz 4.4 wendet man die Dreiecksungleichung an und erhält

$$|\xi_1(S(\underset{\sim}{x}^r_N)) - \xi_1(S(\underset{\sim}{y}_{N,k_2}))| \leq |\xi_1(S(\underset{\sim}{x}^r_N)) - \lambda_1| + |\xi_1(S(\underset{\sim}{y}_{N,k_2})) - \lambda_1|,$$

wobei λ_1 der größte Eigenwert von Σ ist.

Mit der gleichen Argumentation wie vorher gilt einerseits

$$\sup_{\underset{\sim}{x}^0_{k_2} \in \mathrm{out}(\delta_N, \underline{\mu}, \Sigma)} |\xi_1(S(\underset{\sim}{x}^r_N)) - \lambda_1| < \infty,$$

andererseits

$$\sup_{\underset{\sim}{x}^0_{k_2} \in \mathrm{out}(\delta_N, \underline{\mu}, \Sigma)} |\xi_1(S(\underset{\sim}{x}^r_N)) - \xi_1(S(\underset{\sim}{x}_N))| = \infty,$$

woraus entsprechend folgt

$$\sup_{\underset{\sim}{x}^0_{k_2} \in \mathrm{out}(\delta_N, \underline{\mu}, \Sigma)} |\xi_1(S(\underset{\sim}{x}_N)) - \lambda_1| = \infty.$$

Weiter gilt dann wegen $\xi_1(S) = \|S\|_2 \leq \|S - \Sigma\|_2 + \|\Sigma\|_2 = |\zeta(S - \Sigma)| + \lambda_1$ (mit λ_1 größter Eigenwert von Σ und ζ betragsmäßig größter Eigenwert von $S - \Sigma$):

$$\begin{aligned} |\xi_1(S(\underset{\sim}{x}_N)) - \lambda_1| &\leq |\xi_1(S(\underset{\sim}{x}_N))| + |\lambda_1| = \xi_1(S(\underset{\sim}{x}_N)) + \lambda_1 \\ &\leq |\zeta(S(\underset{\sim}{x}_N) - \Sigma)| + 2\lambda_1. \end{aligned}$$

Daraus folgt die Beziehung

$$\sup_{\underset{\sim}{x}^0_{k_2} \in \mathrm{out}(\delta_N, \underline{\mu}, \Sigma)} |\zeta(S(\underset{\sim}{x}_N) - \Sigma)| = \infty.$$

Diese Aussage gilt für jeden Wert von N und kann in entsprechender Weise für jedes $k > k_2$ hergeleitet werden.

Damit folgt die Behauptung.

$\square$

Der maximale asymptotische Bias von Lokations- und Kovarianzschätzern hängt also eng mit den Finite–sample Bruchpunkten der Schätzer zusammen. Das bedeutet, daß auch der Bias eines Ausreißer–Identifizierers mit den Bruchpunkten der in ihn eingehenden Schätzer eng verbunden ist. Speziell kann man aus den vorangegangenen Sätzen sofort folgern, daß $B(\underline{\mathrm{OR}}, \eta, \boldsymbol{\delta}) = \infty$ ist, sobald der Anteil $\frac{k}{N} = \frac{[\eta n]}{N}$ an Ausreißern in der Stichprobe größer ist als der kleinere Finite–sample Bruchpunkt der beiden Schätzer. Will man also einen Identifizierer konstruieren, dessen maximaler asymptotischer Bias für einen bestimmten Anteil $\frac{k}{N}$ von Ausreißern in einer Stichprobe beschränkt ist, so müssen Schätzer mit einem Bruchpunkt größer als $\frac{k}{N}$ verwendet werden. Im folgenden Abschnitt werden weitere Anforderungen an die verwendeten Schätzer untersucht.

4.3 Untersuchung des Wachstums von Normierungskonstanten

In Satz 4.3 zeigte sich, daß der maximale asymptotische Bias eines Identifizierers nicht nur von den verwendeten Schätzern, sondern darüber hinaus offenbar vom Wachstum des Normierungsfaktors $c(p, N, \alpha_N)$ abhängt. Im folgenden wird daher untersucht, wie das Wachstumsverhalten von $c(p, N, \alpha_N)$ mit Eigenschaften der in den Identifizierer eingehenden Schätzer zusammenhängt.
Benutzt man als Normierungsbedingung, daß

$$\mathrm{P}(\underline{X}_i \in I\!R^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha \tag{4.5}$$

für $\alpha \in]0,1[$, $\alpha_N = 1 - (1-\alpha)^{1/N}$, mit $\underset{\sim}{X}_N = (\underline{X}_1, \ldots, \underline{X}_N)$, $\underline{X}_1, \ldots, \underline{X}_N$ stochastisch unabhängig und identisch $N(\underline{\mu}, \Sigma)$–verteilt,
so wird die in Satz 4.3 geforderte Bedingung an $c(p, N, \alpha_N)$ für Identifizierer erfüllt, die auf konsistenten Schätzern $\underline{m}$ und S mit ausreichender Konsistenzordnung ($\sqrt{N}$–Konsistenz) beruhen. Um dies nachzuweisen, muß zunächst die Verteilung der Zufallsvariablen $(\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m})$ bestimmt werden, wobei $(\underline{m}, S)$ ein Paar von geeignet konsistenten Schätzern sei. Die Aussage des folgenden Satzes ist zum Beispiel den Arbeiten von Witting, Nölle (1970, S. 45) und Fahrmeir et al. (1996, S. 31 ff.) zu entnehmen.

Satz 4.6

Seien $\underline{X}_1, \ldots, \underline{X}_N$ stochastisch unabhängige und identisch $N(\underline{\mu}, \Sigma)$–verteilte Zufallsvektoren, sei $(\underline{m}, S)$ ein Paar von $\sqrt{N}$–konsistenten Schätzern für $(\underline{\mu}, \Sigma)$. Dann sind die quadrierten Distanzen $(\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m})$ vom Mahalanobis-Typ asymptotisch $(N \to \infty)$ χ^2-verteilt mit p Freiheitsgraden.

Beweis

Witting, Nölle (1970, S. 45), Fahrmeir et al. (1996, S. 31 ff.)

□

Betrachtet man die Funktion $g(\underline{m}) := (\underline{x} - \underline{m})^T S^{-1} (\underline{x} - \underline{m})$ in Abhängigkeit von $\underline{m}$, so läßt sich die Verteilungsapproximation durch eine Taylorreihenentwicklung zweiter Ordnung verdeutlichen. Entwickelt man g um die Stelle $\underline{\mu}$, dann erhält man die folgende Darstellung:

$$\begin{aligned}
g(\underline{m}) &= g(\underline{\mu}) + \left[\left. \frac{\partial g(\underline{m})}{\partial \underline{m}} \right|_{\underline{m}=\underline{\mu}} \right]^T (\underline{m} - \underline{\mu}) \\
&\quad + (\underline{m} - \underline{\mu})^T \left. \left[\frac{\partial}{\partial \underline{m}} \left(\frac{\partial g(\underline{m})}{\partial \underline{m}} \right)^T \right] \right|_{\underline{m}=\underline{\mu}} (\underline{m} - \underline{\mu}) + \text{ Restglied} \\
&= (\underline{x} - \underline{\mu})^T S^{-1} (\underline{x} - \underline{\mu}) + 2(\underline{\mu} - \underline{x})^T S^{-1} (\underline{m} - \underline{\mu}) \\
&+ 2(\underline{m} - \underline{\mu})^T S^{-1} (\underline{m} - \underline{\mu}) + \text{ Restglied.}
\end{aligned}$$

Die Konsistenzordnung von $\sqrt{N}$ liefert die Vernachlässigbarkeit des Restglieds.
Unter Ausnutzung der Konsistenz von $\underline{m}$ für $\underline{\mu}$ und S für Σ und mit dem bekannten Zusammenhang $(\underline{X}_i - \underline{\mu})^T \Sigma^{-1} (\underline{X}_i - \underline{\mu}) \sim \chi^2_p$ ergibt sich die asymptotische χ^2–Verteilung der quadrierten Distanzen.

Betrachtet man nun die oben genannte Normierungsbedingung, so erhält man

$$\begin{aligned}
&\mathrm{P}(\underline{X}_i \in \mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha \\
\Leftrightarrow\ &\mathrm{P}(\max_{i=1,\ldots,N} (\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m}) < c(p, N, \alpha_N)) = 1 - \alpha.
\end{aligned}$$

Damit ergibt sich die Konstante $c(p, N, \alpha_N)$ als das $(1-\alpha)$–Quantil der Verteilung von $\max_{i=1,\ldots,N} (\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m})$. Unter Ausnutzung von Satz 4.6 kann man diese Verteilung asymptotisch bestimmen.

Satz 4.7

Seien unter den Voraussetzungen von Satz 4.6 die Zufallsvariablen Y_i definiert als $Y_i := (\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m})$, $i = 1, \ldots, N$. Dann gilt:

$$\lim_{N \to \infty} \mathrm{P} \left(N f_{\chi^2_p}(\chi^2_{p;1-1/N})(\max(Y_1, \ldots, Y_N) - \chi^2_{p;1-1/N}) < y \right) = \exp(-\exp(-y)).$$

Dabei bezeichnet $f_{\chi^2_p}$ die Lebesgue–Dichte der χ^2–Verteilung mit p Freiheitsgraden.

Beweis

Zunächst stellt man fest, daß die χ^2–Verteilung im Maximums-Anziehungsbereich der doppelten Exponentialverteilung liegt. Dazu wird z.B. das folgende Resultat aus Galambos (1987, S. 102) herangezogen. Sei F eine Verteilungsfunktion mit Lebesgue–Dichte f und rechtem Trägerrandpunkt $\omega(F) \leq \infty$. Weiterhin existiere eine Zahl $x_1 \in I\!R$, so daß $\forall x : x_1 \leq x < \omega(F)$ die Ableitung $f'(x)$ existiert und $f(x) \neq 0$ ist. Dann liegt F im Maximums-Anziehungsbereich der doppelten Exponentialverteilung, wenn

$$\lim_{x\to\omega(F)} \frac{d}{dx}\left(\frac{1-F(x)}{f(x)}\right) = 0.$$

Für die χ^2–Verteilung ist damit zu überprüfen, ob der obige Grenzwert existiert und gleich Null ist. Die restlichen Bedingungen werden von Verteilungsfunktion und Dichte erfüllt. Es ist $f(x) = f_{\chi^2_p}(x) = 2^{-p/2}\Gamma(p/2)^{-1}x^{p/2-1}e^{-x/2}$, $x > 0$, und damit

$$\frac{d}{dx}\left(\frac{1-F(x)}{f(x)}\right) = -\frac{f'(x)(1-F(x))}{f^2(x)} - 1$$

und

$$\frac{f'(x)(1-F(x))}{f^2(x)} = \underbrace{\left((\frac{p}{2}-1)\frac{1}{x} - \frac{1}{2}\right)}_{\to -1/2 (x\to\infty)} \frac{1-F(x)}{f(x)}.$$

Durch Anwendung der Regel von l'Hospital erhält man

$$\lim_{x\to\infty} \frac{1-F(x)}{f(x)} = -\lim_{x\to\infty} \frac{f(x)}{f'(x)} = -\lim_{x\to\infty} \frac{1}{(\frac{p}{2}-1)\frac{1}{x} - \frac{1}{2}} = 2.$$

Insgesamt folgt daher:

$$\lim_{x\to\infty} \frac{d}{dx}\left(\frac{1-F(x)}{f(x)}\right) = -(-\frac{1}{2})\,2 - 1 = 0.$$

Damit liegt die χ^2–Verteilung im Maximums-Anziehungsbereich der doppelten Exponentialverteilung. Das heißt, es existieren Folgen a_N, b_N ($b_N > 0$), so daß

$$\lim_{N\to\infty} \mathrm{P}\left(\frac{\max(Y_1,\ldots,Y_N) - a_N}{b_N} < y\right) = \exp(-\exp(-y)).$$

Dabei sind die Folgen a_N, b_N wählbar als (Galambos (1987), S. 54, 105)

$$a_N = \inf\{x : 1 - F(x) \leq \frac{1}{N}\},\; b_N = \frac{1-F(a_N)}{f(a_N)}.$$

Damit folgt hier:

$$\begin{aligned} a_N &= \inf\{x : 1 - F(x) \leq \frac{1}{N}\} = \inf\{x : F(x) \geq 1 - \frac{1}{N}\} = F^{-1}(1 - \frac{1}{N}) \\ &= \chi^2_{p;1-1/N}, \end{aligned}$$

wobei F^{-1} die Quantilfunktion der χ^2_p–Verteilung bezeichne, und

$$b_N = \frac{1 - F(\chi^2_{p;1-1/N})}{f(\chi^2_{p;1-1/N})} = \frac{\frac{1}{N}}{f(\chi^2_{p;1-1/N})} = \frac{1}{Nf(\chi^2_{p;1-1/N})}$$

mit F, f Verteilungsfunktion bzw. Lebesgue-Dichte der χ^2_p–Verteilung.
Also gilt:

$$\lim_{N\to\infty} \mathrm{P}\left(Nf_{\chi^2_p}(\chi^2_{p;1-1/N})(\max(Y_1,\ldots,Y_N) - \chi^2_{p;1-1/N}) < y\right) = \exp(-\exp(-y))$$

bzw.

$$\mathrm{P}\left(\max(Y_1,\ldots,Y_N) < y\right) \simeq \exp(-\exp(-Nf_{\chi^2_p}(\chi^2_{p;1-1/N})(y - \chi^2_{p;1-1/N})))$$

für große Stichprobenumfänge N.

□

Unmittelbar ergibt sich das folgende Korollar.

Korollar 4.1

Für einen Identifizierer, der auf konsistenten Schätzern für Lage und Kovarianz beruht ($\sqrt{N}$–Konsistenz), läßt sich unter der oben angegebenen Normierungsbedingung (4.5) die Normierungskonstante $c(p, N, \alpha_N)$ für große Stichprobenumfänge approximieren durch das $(1-\alpha)$–Quantil der doppelten Exponentialverteilung mit Parametern $\chi^2_{p;1-1/N}$ und $\dfrac{1}{Nf_{\chi^2_p}(\chi^2_{p;1-1/N})}$, das heißt:

$$c(p, N, \alpha_N) \simeq \chi^2_{p;1-1/N} - \frac{\ln(-\ln(1-\alpha))}{Nf_{\chi^2_p}(\chi^2_{p;1-1/N})}$$

mit $\alpha_N = 1 - (1-\alpha)^{1/N}$.

Mit dieser Näherung läßt sich nun die Größenordnung des Wachstums der Normierungskonstanten $c(p, N, \alpha_N)$ für Identifizierer bestimmen, die auf konsistenten Schätzern beruhen.

Satz 4.8

Für die Normierungskonstante $c(p,N,\alpha_N)$ eines Identifizierers $\underline{\mathrm{OR}}$, der auf einem Paar $(\underline{m},S)$ von $\sqrt{N}$–konsistenten Schätzern für Lage und Kovarianz beruht, gilt:

$$c(p,N,\alpha_N) = \mathcal{O}(\chi^2_{p;1-\alpha_N})\ (N \to \infty),$$

falls c der Normierungsbedingung (4.5) gehorcht.

Beweis

Nach Korollar 4.1 läßt sich die Normierungskonstante $c(p,N,\alpha_N)$ eines Identifizierers $\underline{\mathrm{OR}}$, der auf $\sqrt{N}$–konsistenten Schätzern beruht, mit wachsendem Stichprobenumfang schreiben als

$$c(p,N,\alpha_N) \simeq \chi^2_{p;1-1/N} - \frac{\ln(-\ln(1-\alpha))}{N f_{\chi^2_p}(\chi^2_{p;1-1/N})}$$

mit $\alpha_N = 1-(1-\alpha)^{1/N}$.Damit ist

$$\begin{aligned}
&\lim_{N\to\infty}\left(\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}\right)\\
=\ &\lim_{N\to\infty}\left(\frac{\chi^2_{p;1-1/N}}{\chi^2_{p;(1-\alpha)^{1/N}}} + \frac{-\ln(-\ln(1-\alpha))}{\chi^2_{p;(1-\alpha)^{1/N}} N f_{\chi^2_p}(\chi^2_{p;1-1/N})}\right).
\end{aligned}$$

Betrachtet man den zweiten Summanden, so erhält man zunächst mit Hilfe der Regeln von l'Hospital

$$\lim_{N\to\infty}\Big(N f_{\chi^2_p}(\chi^2_{p;1-1/N})\Big) = \lim_{N\to\infty}\left(\frac{1}{2} - \frac{p-2}{2\chi^2_{p;1-1/N}}\right) = \frac{1}{2},$$

also

$$\lim_{N\to\infty}\left(\frac{1}{N f_{\chi^2_p}(\chi^2_{p;1-1/N})}\right) = 2$$

und insgesamt

$$\lim_{N\to\infty}\left(\frac{-\ln(-\ln(1-\alpha))}{\chi^2_{p;(1-\alpha)^{1/N}} N f_{\chi^2_p}(\chi^2_{p;1-1/N})}\right) = 0.$$

Auch der Grenzwert des ersten Summanden läßt sich durch Anwendung der l'Hospitalschen Regeln bestimmen. Es ergibt sich

$$
\begin{aligned}
\lim_{N\to\infty}\left(\frac{\chi^2_{p;1-1/N}}{\chi^2_{p;(1-\alpha)^{1/N}}}\right) &= \lim_{N\to\infty}\left(\frac{F^{-1}_{\chi^2_p}(1-1/N)}{F^{-1}_{\chi^2_p}((1-\alpha)^{1/N})}\right)\\
&= \lim_{N\to\infty}\left(\frac{f_{\chi^2_p}(\chi^2_{p;(1-\alpha)^{1/N}})}{-(1-\alpha)^{1/N}\ln(1-\alpha)f_{\chi^2_p}(\chi^2_{p;1-1/N})}\right)\\
&= \lim_{N\to\infty}\left(\frac{\dfrac{p-2}{2\chi^2_{p;(1-\alpha)^{1/N}}}-\dfrac{1}{2}}{-\ln(1-\alpha)f_{\chi^2_p}(\chi^2_{p;1-1/N})+\dfrac{p-2}{2\chi^2_{p;1-1/N}}-\dfrac{1}{2}}\right)\\
&= 1.
\end{aligned}
$$

Insgesamt gilt daher

$$\lim_{N\to\infty}\left(\frac{c(p,N,\alpha_N)}{\chi^2_{p;1-\alpha_N}}\right)=1,$$

und damit ist auch $c(p,N,\alpha_N)=\mathcal{O}(\chi^2_{p;1-\alpha_N})\ (N\to\infty)$.

□

Bei Verwendung $\sqrt{N}$–konsistenter Schätzer und der Normierungsbedingung (4.5) ergibt sich also eine Konstante $c(p,N,\alpha_N)$, die die in Satz 4.3 geforderte Wachstumsbedingung erfüllt. Zusammen mit den bisherigen Ergebnissen kommt man daher zu der Empfehlung, für einen Identifizierer $\sqrt{N}$–konsistente Lage- und Kovarianzschätzer mit hohen Bruchpunkten zu verwenden, deren maximaler asymptotischer Bias jeweils beschränkt ist.

4.4 Der maximale asymptotische Bias eines klassischen und eines robusten Identifizierers

Zur Beurteilung der Größe des maximalen asymptotischen Bias eines Identifizierers wird zunächst das Beispiel des klassischen Ausreißer–Identifizierers $\underline{\mathrm{OR}}_{\mathrm{MD}}$ betrachtet. Dieser Identifizierer beruht auf den klassischen Schätzern für $\underline{\mu}$ und Σ, dem arithmetischen Mittel $\overline{\underline{x}}_N$ und der Stichprobenkovarianzmatrix S_N der Beobachtungen. Er ist definiert als

$$\underline{\mathrm{OR}}_{\mathrm{MD}}(\underset{\sim}{x}_N, \alpha_N) := \{\underline{x} \in I\!R^p : (\underline{x} - \overline{\underline{x}}_N)^T S_N^{-1} (\underline{x} - \overline{\underline{x}}_N) \geq \chi^2_{p;1-\alpha_N}\}.$$

Dabei sind $\overline{\underline{x}}_N = \frac{1}{N}\sum_{i=1}^N \underline{x}_i$ und $S_N = \frac{1}{N-1}\sum_{i=1}^N (\underline{x}_i - \overline{\underline{x}}_N)(\underline{x}_i - \overline{\underline{x}}_N)^T$ bekanntermaßen konsistent für $\underline{\mu}$ und Σ mit ausreichender Konsistenzordnung (vgl. Lehmann (1983), S. 438 ff.). Dennoch hat $\underline{\mathrm{OR}}_{\mathrm{MD}}$ einen unbeschränkten maximalen asymptotischen Bias, da die Biaswerte der beiden Schätzer nicht beschränkt sind. Dies ist eine direkte Schlußfolgerung aus den Ergebnissen des Abschnitts 4.2.

Korollar 4.2

(a) Für das arithmetische Mittel $\overline{\underline{x}}_N$ gilt:

$$b(\overline{\underline{x}}_N, \eta, \boldsymbol{\delta}) \quad = \quad \infty \qquad \forall \eta, 0 < \eta < 1.$$

(b) Falls die Stichprobe $\underset{\sim}{x}_n^r = (\underline{x}_1^r, \ldots, \underline{x}_n^r)$ der regulären Beobachtungen für jede Zahl $n \in I\!N$ mit $n > p$ in allgemeiner Lage ist, gilt für die Stichprobenkovarianzmatrix S_N:

$$b(S_N, \eta, \boldsymbol{\delta}) \quad = \quad \infty \qquad \forall \eta, 0 < \eta < 1.$$

Beweis

Beachtet man, daß für die Finite–sample Bruchpunkte der beiden Schätzer gilt $\varepsilon^*(\overline{\underline{x}}_N) = \varepsilon^*(S_N) = \frac{1}{N}$, so folgen die Behauptungen unmittelbar aus den Sätzen 4.4 und 4.5.

□

Mit den Ergebnissen aus Korollar 4.2 kann der maximale asymptotische Bias des klassischen Ausreißer–Identifizierers $\underline{\mathrm{OR}}_{\mathrm{MD}}$ angegeben werden.

Korollar 4.3

Unter den Voraussetzungen von Korollar 4.2 gilt für den Identifizierer $\underline{\mathrm{OR}}_{\mathrm{MD}}$:

$$B(\underline{\mathrm{OR}}_{\mathrm{MD}}, \eta, \boldsymbol{\delta}) \quad = \quad \infty$$

$\forall \eta, 0 < \eta < 1.$

Im maximalen asymptotischen Bias des Identifizierers $\underline{\mathrm{OR}}_{\mathrm{MD}}$ spiegeln sich die niedrigen Bruchpunkte der Schätzer $\underline{\overline{x}}_N$ und S_N wider. Der klassische Identifizierer kann also selbst für kleinste Anteile an Ausreißern als Schätzer für den Ausreißer–Bereich völlig versagen, unabhängig davon, wie groß der Stichprobenumfang ist.
Im Gegensatz dazu hat ein Identifizierer, der auf den in Kapitel 3 definierten $\mathrm{S_{MB}}$–Schätzern beruht, einen beschränkten maximalen asymptotischen Bias.

Korollar 4.4

Für einen Identifizierer $\underline{\mathrm{OR}}$, der auf $\mathrm{S_{MB}}$–Schätzern für Lage und Kovarianz beruht, gilt bei Normierung gemäß (4.5):

$$B(\underline{\mathrm{OR}}, \eta, \boldsymbol{\delta}) \quad < \quad \infty$$

$\forall \eta, 0 < \eta < 1.$

Beweis
Wegen der $\sqrt{N}$–Konsistenz von $\mathrm{S_{MB}}$–Schätzern (Davies (1987), S. 1282, 1284) gilt nach Satz 4.8 bei Normierung gemäß (4.5) für die Normierungskonstante des resultierenden Identifizierers die Wachstumsbedingung $c(p, N, \alpha_N) = \mathcal{O}(\chi^2_{p;1-\alpha_N})$ $(N \to \infty)$.
Weiterhin folgt aus der Herleitung des Finite–sample Bruchpunkts der $\mathrm{S_{MB}}$–Schätzer $(\underline{m}, S)$ in Davies (1987, S. 1287 ff.), daß $\|\underline{m}\|$ nach oben beschränkt ist und die Eigenwerte von S durch Konstanten $c_1 > 0$ und $c_2 < \infty$ beschränkt werden, wenn der Anteil an Ausreißern in der Stichprobe unterhalb des Finite–sample Bruchpunkts bleibt. Damit besitzen in diesem Fall $\underline{m}$ und S beschränkten maximalen asymptotischen Bias, wobei für eine feste Stichprobe vom Umfang $N = n + [n\eta]$ mit n regulären Beobachtungen die Zahl η zwischen 0 und $1 - \dfrac{p-1}{n}$ liegen darf; mit $N \to \infty$ und damit auch $n \to \infty$ gilt die Aussage also für $0 < \eta < 1$.
Insgesamt sind die Bedingungen von Satz 4.3 erfüllt, und daraus folgt die Behauptung.

□

Ein auf $\mathrm{S_{MB}}$–Schätzern basierender Ausreißer–Identifizierer genügt also sowohl den in Kapitel 3 als auch den in Kapitel 4 gestellten Anforderungen. Der in Kapitel 3 vorgestellte Identifizierer $\underline{\mathrm{OR}}_{\mathrm{MVE}}$, der MVE–Schätzer benutzt, kann zwar bezüglich der

Bruchpunktkriterien mithalten, jedoch nicht bezüglich des Biaskriteriums. Das liegt an der Konsistenzordnung der MVE–Schätzer. Zwar handelt es sich auch hier um konsistente Schätzer, jedoch lediglich mit einer Ordnung von $N^{1/3}$ (Davies (1992)).
Mit den Untersuchungen dieses Kapitels zeigt sich, daß bei der Identifizierung von Ausreißern die Verwendung hochrobuster Schätzer allein nicht ausreicht, um Verfahren mit vernünftigen Eigenschaften zu konstruieren. Vielmehr müssen auch asymptotische Verhaltensweisen der benutzten Schätzer in Betracht gezogen werden. Ein Kompromiß zwischen hoher Resistenz gegen das Versagen beim Auftreten großer Anzahlen von Ausreißern und schneller Konvergenz der Schätzer gegen die zugrundeliegenden Parameter liefert Identifizierungsverfahren, die den hier entwickelten Gütekriterien genügen. Die in Kapitel 3 eingeführten Identifizierer, die auf S–Schätzern mit maximalem Bruchpunkt beruhen, fallen in diese Kategorie. Sie sind nach den bisher gewonnenen Erkenntnissen zur Entdeckung von Ausreißern in multivariaten Datensätzen bei zugrundeliegender Normalverteilung zu empfehlen. Dabei sei an dieser Stelle nochmals darauf hingewiesen, daß in dieser Arbeit Verfahren speziell für bis zu ca. 5 Dimensionen der Daten untersucht worden sind. Bei höherdimensionalen Datensätzen treten Schwierigkeiten, ähnlich wie bei der Analyse von Strukturen in Daten, dadurch auf, daß der Stichprobenraum selbst mit einer großen Stichprobe nicht mehr ausreichend dicht besetzt werden kann und die hier benutzten Schätzer numerisch nicht mehr stabil zu bestimmen sind (vgl. dazu auch Rocke (1993)). Daher müssen Verfahren zur Identifizierung von Ausreißern gegebenenfalls den sich in solchen Fällen zum Beispiel durch den Fluch der Dimension ergebenden Erfordernissen angepaßt werden.

5 Anwendung von Identifizierern

Die Ergebnisse der bisherigen Kapitel führen zu der Empfehlung, für Ausreißer–Identifizierer robuste Schätzer mit hohem Bruchpunkt zu verwenden, die gleichzeitig eine ausreichend schnelle Konvergenz gegen die Parameter der zugrundeliegenden Normalverteilung aufweisen und damit beschränkten maximalen asymptotischen Bias besitzen. Speziell besitzen S–Schätzer die entsprechenden theoretischen Eigenschaften, die sie zur Verwendung in Identifizierern prädestinieren. In diesem Kapitel wird untersucht, ob auf S–Schätzern basierende Identifizierer in praktischen Situationen halten können, was ihre theoretischen Eigenschaften versprechen. Dazu werden vier aus der Literatur bekannte Datensätze herangezogen, in denen bereits mehrfach einige Beobachtungen als Ausreißer deklariert wurden. Sie werden mit Hilfe von zwei Varianten der in dieser Arbeit vorgeschlagenen Identifizierungsprozedur erneut auf Ausreißer untersucht, und die Ergebnisse werden mit den bereits bekannten verglichen. Zum Vergleich der beiden Verfahren wird abschließend kurz darauf eingegangen, wie „groß“ Ausreißer in einer Stichprobe sein können, ohne daß die Identifizierer sie erkennen.
In den folgenden Abschnitten werden die verwendeten Identifizierer vorgestellt und das Verfahren zur Gewinnung der benötigten kritischen Werte $c(p, N, \alpha_N)$ erläutert.

5.1 Die Identifizierer $\underline{\textbf{OR}}_{\text{BW}}$ und $\underline{\textbf{OR}}_{\text{TW}}$

Wie oben erwähnt, werden für die hier betrachteten Identifizierer S–Schätzer als Grundlage verwendet. In Kapitel 3.4 wurde eine Definition für S–Schätzer eingeführt. Alternativ dazu kann ein Paar $(\underline{m}, S)$ von S–Schätzern für Lage und Kovarianz auch anders bestimmt werden (vgl. Lopuhaä (1989), S. 1664 f.).

Lemma 5.1 (Lopuhaä (1989))

Seien $\underline{X}_1, \ldots, \underline{X}_N$ unabhängige, identisch $N(\underline{\mu}, \Sigma)$–verteilte Zufallsvariablen.
Die Lösungen $(\underline{m}, S)$ des folgenden Minimierungsproblems sind S–Schätzer für Lage und Kovarianz:

$$\min_{S \in \text{PDS}(p)} \det(S)$$

unter der Nebenbedingung

$$\frac{1}{N}\sum_{i=1}^{N}\rho\left(\sqrt{(\underline{X}_i-\underline{m})^T S^{-1}(\underline{X}_i-\underline{m})}\right) \;=\; b_0. \tag{5.1}$$

Dabei ist $\rho : I\!R \mapsto I\!R$ symmetrisch, besitzt eine stetige Ableitung ψ, und es ist $\rho(0)=0$. Weiterhin existiert eine Konstante $c_0>0$, so daß ρ streng monoton wachsend ist auf $[0,c_0]$ und konstant auf $]c_0,\infty[$. Die Konstante b_0 wird bestimmt durch $\mathrm{E}(\rho(D))=b_0$, $D^2\sim\chi^2_p$.
Setzt man $a_0:=\rho(c_0)=\sup\rho$ und $\kappa(y):=1-\rho(\sqrt{y})/a_0$, so erhält man das Minimierungsproblem für S–Schätzer aus Definition 3.5, wenn man die Bedingungen an ρ abschwächt. Es reicht zu fordern, daß ρ linksseitig stetig auf $]0,\infty[$, stetig in 0 und monoton wachsend auf $[0,c_0]$ ist. Außerdem wird die Gleichheit in (5.1) durch die Beziehung „$\leq$“ ersetzt. Unter diesen Bedingungen sind die beiden Definitionen äquivalent (Lopuhaä (1989), S. 1665).

Damit sind Lösungen des obigen Minimierungsproblems stets S–Schätzer gemäß Definition 3.5. Gleichzeitig eignet sich die oben angegebene Form des Problems besser zur konkreten Berechnung von S–Schätzern, denn aus ihr läßt sich ein Zusammenhang von S–Schätzern zu M–Schätzern herleiten (Lopuhaä (1989), S. 1665 ff.). Daher können Algorithmen, die zur Bestimmung von M–Schätzern dienen, zur Berechnung von S–Schätzern herangezogen werden.

Der erste hier betrachtete Identifizierer benutzt eine Funktion ρ, deren Ableitung ψ als Tukey's Biweight bekannt ist (Beaton, Tukey (1974)).

Definition 5.1

Der Ausreißer–Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$ ist definiert als

$$\underline{\mathrm{OR}}_{\mathrm{BW}} := \{\underline{x}\in I\!R^p : (\underline{x}-\underline{m}_{\mathrm{BW}})^T S_{\mathrm{BW}}^{-1}(\underline{x}-\underline{m}_{\mathrm{BW}}) \geq c_{\mathrm{BW}}(p,N,\alpha_N)\}.$$

Dabei sind $\underline{m}_{\text{BW}}$ und S_{BW} Lösungen des in Lemma 5.1 genannten Minimierungsproblems mit der folgenden Funktion $\rho = \rho_{\text{BW}} : I\!R_+ \mapsto I\!R$:

$$\rho_{\text{BW}}(d) = \begin{cases} \dfrac{d^2}{2} - \dfrac{d^4}{2c_0^2} + \dfrac{d^6}{6c_0^4} & , \quad 0 \leq d \leq c_0 \\ \dfrac{c_0^2}{6} & , \quad d > c_0 \end{cases} .$$

Weiterhin ist $c_0 \in I\!R$ so festgelegt, daß der Finite–sample Bruchpunkt von S_{BW} maximal ist, d. h., c_0 löst die Gleichung $\mathrm{E}(\rho(D)) = r\rho(c_0)$ mit $r = \frac{[\frac{N-p+1}{2}]}{N}$. Dabei ist D eine reelle Zufallsvariable und $D^2 \sim \chi_p^2$. Die Zahl b_0 aus (5.1) wird bestimmt durch $\mathrm{E}(\rho(D)) = b_0$ (vgl. Rocke (1993)).

Der kritische Wert $c_{\text{BW}}(p, N, \alpha_N)$ wird aus der Normierungsgleichung

$$P(\underline{X}_i \in I\!R^p \backslash \underline{\text{OR}}_{\text{BW}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha$$

für $\underline{X}_1, \ldots, \underline{X}_N$ i.i.d., $\underline{X}_i \sim N(\underline{\mu}, \Sigma)$, gewonnen, wobei α vorgegeben wird und $\alpha_N = 1 - (1-\alpha)^{1/N}$ gilt.

Die konkrete Bestimmung der Schätzer $\underline{m}_{\text{BW}}$ und S_{BW} erfolgt über ein iteratives Verfahren, das von Rocke (1993) allgemein beschrieben wird. Ausgehend von Startschätzern $(\underline{m}^{(0)}, S^{(0)})$ werden im j-ten Iterationsschritt die Distanzen

$$d_i^{(j)} = \sqrt{(\underline{x}_i - \underline{m}^{(j-1)})^T (S^{(j-1)})^{-1} (\underline{x}_i - \underline{m}^{(j-1)})}$$

und die Skalierungskonstante $k^{(j)}$ aus der Beziehung

$$\frac{1}{N} \sum_{i=1}^{N} \rho\left(\frac{d_i^{(j)}}{k^{(j)}}\right) = b_0$$

bestimmt.

Daraus werden die iterierten Schätzer

$$\underline{m}^{(j)} = \frac{\sum_{i=1}^{N} w\left(\frac{d_i^{(j)}}{k^{(j)}}\right) \underline{x}_i}{\sum_{i=1}^{N} w\left(\frac{d_i^{(j)}}{k^{(j)}}\right)}$$

und

$$S^{(j)} = \frac{p\sum_{i=1}^{N} w\left(\frac{d_i^{(j)}}{k^{(j)}}\right)(\underline{x}_i - \underline{m}^{(j)})(\underline{x}_i - \underline{m}^{(j)})^T}{\sum_{i=1}^{N} v\left(\frac{d_i^{(j)}}{k^{(j)}}\right)}$$

berechnet, wobei w die Gewichtsfunktion der Schätzer ist mit $w(d) = \frac{\psi(d)}{d}$, und weiterhin $v(d) = d\psi(d)$ und $\psi(d) = \rho'(d)$ ist.
Als Startschätzer für die Iteration müssen Schätzer mit hohem Bruchpunkt verwendet werden, da sonst die hohen Bruchpunkte der aus der Iteration resultierenden S–Schätzer nicht gewährleistet sind. Daher werden hier die MVE–Schätzer als Startwerte herangezogen. Sie werden approximiert nach dem von Rousseeuw und van Zomeren (1990) vorgestellten Algorithmus, der auf der zufälligen Ziehung von Unterstichproben aus den vorhandenen Beobachtungen beruht.

Der zweite Identifizierer, der in diesem Kapitel untersucht wird, ist eine Modifikation von $\underline{\mathrm{OR}}_{\mathrm{BW}}$. Wie in Rocke (1993, S. 3 ff.) beschrieben, können die Schätzer $\underline{m}_{\mathrm{BW}}$ und S_{BW} noch stark von beliebig schlecht plazierten Beobachtungen beeinflußt werden, obwohl sie den maximal möglichen Finite–sample Bruchpunkt besitzen. Dieses Phänomen tritt verstärkt in höheren Dimensionen auf. Die Ursache liegt in der zur Funktion ρ_{BW} gehörenden Gewichtsfunktion w_{BW}, die nur extrem weit außen liegenden Beobachtungen das Gewicht Null zuweist.
Um diesem Problem zu begegnen, entwickelt Rocke eine Modifikation von ρ_{BW}, die sogenannte Translated-Biweight-Funktion, hier mit ρ_{TW} bezeichnet. Sie hängt im Gegensatz zu ρ_{BW} von zwei Parametern c_0 und M ab. Diese können so gewählt werden, daß einerseits die resultierenden Schätzer einen gegebenen Finite–sample Bruchpunkt besitzen. Gleichzeitig kann erreicht werden, daß die Wahrscheinlichkeit für eine „gute" Beobachtung, das Gewicht Null zu erhalten, einen vorgegebenen Wert β annimmt. Dabei soll β einerseits hinreichend klein sein, um die Information, die die guten Beobachtungen beinhalten, möglichst vollständig für die Schätzung zu nutzen. Andererseits

darf β nicht zu klein gesetzt werden, da sonst wie bei der Biweight-Funktion die Gefahr besteht, daß auch weit außen liegende Punkte noch zu großes Gewicht für die Schätzung bekommen. Rocke schlägt für β einen Wert von 0.01 vor.

Definition 5.2

Der Ausreißer–Identifizierer $\underline{\mathrm{OR}}_{\mathrm{TW}}$ ist definiert als

$$\underline{\mathrm{OR}}_{\mathrm{TW}} := \{\underline{x} \in I\!R^p : (\underline{x} - \underline{m}_{\mathrm{TW}})^T S_{\mathrm{TW}}^{-1}(\underline{x} - \underline{m}_{\mathrm{TW}}) \geq c_{\mathrm{TW}}(p, N, \alpha_N)\}.$$

Dabei sind $\underline{m}_{\mathrm{TW}}$ und S_{TW} Lösungen des in Lemma 5.1 genannten Minimierungsproblems mit der folgenden Funktion $\rho = \rho_{\mathrm{TW}} : I\!R_+ \mapsto I\!R$:

$$\rho_{\mathrm{TW}}(d) = \begin{cases} \dfrac{d^2}{2} & , \quad 0 \leq d \leq M \\ \dfrac{M^2}{2} - \dfrac{M^2(M^4 - 5M^2c_0^2 + 15c_0^4)}{30c_0^4} & \\ \quad + d^2(\dfrac{1}{2} + \dfrac{M^4}{2c_0^4} - \dfrac{M^2}{c_0^2}) + d^3(\dfrac{4M}{3c_0^2} - \dfrac{4M^3}{3c_0^4}) & \\ \quad + d^4(\dfrac{3M^2}{2c_0^4} - \dfrac{1}{2c_0^2}) - d^5\dfrac{4M}{5c_0^4} + d^6\dfrac{1}{6c_0^4} & , \quad M \leq d \leq M + c_0 \\ \dfrac{M^2}{2} + \dfrac{c_0(5c_0 + 16M)}{30} & , \quad d > M + c_0 \end{cases} .$$

Dabei werden $c_0 \in I\!R$ und $M \in I\!R$ so festgelegt, daß die oben genannte Wahrscheinlichkeit β einen Wert von 0.01 annimmt und der Finite–sample Bruchpunkt von S_{TW} maximal ist (vgl. Rocke (1993), S. 6 f.).
Der kritische Wert $c_{\mathrm{TW}}(p, N, \alpha_N)$ wird wie bei $\underline{\mathrm{OR}}_{\mathrm{BW}}$ aus der Normierungsgleichung

$$P(\underline{X}_i \in I\!R^p \backslash \underline{\mathrm{OR}}_{\mathrm{TW}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha$$

für $\underline{X}_1, \ldots, \underline{X}_N$ i.i.d., $\underline{X}_i \sim N(\underline{\mu}, \Sigma)$, gewonnen, wobei α vorgegeben wird und $\alpha_N = 1 - (1 - \alpha)^{1/N}$ gilt.

Das Verfahren zur Bestimmung der beiden Schätzer $\underline{m}_{\mathrm{TW}}$ und S_{TW} ist das gleiche wie für $\underline{m}_{\mathrm{BW}}$ und S_{BW} bereits beschrieben.

Die Stichprobenumfänge der im weiteren betrachteten Beispieldatensätze sind nicht groß genug dafür, daß die im vorigen Kapitel erarbeitete Asymptotik schon greift. Die Normierungskonstanten $c_{\mathrm{BW}}(p, N, \alpha_N)$ und $c_{\mathrm{TW}}(p, N, \alpha_N)$ müssen daher numerisch gewonnen werden. Das dazu benutzte Verfahren wird im folgenden Abschnitt vorgestellt.

5.2 Bestimmung der Normierungskonstanten

Die Normierungskonstante $c(p, N, \alpha_N)$ eines Identifizierers $\underline{\mathrm{OR}}$ wird hier, wie bereits erwähnt, durch den Zusammenhang

$$P(\underline{X}_i \in \mathbb{R}^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha \tag{5.2}$$

bestimmt, wobei die Zufallsvariablen $\underline{X}_i, i = 1, \ldots, N$, stochastisch unabhängig und identisch normalverteilt sind und $\alpha_N = 1 - (1 - \alpha)^{1/N}$ gilt, α gegeben.
Gleichung (5.2) ist äquivalent zu

$$\begin{aligned} & P((\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m}) < c(p, N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha \\ \Leftrightarrow \quad & P(\max_{i=1,\ldots,N} (\underline{X}_i - \underline{m})^T S^{-1} (\underline{X}_i - \underline{m}) < c(p, N, \alpha_N)) = 1 - \alpha. \end{aligned}$$

Zur Berechnung von $c(p, N, \alpha_N)$ werden 100 Stichproben vom Umfang N aus einer p–variaten Standardnormalverteilung gezogen. Aus jeder Stichprobe wird das Maximum der quadrierten Distanzen $(\underline{x}_i - \underline{m})^T S^{-1} (\underline{x}_i - \underline{m})$ bestimmt. Auf diese Weise erhält man 100 Maxima, aus denen das $(1 - \alpha)$–Quantil berechnet wird. Dieses Vorgehen wird hundertmal wiederholt. Die Normierungskonstante berechnet man als arithmetisches Mittel der Quantile aus diesen 100 Wiederholungen.
Die Berechnungen der Normierungskonstanten und der Schätzer für die Beispiele des folgenden Abschnitts wurden auf einem IBM–kompatiblen PC mit einem Prozessor 80486/33 mit Hilfe des statistischen Programmpakets S–Plus für Windows (Version 3.2) durchgeführt. Die Routinen zur Berechnung der verwendeten S–Schätzer wurden freundlicherweise von D. Rocke zur Verfügung gestellt.

5.3 Beispiele

Die beiden in Abschnitt 5.1 vorgestellten Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$ und $\underline{\mathrm{OR}}_{\mathrm{TW}}$ werden im folgenden auf vier Datensätze angewandt, die bereits mehrfach auf das Vorhandensein von Ausreißern untersucht wurden. Es handelt sich dabei um Datensätze mit bis zu vier Variablen, so daß die Anwendung von Identifizierern möglich ist, ohne daß der „Fluch der Dimension“ zum Tragen kommt (vgl. Kapitel 1).

5.3.1 Körper- und Hirngewicht von 28 Tierarten

Der erste betrachtete Datensatz wurde bereits in der Einleitung beschrieben. Von 28 ausgewählten Tierarten wurde das Körper- und Hirngewicht (in Kilogramm bzw. Gramm) ermittelt, vgl. Rousseeuw, Leroy (1987, S. 58). Der Datensatz enthält die jeweils logarithmierten Gewichte. Damit handelt es sich um $N = 28$ Beobachtungen in $p = 2$ Dimensionen. In einer Arbeit von Rousseeuw und van Zomeren (1990) werden insgesamt fünf dieser Beobachtungen als Ausreißer identifiziert. Das dort verwendete Verfahren ähnelt der Anwendung eines Ausreißer–Identifizierers. Es werden für alle Beobachtungen die robusten Mahalanobisdistanzen bezüglich der MVE–Schätzer bestimmt. Als Ausreißer erkannt werden diejenigen Beobachtungen, deren robuste Distanzen größer sind als ein kritischer Wert. Dieser wird allerdings nicht aus einer speziellen Normierungsbedingung für das Identifizierungsverfahren gewonnen, sondern aus dem klassischen Identifizierungsansatz übernommen. Genauer gesagt, benutzen Rousseeuw und van Zomeren den Wert $\sqrt{\chi^2_{2;0.975}}$ als kritischen Wert für die robusten Distanzen. Überträgt man dies in den Zusammenhang der Ausreißer–Identifizierer mit der in Abschnitt 5.1 genannten Normierungsbedingung

$$P(\underline{X}_i \in I\!R^p \backslash \underline{\mathrm{OR}}(\underset{\sim}{X}_N, \alpha_N), i = 1, \ldots, N) = 1 - \alpha,$$

so erhält man $\alpha_N = 0.025$, und damit über den Zusammenhang $\alpha_N = 1 - (1-\alpha)^{1/N}$ einen Wert von $\alpha = 0.507814$. Das bedeutet aber, daß die Wahrscheinlichkeit, bei Vorliegen von normalverteilten Beobachtungen keine von ihnen als Ausreißer zu identifizieren, bei etwa 49% liegt und somit intuitiv viel zu klein ist. Es ist daher zu vermuten,

Tabelle 5.1: Kritische Werte für die verschiedenen Identifizierungsverfahren

$N = 28,\ p = 2$		
	$\alpha = 0.05$	$\alpha = 0.1$
Verfahren	$c(2, 28, 0.0018)$	$c(2, 28, 0.0038)$
$\underline{\text{OR}}_{\text{BW}}$	23.06993	18.27196
$\underline{\text{OR}}_{\text{TW}}$	23.23123	18.73179
$\underline{\text{OR}}_{\text{MD}}$	12.60663	11.16891
	$\chi^2_{2;0.9982}$	$\chi^2_{2;0.9962}$
ROUZO	12.60663	11.16891

daß Rousseeuw und van Zomeren mit ihrem Verfahren zu viele der Beobachtungen als Ausreißer identifizieren. Um einen direkten Vergleich der Identifizierer $\underline{\text{OR}}_{\text{BW}}$ und $\underline{\text{OR}}_{\text{TW}}$ mit dem Verfahren nach Rousseeuw und van Zomeren zu ermöglichen, werden für alle Verfahren die Normierungskonstanten zu Werten von $\alpha = 0.05$ und $\alpha = 0.1$ bestimmt. Es ergeben sich zugehörige Werte von $\alpha_N \approx 0.0018$ und $\alpha_N \approx 0.0038$. Tabelle 5.1 enthält die entsprechenden Normierungskonstanten und die Quantile der χ^2–Verteilung, die für das Verfahren von Rousseeuw und van Zomeren (ROUZO) und für den klassischen Identifizierer $\underline{\text{OR}}_{\text{MD}}$ benötigt werden.

Die Anwendung von $\underline{\text{OR}}_{\text{BW}}$ und $\underline{\text{OR}}_{\text{TW}}$ sowie zum Vergleich auch von $\underline{\text{OR}}_{\text{MD}}$ auf den Datensatz liefert die im folgenden dargestellten Ergebnisse. Die quadrierten Distanzen der Beobachtungen bezüglich der in den Identifizierern verwendeten Schätzer sind in Tabelle 5.2 zusammengefaßt, die Werte für das Verfahren von Rousseeuw und van Zomeren sind deren Arbeit (1990, S. 58) entnommen.

In der Arbeit von Rousseeuw und van Zomeren, wo mit der (unzureichenden) Normierung auf $\alpha \approx 0.51$ gearbeitet wird, werden insgesamt fünf Beobachtungen als Ausreißer identifiziert. Es handelt sich dabei um die Daten von drei Dinosauriern (Beobachtungen Nr. 6, 16, 25), dem Rhesus–Affen (Nr. 17) und dem Menschen (Nr. 14).

Tabelle 5.2: Datensatz „Körpergewicht–Hirngewicht“; quadrierte Distanzen der Beobachtungen bei Verwendung der verschiedenen Identifizierungsverfahren

Nummer der	Quadrierte Distanzen für			
Beobachtung	$\underline{\mathrm{OR}}_{\mathrm{BW}}$	$\underline{\mathrm{OR}}_{\mathrm{TW}}$	ROUZO	$\underline{\mathrm{OR}}_{\mathrm{MD}}$
1	0.42	0.44	0.29	1.05
2	1.17	1.20	0.29	0.50
3	0.09	0.08	0.16	0.09
4	0.34	0.33	0.40	0.15
5	0.66	0.70	0.55	1.36
6	81.92	82.39	46.65	7.25
7	2.16	2.17	2.53	3.03
8	0.29	0.29	0.41	0.52
9	0.50	0.51	0.23	0.76
10	3.82	3.78	2.79	0.66
11	0.79	0.78	0.48	0.49
12	0.48	0.50	0.25	0.79
13	0.23	0.22	0.27	0.48
14	14.64	14.59	11.49	3.07
15	1.50	1.54	1.30	3.22
16	66.71	67.11	37.33	5.82
17	10.58	10.52	7.40	1.55
18	1.00	1.03	0.45	0.04
19	1.66	1.75	1.42	3.57
20	2.35	2.45	1.54	5.33
21	0.29	0.31	0.22	0.72
22	0.19	0.18	0.30	0.18
23	0.42	0.44	0.08	0.07
24	4.48	4.44	3.80	1.13
25	95.97	96.51	52.71	8.79
26	1.26	1.33	1.08	2.61
27	3.26	3.29	1.42	2.60
28	1.65	1.69	0.56	0.16

Wie vermutet, werden bei einer geeigneteren Wahl von α ($\alpha = 0.1$, $\alpha = 0.05$) weniger Beobachtungen als Ausreißer erkannt. Bei einer Wahl von $\alpha = 0.1$ wird Beobachtung Nr. 17 nicht mehr identifiziert, bei $\alpha = 0.05$ wird auch Beobachtung Nr. 14 nicht mehr als Ausreißer erkannt. Die beiden Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$ und $\underline{\mathrm{OR}}_{\mathrm{TW}}$ liefern für beide Normierungen ein einheitliches Ergebnis: lediglich die Daten der drei Dinosaurier werden als Ausreißer identifiziert. Betrachtet man Abbildung 5.1, so ist dieses

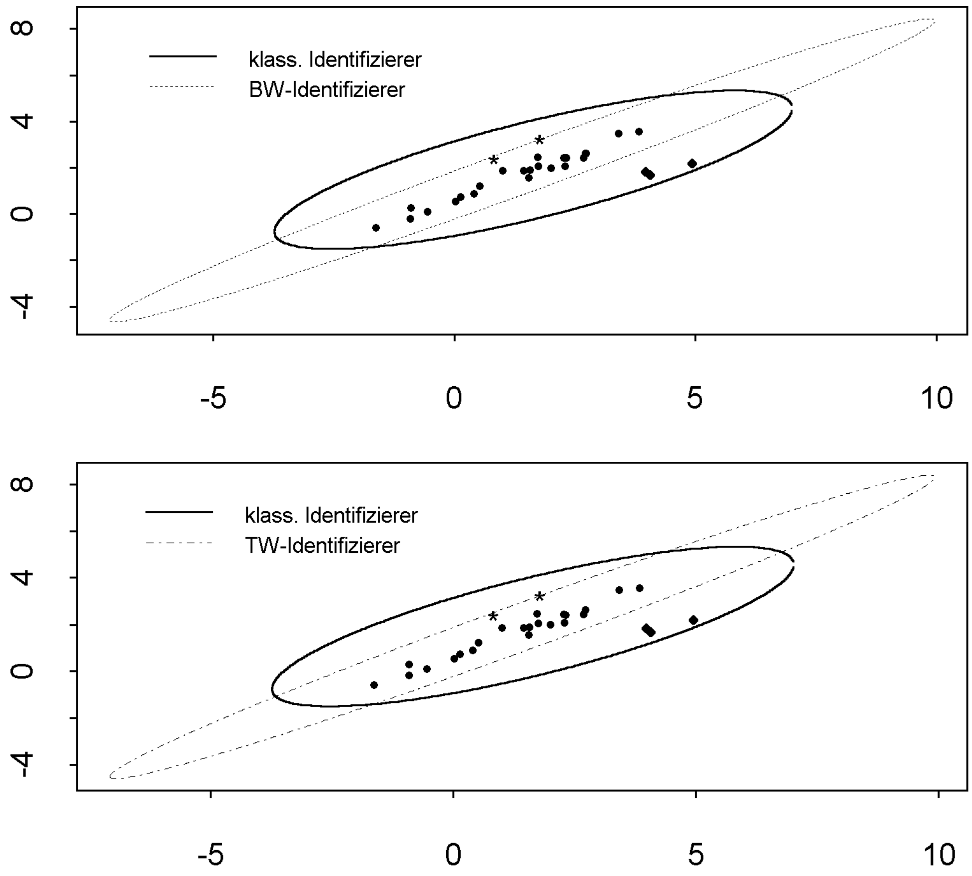


Abbildung 5.1: Logarithmiertes Körper- und Hirngewicht von 28 Spezies; eingetragen sind die Ellipsen der Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$, $\underline{\mathrm{OR}}_{\mathrm{TW}}$ und $\underline{\mathrm{OR}}_{\mathrm{MD}}$, jeweils normiert auf $\alpha = 10\%$

Resultat auch das plausibelste. Dort sind die Daten mit den zu $\underline{\mathrm{OR}}_{\mathrm{BW}}$, $\underline{\mathrm{OR}}_{\mathrm{TW}}$ und $\underline{\mathrm{OR}}_{\mathrm{MD}}$ gehörenden Ellipsen bei einer Normierung auf $\alpha = 0.1$ eingetragen. Die zu den Dinosauriern gehörenden Beobachtungen sind durch Kästchen markiert, die Beobachtungen von Mensch und Rhesus–Affen durch Sterne. Wie man sieht, sind die drei als Kästchen markierten Beobachtungen auch optisch auffällig, während die Sterne weniger den Eindruck vermitteln, aus der Punktwolke herauszufallen. Beide robusten Identifizierer führen hier zu sehr ähnlichen Ellipsen. Die zum Vergleich eingetragene Ellipse des klassischen Identifizierers umfaßt alle Beobachtungen. Daran zeigt sich, daß $\underline{\mathrm{OR}}_{\mathrm{MD}}$ hier völlig dem Masking–Effekt unterliegt und keine einzige Beobachtung als Ausreißer erkennt.

Insgesamt liefern die beiden Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$ und $\underline{\mathrm{OR}}_{\mathrm{TW}}$ bei einer Normierung auf $\alpha = 0.1$ bessere Resultate als die beiden Verfahren ROUZO und $\underline{\mathrm{OR}}_{\mathrm{MD}}$. Es werden genau die auch optisch auffälligen Beobachtungen erkannt.

5.3.2 Kosten für den Transport von Milch

Die Daten für das zweite hier betrachtete Beispiel stammen aus einer Studie über die Kosten für den Transport von Milch von Farmen zu Molkereien (Johnson, Wichern (1982), S. 280 f.). Untersucht wurden die Kosten für 36 benzingetriebene und 23 dieselgetriebene Fahrzeuge. Für die Ausreißer–Identifizierung werden nur die Kosten für die benzingetriebenen Fahrzeuge herangezogen. Der Datensatz besteht daher aus $N = 36$ Beobachtungen in $p = 3$ Variablen, wobei die erhobenen Variablen gegeben sind als Kosten für Treibstoff, Reparaturkosten und Investitionskosten, jeweils in US$ pro gefahrener Meile.
Dieser Datensatz wird in den Arbeiten von Bacon–Shone, Fung (1987) sowie Caroni, Prescott (1992) auf Ausreißer untersucht. Die dazu verwendeten Verfahren, die beide auf der Teststatistik von Wilks (1963) beruhen, wurden bereits in Kapitel 2 beschrieben. Bacon–Shone und Fung verwenden eine graphische Methode, ohne eine konkrete Normierung anzugeben. Bei Caroni und Prescott werden Normierungen benutzt, die Werten von $\alpha = 0.05$ und $\alpha = 0.1$ entsprechen. Beide Arbeiten kommen übereinstimmend zu dem Ergebnis, daß zwei Beobachtungen (Nr. 9, 21) als Ausreißer zu betrachten sind.
Dieses Ergebnis wird bei einer Normierung auf $\alpha = 0.05$ von $\underline{\mathrm{OR}}_{\mathrm{BW}}$ und von $\underline{\mathrm{OR}}_{\mathrm{TW}}$ nicht ganz bestätigt. Beide Identifizierer erkennen lediglich eine Beobachtung (Nr. 9) als Ausreißer. Dasselbe Ergebnis liefert auch der hier zum Vergleich angewendete klassische Identifizierer $\underline{\mathrm{OR}}_{\mathrm{MD}}$ für $\alpha = 0.05$. Die Normierungskonstanten der drei Verfahren für $\alpha = 0.05$ und für $\alpha = 0.1$ sind in Tabelle 5.3 zusammengestellt.

Tabelle 5.3: Kritische Werte für die verschiedenen Identifizierungsverfahren

$N = 36,\ p = 3$		
	$\alpha = 0.05$	$\alpha = 0.1$
Verfahren	$c(3, 36, 0.0014)$	$c(3, 36, 0.0029)$
$\underline{\mathrm{OR}}_{\mathrm{BW}}$	25.49317	21.65683
$\underline{\mathrm{OR}}_{\mathrm{TW}}$	26.56559	22.23688
$\underline{\mathrm{OR}}_{\mathrm{MD}}$	15.51749	13.98736

Tabelle 5.4: Datensatz „Transportkosten für Milch"; quadrierte Distanzen der Beobachtungen bei Verwendung der verschiedenen Identifizierungsverfahren

Nr. der	Quadrierte Distanzen für			Nr. der	Quadrierte Distanzen für		
Beob.	$\underline{\mathrm{OR}}_{\mathrm{BW}}$	$\underline{\mathrm{OR}}_{\mathrm{TW}}$	$\underline{\mathrm{OR}}_{\mathrm{MD}}$	Beob.	$\underline{\mathrm{OR}}_{\mathrm{BW}}$	$\underline{\mathrm{OR}}_{\mathrm{TW}}$	$\underline{\mathrm{OR}}_{\mathrm{MD}}$
1	3.07	3.04	1.21	19	0.40	0.42	0.42
2	2.78	2.74	3.28	20	7.57	7.84	6.53
3	2.18	2.25	2.47	21	22.79	22.87	11.04
4	6.25	6.44	3.36	22	1.33	1.33	0.51
5	0.55	0.59	0.95	23	5.87	5.85	5.24
6	2.06	2.14	1.38	24	2.26	2.27	1.93
7	0.86	0.83	0.52	25	6.73	7.01	6.01
8	2.95	2.87	3.28	26	2.04	2.01	2.35
9	58.79	60.67	18.15	27	2.92	2.89	3.70
10	0.26	0.27	0.12	28	2.42	2.42	1.41
11	1.53	1.57	1.07	29	1.76	1.78	2.47
12	0.73	0.76	1.28	30	1.68	1.67	1.79
13	0.43	0.47	0.71	31	4.61	4.60	3.11
14	0.36	0.36	0.40	32	2.63	2.67	2.49
15	4.57	4.66	4.32	33	2.34	2.32	1.32
16	1.86	1.90	2.36	34	1.37	1.41	1.12
17	1.18	1.27	1.05	35	1.50	1.57	2.14
18	2.94	3.07	3.12	36	11.88	12.32	4.40

Die quadrierten Distanzen der einzelnen Beobachtungen für die drei Verfahren finden sich in Tabelle 5.4, aus der man auch ablesen kann, daß bei einer Wahl von $\alpha = 0.1$ beide robusten Identifizierer die Ergebnisse aus der oben genannten Literatur reproduzieren, während $\underline{\mathrm{OR}}_{\mathrm{MD}}$ auch in diesem Fall die zweite auffällige Beobachtung nicht entdeckt.
Bei Betrachtung des zugrundeliegenden Datensatzes findet man die folgende Interpretation dieses Resultats: Bei Beobachtung 9 handelt es sich um ein Fahrzeug, bei dem die Benzinkosten im Vergleich zu den übrigen Fahrzeugen (ausgenommen Nr. 21) extrem hoch sind. Gleichzeitig bewegen sich die Reparaturkosten im oberen Bereich, während die Investitionskosten sehr niedrig sind. Es könnte sich hier also um ein älteres Fahrzeug handeln, das dementsprechend reparaturanfälliger ist und deutlich mehr Treibstoff verbraucht. Das Fahrzeug mit der Nummer 21 zeichnet sich ebenfalls durch sehr hohe Treibstoff- und Reparaturkosten aus, gleichzeitig findet man hier auch noch Investitionskosten, die relativ hoch sind. Die Interpretation ist ähnlich wie bei Beobachtung 9, wobei hier offenbar höhere Anschaffungskosten anzusetzen sind.

5.3.3 Der Datensatz von Hawkins, Bradu und Kass

Bei den Daten von Hawkins, Bradu und Kass (1984) handelt es sich um einen künstlich erzeugten Datensatz, an dem die Autoren den Vorteil der Verwendung robuster Prozeduren für die multiple Regression illustrieren. Der Datensatz enthält 75 Beobachtungen in 4 Variablen (drei Regressoren, ein Regressand). Die ersten 14 Beobachtungen sind als Ausreißer konstruiert. Rousseeuw und Leroy (1987, S. 266 ff.) behandeln am Beispiel dieser Daten das Problem der Ausreißererkennung in multivariaten Stichproben, indem sie nur die Regressoren betrachten, d. h., sie betrachten $N = 75$ Beobachtungen in $p = 3$ Variablen. Mit dem gleichen Verfahren, wie es in Rousseeuw und van Zomeren (1990) beschrieben wird (vgl. 5.3.1), erkennen sie die ersten 14 Beobachtungen als deutliche Ausreißer („far away outliers“). Ähnlich wie in Beispiel 5.3.1 ist in diesem Fall die von Rousseeuw und Leroy gewählte Normierung der Identifizierungsprozedur unzureichend. Durch die Wahl von $\chi^2_{3;0.95}$ als kritischen Wert für die quadrierten Distanzen liegt implizit eine Normierung auf $\alpha \approx 0.98$ vor (mit $\alpha_N = 0.05$). Korrigiert man dies zu einem sinnvolleren Wert von $\alpha = 0.1$, so ist $\chi^2_{3;0.9986}$ als Vergleichswert zu benutzen. Die ersten 14 Beobachtungen sind allerdings so deutliche Ausreißer, daß sie auch mit dieser Normierung vom Verfahren nach Rousseeuw und van Zomeren klar identifiziert werden.

Tabelle 5.5: Kritische Werte für die verschiedenen Identifizierungsverfahren

$N = 75,\ p = 3$		
	$\alpha = 0.05$	$\alpha = 0.1$
Verfahren	$c(3, 75, 0.00068)$	$c(3, 75, 0.0014)$
$\underline{\mathrm{OR}}_{\mathrm{BW}}$	21.67252	19.08589
$\underline{\mathrm{OR}}_{\mathrm{TW}}$	22.08527	19.21467
$\underline{\mathrm{OR}}_{\mathrm{MD}}$	17.07006	15.54748

Tabelle 5.6: Datensatz von Hawkins, Bradu und Kass; quadrierte Distanzen der Beobachtungen bei Verwendung der verschiedenen Identifizierungsverfahren

Nr. der	Quadrierte Distanzen für			Nr. der	Quadrierte Distanzen für		
Beob.	$\underline{\mathrm{OR}}_{\mathrm{BW}}$	$\underline{\mathrm{OR}}_{\mathrm{TW}}$	$\underline{\mathrm{OR}}_{\mathrm{MD}}$	Beob.	$\underline{\mathrm{OR}}_{\mathrm{BW}}$	$\underline{\mathrm{OR}}_{\mathrm{TW}}$	$\underline{\mathrm{OR}}_{\mathrm{MD}}$
1	524.81	539.17	3.72	16	2.93	3.01	4.69
2	552.77	567.84	3.49	17	2.20	2.28	1.94
3	615.54	632.49	5.43	18	0.40	0.40	0.72
4	654.24	672.16	5.04	19	0.95	0.99	1.34
5	630.94	648.22	4.47	20	2.62	2.68	2.57
6	566.39	581.86	4.67	21	0.69	0.70	1.20
7	570.03	585.55	4.10	22	1.80	1.87	2.43
8	537.67	552.40	3.73	23	0.81	0.83	1.19
9	618.50	635.47	5.00	24	1.00	1.04	0.96
10	579.99	595.95	5.52	25	2.46	2.51	0.65
11	813.21	835.45	6.07	26	1.74	1.80	1.38
12	871.92	896.02	9.79	27	2.42	2.48	2.13
13	825.94	848.23	7.18	28	0.73	0.73	0.76
14	1027.40	1054.09	41.28	29	0.80	0.81	0.34
15	2.46	2.53	3.34	30	2.94	2.99	2.49

Fortsetzung von Tabelle 5.6

Nr. der Beob.	Quadrierte Distanzen für			Nr. der Beob.	Quadrierte Distanzen für		
	$\underline{\mathrm{OR}}_{\mathrm{BW}}$	$\underline{\mathrm{OR}}_{\mathrm{TW}}$	$\underline{\mathrm{OR}}_{\mathrm{MD}}$		$\underline{\mathrm{OR}}_{\mathrm{BW}}$	$\underline{\mathrm{OR}}_{\mathrm{TW}}$	$\underline{\mathrm{OR}}_{\mathrm{MD}}$
31	1.84	1.88	3.43	54	2.13	2.21	2.03
32	1.79	1.86	1.73	55	1.08	1.12	1.53
33	0.95	0.98	0.98	56	1.62	1.67	1.80
34	2.63	2.68	1.40	57	1.15	1.18	0.70
35	2.16	2.22	1.57	58	1.85	1.90	2.00
36	0.83	0.85	0.73	59	1.04	1.06	0.35
37	2.48	2.56	3.40	60	2.88	2.94	3.62
38	1.33	1.36	0.57	61	3.07	3.16	2.84
39	2.06	2.09	1.62	62	2.45	2.50	0.58
40	0.68	0.71	1.25	63	1.90	1.97	1.69
41	2.57	2.63	2.93	64	2.08	2.14	0.96
42	2.18	2.21	3.16	65	1.57	1.60	1.34
43	2.72	2.82	3.54	66	1.25	1.29	1.70
44	2.67	2.72	2.04	67	0.16	0.17	0.40
45	2.16	2.20	1.17	68	2.85	2.90	2.43
46	2.30	2.33	1.83	69	1.88	1.93	1.16
47	3.25	3.33	3.92	70	1.18	1.21	1.01
48	2.18	2.24	2.06	71	0.60	0.62	0.42
49	1.55	1.59	2.50	72	0.50	0.52	1.12
50	1.31	1.35	0.18	73	1.14	1.16	2.20
51	1.41	1.45	1.72	74	1.38	1.43	2.75
52	2.74	2.80	4.37	75	2.85	2.90	3.66
53	4.21	4.28	4.95				

Auch die Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$ und $\underline{\mathrm{OR}}_{\mathrm{TW}}$ finden 14 Ausreißer, und zwar sowohl bei Normierung auf $\alpha = 0.1$ als auch auf $\alpha = 0.05$. Die Normierungskonstanten der Identifizierer sind in Tabelle 5.5 zusammengefaßt, die quadrierten Distanzen der Beobachtungen in Tabelle 5.6. In dieser Tabelle sieht man auch, daß der klassische Identifizierer für beide Normierungen nur einen einzigen Ausreißer identifizieren kann. Auch hier zeigt sich der Masking–Effekt.

5.3.4 Oxidation von Ammoniak zu Salpetersäure

Auch der vierte untersuchte Datensatz, der unter dem Namen „Stackloss Data“ bekannt ist, wurde bisher meist im Zusammenhang der Ausreißererkennung bei multipler Regression betrachtet. Es handelt sich um eine Erhebung über den Betrieb einer Apparatur, die Ammoniak zu Salpetersäure oxidiert (vgl. Brownlee (1965), S. 454). Die aus dem in die Anlage einströmenden Ammoniak gewonnenen Stickoxide werden in einer Gegenstrom–Absorptionssäule absorbiert. Die vier erhobenen Variablen sind die Strömungsgeschwindigkeit des Ammoniaks, die Ausgangstemperatur des in der Absorptionssäule zirkulierenden Kühlwassers, die Konzentration der Säure sowie der Anteil des Ammoniaks, der unabsorbiert entweicht. Beim Regressionsansatz soll dieser Anteil durch die drei anderen Variablen erklärt werden. Man kann aber die Beobachtungen auch als unstrukturierten multivariaten Datensatz auffassen. In diesem Fall wird eine Stichprobe aus $N = 21$ Beobachtungen in $p = 4$ Variablen auf Ausreißer untersucht.
Die Anwendung der zwei robusten und des klassischen Identifizierers führen zu den im folgenden dargestellten Ergebnissen. Die Normierungskonstanten der Identifizierer und die quadrierten Distanzen der Beobachtungen sind den folgenden Tabellen 5.7 und 5.8 zu entnehmen.

Tabelle 5.7: Kritische Werte für die verschiedenen Identifizierungsverfahren

$N = 21,\ p = 4$		
	$\alpha = 0.05$	$\alpha = 0.1$
Verfahren	$c(4, 21, 0.0024)$	$c(4, 21, 0.0050)$
$\underline{\text{OR}}_{\text{BW}}$	40.52733	31.57103
$\underline{\text{OR}}_{\text{TW}}$	48.56629	38.89749
$\underline{\text{OR}}_{\text{MD}}$	16.47884	14.85817

Bei einer Normierung auf $\alpha = 0.05$ identifiziert $\underline{\text{OR}}_{\text{BW}}$ zwei Ausreißer (Beobachtungen Nr. 1, 3), wogegen der modifizierte Identifizierer $\underline{\text{OR}}_{\text{TW}}$ nur eine Beobachtung (Nr. 1) erkennt. Das klassische Verfahren versagt wiederum und identifiziert keinen Ausreißer. Auch bei einer großzügigeren Normierung mit $\alpha = 0.1$ kann $\underline{\text{OR}}_{\text{MD}}$ keinen Ausreißer finden, während in diesem Fall $\underline{\mathrm{OR}}_{\mathrm{TW}}$ drei (Nr. 1, 3, 4) und $\underline{\mathrm{OR}}_{\mathrm{BW}}$ sogar vier (Nr. 1, 3, 4, 21) auffällige Beobachtungen ausmacht.

Die meisten Artikel, die sich mit diesem oft analysierten Datensatz befassen, stimmen darin überein, daß die Beobachtungen 3, 4 und 21 als Ausreißer zu betrachten sind. Dagegen besteht keine Einigkeit bezüglich der Beobachtungen 1 und 2. So identifizieren beispielsweise Daniel, Wood (1971), Li (1985) und Andrews (1974) die auch von $\underline{\mathrm{OR}}_{\mathrm{BW}}$ entdeckten vier Beobachtungen (Nr. 1, 3, 4, 21) als Ausreißer, während Carroll und Ruppert (1985) statt dessen die Beobachtungen 2, 3, 4, 21 als verdächtig einstufen. In den Arbeiten von Dempster, Gasko–Green (1981) und Andrews, Pregibon (1978) werden sogar alle fünf (Nr. 1, 2, 3, 4, 21) als Ausreißer deklariert. Die Identifizierung der Beobachtung Nr. 2 als Ausreißer ist nach den Ergebnissen der beiden hier verwendeten robusten Prozeduren $\underline{\mathrm{OR}}_{\mathrm{BW}}$ und $\underline{\mathrm{OR}}_{\mathrm{TW}}$ nicht gerechtfertigt. Zwar fällt diese Beobachtung hier ebenfalls auf, weil sie eine deutlich höhere Distanz aufweist als die restlichen, nicht als Ausreißer identifizierten. Allerdings bleibt die Größe dieser Distanz klar unter den kritischen Werten der hier verwendeten Verfahren.

Tabelle 5.8: „Stackloss Data“; quadrierte Distanzen der Beobachtungen bei Verwendung der verschiedenen Identifizierungsverfahren

Nr. der	Quadrierte Distanzen für			Nr. der	Quadrierte Distanzen für		
Beob.	$\underline{\mathrm{OR}}_{\mathrm{BW}}$	$\underline{\mathrm{OR}}_{\mathrm{TW}}$	$\underline{\mathrm{OR}}_{\mathrm{MD}}$	Beob.	$\underline{\mathrm{OR}}_{\mathrm{BW}}$	$\underline{\mathrm{OR}}_{\mathrm{TW}}$	$\underline{\mathrm{OR}}_{\mathrm{MD}}$
1	51.22	58.45	6.56	11	1.44	1.68	3.07
2	22.30	26.21	6.11	12	1.78	2.08	4.47
3	41.44	47.03	5.10	13	3.51	3.71	2.55
4	38.49	42.39	5.51	14	2.36	2.60	3.32
5	0.78	0.94	0.44	15	1.41	2.78	3.65
6	1.16	1.40	1.69	16	1.55	1.81	1.85
7	1.52	1.84	4.27	17	2.79	3.14	7.93
8	1.98	2.30	3.83	18	1.03	1.21	2.40
9	1.13	1.38	3.10	19	1.30	1.53	2.71
10	2.03	2.32	3.39	20	2.35	2.63	0.92
				21	32.19	34.91	11.13

Ein deutlicher Unterschied zu den Ergebnissen der oben angegebenen Arbeiten besteht in der Bedeutung der Beobachtung Nr. 21. Alle genannten Autoren stimmen darin überein, daß diese Beobachtung als stärkster Ausreißer anzusehen ist. Betrachtet man dagegen die in Tabelle 5.8 eingetragenen Distanzen, so erkennt man, daß mit den hier verwendeten robusten Identifizierern diese Aussage nicht bestätigt werden kann. Gleichzeitig ist aber bei Verwendung des klassischen Identifizierers $\underline{\mathrm{OR}}_{\mathrm{MD}}$ die Distanz von Beobachtung Nr. 21 deutlich größer als die aller anderen Beobachtungen. Damit zeigt $\underline{\mathrm{OR}}_{\mathrm{MD}}$ tendenziell ein ähnliches Verhalten wie die Prozeduren in den oben genannten Arbeiten, auch wenn die Größe der Distanz in diesem Fall nicht ausreicht, um Beobachtung 21 als Ausreißer zu identifizieren. Da gleichzeitig bei Verwendung von $\underline{\mathrm{OR}}_{\mathrm{MD}}$ die Beobachtungen 1, 3 und 4 keine auffällig großen Distanzen aufweisen, liegt der Schluß nahe, daß die starken Ausreißer (Nr. 1, 3, 4) sich hier nicht nur gegenseitig maskieren, sondern auch dafür sorgen, daß die weniger abweichende Beobachtung (Nr. 21) stärker auffällt. Aus der Ähnlichkeit des Verhaltens von $\underline{\mathrm{OR}}_{\mathrm{MD}}$ und den Prozeduren in den oben genannten Arbeiten läßt sich der Schluß ziehen, daß auch die dort verwendeten Verfahren noch immer dem Einfluß der starken Ausreißer unterliegen, wenn sie Beobachtung Nr. 21 als größten Ausreißer ausmachen. Der hier verwendete robuste Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$ dagegen reagiert weniger auf den Einfluß der Beobachtungen 1, 3 und 4 und identifiziert Beobachtung 21 nicht mehr als den stärksten Ausreißer.

Die Interpretation des Ergebnisses ist hier nicht so unmittelbar zugänglich wie in den vorangegangenen Beispielen. Betrachtet man die zugrundeliegenden Daten, so zeigen sich die Beobachtungen 1 und 3 ähnlich. Bei beiden findet man hohe Strömungsgeschwindigkeit und hohe Kühlwassertemperatur bei gleichzeitig großer Menge an Ammoniak, der unabsorbiert entweicht. Beobachtung 2, die ja hier auch auffällt, ohne jedoch identifiziert zu werden, ist von ähnlicher Qualität. Die als Ausreißer identifizierte Beobachtung mit der Nummer 21 ist dagegen anders gelagert. Hier sieht man bei ebenfalls eher hoher Strömungsgeschwindigkeit eine niedrige Kühlwassertemperatur und eine geringe Menge nicht absorbierten Ammoniaks, wobei die Abweichungen vom Rest der Daten weniger gravierend erscheinen als bei den Beobachtungen 1 und 3. Dies bestätigt auch das Ergebnis der Identifizierungsprozedur $\underline{\mathrm{OR}}_{\mathrm{BW}}$. Bei Beobachtung Nummer 4 handelt es sich um einen Ausreißer, der in keiner der einzelnen Variablen auffällig ist. Lediglich in Kombination aller vier betrachteten Variablen ist diese Beobachtung als Ausreißer zu klassifizieren.

5.4 Diskussion der Ergebnisse

Insgesamt kann man als Resultat der in den vorigen Abschnitten betrachteten Beispiele festhalten, daß sich bei einer Normierung der robusten Identifizierer auf $\alpha = 0.1$ die aus der Literatur bekannten Ergebnisse im wesentlichen reproduzieren und in einigen Fällen verbessern lassen. Eine Wahl der Normierungskonstanten für einen Wert von $\alpha = 0.1$ entspricht auch der Idee des Einsatzes von Identifizierern. Das Ziel besteht nicht ausschließlich darin, extrem von der Hauptmasse der Daten abweichende Beobachtungen zu finden, sondern ebenso darin, Ungewöhnliches zu entdecken. Bei einer zu restriktiven Wahl von α werden aber nur ganz extreme Auffälligkeiten gefunden, während weniger offenkundige Abweichungen nicht aufgedeckt werden können.

Die hier betrachteten Beispiele zeigen, daß die beiden verwendeten robusten Identifizierer bei geeigneter Normierung gute Ergebnisse liefern. Der auf Tukeys Biweight–Funktion basierende Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$ schneidet bei den behandelten Beispielen insgesamt etwas besser ab als seine Modifikation $\underline{\mathrm{OR}}_{\mathrm{TW}}$. Das liegt speziell an den Ergebnissen aus Abschnitt 5.3.4. Bei den „Stackloss Data“ erkennt $\underline{\mathrm{OR}}_{\mathrm{TW}}$ selbst bei der als angemessen angesehenen Normierung auf $\alpha = 0.1$ weniger Ausreißer als $\underline{\mathrm{OR}}_{\mathrm{BW}}$. Dieses Resultat erscheint zunächst widersprüchlich. Einerseits wurden die für $\underline{\mathrm{OR}}_{\mathrm{TW}}$ verwendeten Schätzer gerade für höhere Dimensionen konstruiert, andererseits ist $\underline{\mathrm{OR}}_{\mathrm{TW}}$ bei den zwei- und dreidimensionalen Beispieldatensätzen genauso gut wie $\underline{\mathrm{OR}}_{\mathrm{BW}}$, wird aber bei dem vierdimensionalen Datensatz schlechter.
Bei genauerem Hinsehen ist das Ergebnis mit der Arbeitsweise der TW–Schätzer erklärbar. Damit für die Schätzung weit außen liegende Beobachtungen kein Gewicht erhalten, wird in Kauf genommen, daß auch „gute“ Datenpunkte mit einer gewissen Wahrscheinlichkeit mit Gewicht Null in die Schätzung eingehen. Dies ist sinnvoll und funktioniert gut, wenn Ausreißer in den Daten vorhanden sind. Für die Bestim-

mung der Normierungskonstanten des Identifizierers werden aber Stichproben aus dem Nullmodell herangezogen, so daß diese fast keine Ausreißer enthalten. Bestimmt man aus diesen Stichproben die TW–Schätzer, so wird insbesondere die Kovarianzmatrix unterschätzt, weil „zu viele" Punkte mit Gewicht Null in die Schätzung einbezogen werden. Dadurch muß die Normierungskonstante größer werden, um die korrekte Normierung zu gewährleisten. Dieser Effekt wird verstärkt durch einen relativ kleinen Stichprobenumfang, wie es ja bei den „Stackloss Data" mit $N = 21$ für $p = 4$ Dimensionen der Fall ist. Insgesamt wird die Normierungskonstante so groß, daß bei Anwendung des Identifizierers auf Daten, die Ausreißer enthalten, das Ellipsoid $I\!R^p \backslash \underline{\mathrm{OR}}_{\mathrm{TW}}$ so stark aufgebläht wird, daß mäßig extreme Aureißer nicht mehr erkannt werden. Daher läßt sich die Empfehlung formulieren, den Identifizierer $\underline{\mathrm{OR}}_{\mathrm{TW}}$ in höheren Dimensionen nicht zu verwenden, wenn der Stichprobenumfang relativ gering ist.
Diese sich aus den Beispielen ergebenden Beobachtungen motivieren die Untersuchung der Frage, wie „groß" Ausreißer werden können, bis sie von den Identifizierungsprozeduren entdeckt werden.

5.5 Der größte nicht entdeckte Ausreißer

Bei multivariaten Datensätzen kann man nicht ohne weiteres von der Größe einer Beobachtung sprechen. Bevor man untersuchen kann, wie groß Ausreißer werden können, ohne identifiziert zu werden, muß der Begriff der Größe präzisiert werden. Da ohne Beschränkung der Allgemeinheit vorausgesetzt werden kann, daß die Beobachtungen unter dem Nullmodell Realisationen $N(\underline{0}, \mathcal{I})$-verteilter Zufallsgrößen sind, läßt sich die Größe einer Beobachtung mit ihrem euklidischen Abstand vom Ursprung gleichsetzen. Ziel der folgenden Untersuchung ist es daher herauszufinden, welchen Abstand vom Ursprung Ausreißer noch haben können, ohne daß sie von den in den vorigen Abschnitten behandelten Prozeduren identifiziert werden. Dabei sei vorausgesetzt, daß die Anzahl von Ausreißern in einem Datensatz unterhalb der Zahl bleibt, mit der ein Zusammenbruch der Identifizierungsprozedur erreicht werden kann.
Erster Schritt hierzu ist die Überlegung, welche Datenkonstellationen für die Identifizierer besonders ungünstig sind, das heißt, durch welche Plazierung der Ausreißer

erreicht werden kann, daß auch große Ausreißer nicht erkannt werden.

Betrachtet man beispielsweise den Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$, so kann man durch eine andere Darstellung der Nebenbedingung des zugrundeliegenden Minimierungsproblems einige Aussagen treffen. Die Schätzer für Lage und Kovarianz, die in $\underline{\mathrm{OR}}_{\mathrm{BW}}$ eingehen, sind nach Lemma 5.1 und Definition 5.1 gegeben als Lösung des Problems

$$\min_{S \in \mathrm{PDS}(p)} \det(S)$$

unter der Nebenbedingung

$$\frac{1}{N}\sum_{i=1}^{N} \rho\left(\sqrt{(\underline{X}_i - \underline{m})^T S^{-1}(\underline{X}_i - \underline{m})}\right) \;=\; b_0$$

mit

$$\rho(d) = \begin{cases} \dfrac{d^2}{2} - \dfrac{d^4}{2c_0^2} + \dfrac{d^6}{6c_0^4} & , \quad 0 \le d \le c_0 \\ \dfrac{c_0^2}{6} & , \quad d > c_0 \end{cases} .$$

Unter Beachtung der Zusammenhänge $b_0 = \mathrm{E}(\rho(D))$ und $\frac{\left[\frac{N-p+1}{2}\right]}{N}\rho(c_0) = \mathrm{E}(\rho(D))$ $(D^2 \sim \chi_p^2)$ sowie $\rho(c_0) = \dfrac{c_0^2}{6}$ läßt sich die Nebenbedingung unmittelbar umformulieren in

$$\sum_{i=1}^{N}\left(1 - \frac{d_i^2}{c_0^2}\right)^3 1\!\mathrm{I}_{\{\underline{X}_i \in E\}} \;=\; \left[\frac{N+p}{2}\right].$$

Dabei ist $d_i^2 = (\underline{X}_i - \underline{m})^T S^{-1}(\underline{X}_i - \underline{m})$, $i = 1, \ldots, N$, und
$E = \{\underline{x} \in I\!R^p : (\underline{x} - \underline{m})^T S^{-1}(\underline{x} - \underline{m}) \le c_0^2\}$, $1\!\mathrm{I}$ bezeichnet die Indikatorfunktion.

Damit gehen alle Beobachtungen, die innerhalb des Ellipsoids E liegen, mit positivem Gewicht in die Schätzung von $\underline{\mu}$ und Σ ein.

Aus der obigen Schreibweise der Nebenbedingung folgt unmittelbar:

- Es werden insgesamt mindestens $\left[\dfrac{N+p}{2}\right] + 1$ der Beobachtungen zur Bestimmung von $\underline{m}$ und S herangezogen.

- Damit enthält die Menge E mindestens $\left[\dfrac{N+p}{2}\right] + 1 - k$ reguläre Beobachtungen; dabei ist k die Anzahl der Ausreißer im Datensatz. Unter der Voraussetzung, daß $k < \left[\dfrac{N-p+1}{2}\right]$, so daß die Schätzer $\underline{m}$ und S nicht zusammenbrechen, gehen also mindestens $p+1$ reguläre Beobachtungen in die Schätzung ein.

Eine besonders ungünstige Datenkonstellation erhält man also dann, wenn es gelingt, die Bestimmung von $\underline{m}$ und S auf allen nichtregulären und möglichst wenigen regulären Beobachtungen beruhen zu lassen. Das heißt, das Ellipsoid E muß alle Ausreißer enthalten, dazu mindestens $p+1$ reguläre Beobachtungen, und es muß minimales Volumen besitzen unter allen möglichen Ellipsoiden, die die obige Nebenbedingung erfüllen. Plaziert man alle nichtregulären Beobachtungen auf denselben Punkt in einer gewissen Entfernung vom Erwartungswert $\underline{\mu}$ der regulären Variablen, so läßt sich genau dies erreichen. Das Ellipsoid E, innerhalb dessen die zur Schätzung von $\underline{\mu}$ und Σ herangezogenen Beobachtungen liegen, zeichnet sich in diesem Fall in der Regel durch eine extrem kleine kürzeste Hauptachse aus, wodurch das Volumen entsprechend klein bleibt.

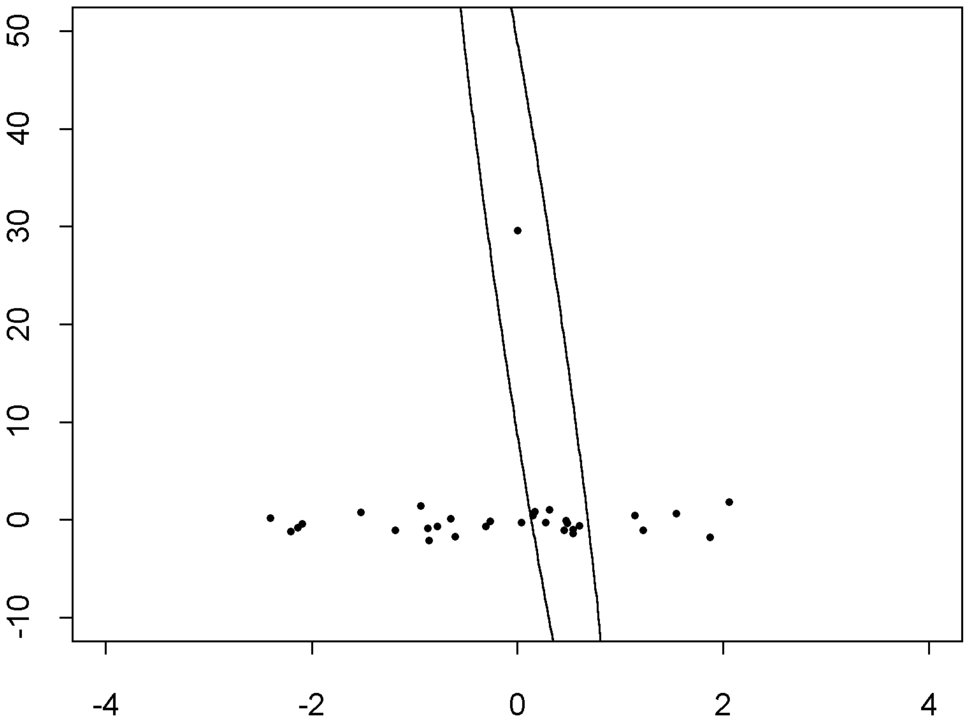


Abbildung 5.2: Datensatz mit $N=50$ Beobachtungen, davon $k=20$ Ausreißer im Punkt $(0,30)$; die innerhalb der Ellipse liegenden Beobachtungen werden zur Bestimmung von $\underline{m}$ und S für den Identifizierer $\underline{\mathrm{OR}}_{\mathrm{BW}}$ herangezogen

Abbildung 5.2 zeigt ein Beispiel für eine solche Situation: in einer Stichprobe vom Umfang $N = 50$ befinden sich $k = 20$ nichtreguläre Beobachtungen, die nach der oben beschriebenen Idee alle auf den Punkt $(0,30)$ plaziert wurden. Die regulären

Beobachtungen entstammen hier einer bivariaten Standardnormalverteilung. Die eingezeichnete Ellipse umfaßt gerade diejenigen Beobachtungen aus dem Datensatz, die zur Bestimmung der beiden Schätzer $\underline{m}$ und S für den Ausreißer–Identifizierer $\underline{\text{OR}}_{\text{BW}}$ benutzt werden. Wie man sieht, enthält die Ellipse den Punkt $(0, 30)$, das heißt, alle Ausreißer gehen in die Schätzung ein.

Nach den vorangegangenen Überlegungen läßt sich die Größe des größten nicht identifizierten Ausreißers anhand von Simulationen schätzen. Eine vorgegebene Anzahl k von nichtregulären Beobachtungen wird auf einen Punkt gelegt, und es wird bestimmt, wie weit dieser Punkt vom Ursprung entfernt werden kann, so daß die Identifizierungsprozedur die Ausreißer noch nicht erkennt. Die folgende Tabelle stellt die Ergebnisse einer Simulationsstudie dar, in der die Größen der größten nicht identifizierten Ausreißer für verschiedene Stichprobenumfänge N, Dimensionen p und Anzahlen k von nichtregulären Beobachtungen zusammengetragen sind. Für jede Kombination von N, p und k wurde in die Tabelle jeweils das arithmetische Mittel der Größen von 1000 Simulationsläufen eingetragen. Beide Identifizierungsprozeduren wurden entsprechend der Überlegungen von Abschnitt 5.4 auf einen Wert von $\alpha = 0.1$ gemäß Beziehung (5.2) normiert.

Tabelle 5.9: Mittlere Entfernung des größten nicht identifizierten Ausreißers vom Ursprung für die Identifizierer $\underline{\text{OR}}_{\text{BW}}$ und $\underline{\text{OR}}_{\text{TW}}$

$\underline{\text{OR}}_{\text{BW}}$	$N = 20$			$N = 50$		
k	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
1	4.44	4.80	5.25	4.03	4.33	4.62
5	7.96	16.16	36.21	4.41	4.74	5.07
7	50.42	65.81	233.38	/	/	/
21	/	/	/	76.61	294.26	781.39

$\underline{\text{OR}}_{\text{TW}}$	$N = 20$			$N = 50$		
k	$p = 2$	$p = 3$	$p = 4$	$p = 2$	$p = 3$	$p = 4$
1	4.37	5.02	5.67	3.99	4.38	4.69
5	7.92	15.74	41.36	4.40	4.75	5.24
7	58.67	79.23	580.33	/	/	/
21	/	/	/	88.76	449.95	1186.83

Man sieht, daß sich beide in diesem Kapitel betrachteten Identifizierungsprozeduren für kleine Anteile an Ausreißern ungefähr gleich verhalten. Größere Ausreißer werden durch beide Prozeduren identifiziert. Bei größeren Anteilen an nichtregulären Beobachtungen zeigen beide Identifizierungsverfahren ein weniger gutes Verhalten. Obwohl der maximale asymptotische Bias und damit auch die Entfernung des größten nicht identifizierten Ausreißers in beiden Fällen beschränkt ist, können auch ziemlich weit entfernte Beobachtungen nicht als Ausreißer identifiziert werden. Dabei werden durch die Simulation die Schlußfolgerungen aus den in den vorigen Abschnitten betrachteten Beispielen bestätigt: der modifizierte Identifizierer $\underline{\mathrm{OR}}_{\mathrm{TW}}$, der auf der von Rocke (1993) vorgeschlagenen geänderten Biweight–Funktion beruht, schneidet für große Anteile an Ausreißern durchweg schlechter ab als $\underline{\mathrm{OR}}_{\mathrm{BW}}$, das heißt, er erkennt im Mittel noch größere Beobachtungen nicht als Ausreißer. Dieser Unterschied verstärkt sich deutlich mit wachsender Dimension bei gleichem Stichprobenumfang. Wie bereits in Abschnitt 5.4 diskutiert, eignet sich $\underline{\mathrm{OR}}_{\mathrm{TW}}$ offenbar weniger zur Anwendung in Datensätzen mit relativ wenigen Beobachtungen im Vergleich zur Anzahl der Dimensionen. In solchen Fällen sollte daher auf $\underline{\mathrm{OR}}_{\mathrm{BW}}$ zurückgegriffen werden.

6 Ausblick

Die vorliegende Arbeit befaßte sich mit der Thematik der Identifizierung von Ausreißern in multivariaten Datensätzen. Es wurde dabei die Annahme getroffen, daß der größte Teil der Beobachtungen in einer zu untersuchenden Stichprobe aus einer gemeinsamen multivariaten Normalverteilung stammt. Unter dieser Voraussetzung können sogenannte Ausreißer–Identifizierer definiert werden, die die Position einer Beobachtung in einem Datensatz als Kriterium dafür heranziehen, ob sie als Ausreißer zu deklarieren ist.

Es wurden verschiedene Beurteilungsmöglichkeiten, wie Bruchpunkt- und Biaskriterien sowie die Größe des größten nicht erkannten Ausreißers, für solche Identifizierungsprozeduren untersucht. Zwei in dieser Arbeit vorgestellte Verfahren zeigten gutes Verhalten bezüglich der hier entwickelten Kriterien. Wie weit Verbesserungen dieser beiden Prozeduren möglich sind, bleibt Gegenstand weiterer Forschungen.
Darüber hinaus beschreiben die hier erarbeiteten Kriterien das Verhalten von Identifizierern nicht vollständig. Weitere Instrumente zur Beurteilung können in Analogie zu solchen konstruiert werden, wie sie bereits für robuste Lokations- und Kovarianzschätzer bekannt sind. Dazu gehört beispielsweise die Influenzfunktion, die den Einfluß einer geringfügigen „Verunreinigung" einer Stichprobe in einem Punkt auf die Schätzer beschreibt (vgl. Hampel et al. (1986), Radhakrishnan, Kshirsagar (1981)). Auch eine Verallgemeinerung des Begriffs der asymptotischen Varianz (Hampel et al. (1986)) auf Ausreißer–Identifizierer ist denkbar. Damit kann die Effizienz eines Identifizierers beurteilt werden. In diesem Zusammenhang bleibt noch zu untersuchen, ob hohe Robustheit und Effizienz für Identifizierer vereinbar sind oder ob sich das Kernproblem der robusten Schätzer, daß starke Robustheit zu Effizienzverlust führt, auch hier fortsetzt. Sollte dies der Fall sein, so muß für das Ziel der Identifizierung von Ausreißern die Effizienz geopfert werden – mit den Worten von Edgeworth (1887, S. 269), die er in einem anderen Zusammenhang benutzt – „[...] a sort of sacrifice which has often to be made by those who sail upon the stormy seas of Probability".

Symbolverzeichnis

$\underline{X}$, $\underline{X}_i$	$\mathbb{R}^p$–Zufallsvektoren	S. 12
$N(\underline{\mu}, \Sigma)$	multivariate Normalverteilung mit Erwartungswertvektor $\underline{\mu} \in \mathbb{R}^p$ und Kovarianzmatrix $\Sigma \in \mathbb{R}^{p \times p}$	S. 12
$\underline{x}$, $\underline{x}_i$	Vektoren des $\mathbb{R}^p$, auch Realisationen von $\mathbb{R}^p$–Zufallsvektoren	S. 13
$\chi^2_{p;1-\alpha}$	$(1-\alpha)$–Quantil der χ^2_p–Verteilung	S. 20
$\mathrm{out}(\alpha, \underline{\mu}, \Sigma)$	α–Ausreißer–Bereich der $N(\underline{\mu}, \Sigma)$–Verteilung	S. 23
$\underset{\sim}{x}_N = (\underline{x}_1, \ldots, \underline{x}_N)$	Stichprobe vom Umfang N	S. 24
$\underline{\mathrm{OR}}(\underset{\sim}{x}_N, \alpha_N)$	multivariater α_N–Ausreißer–Identifizierer; auch Teilmenge des $\mathbb{R}^p$, die als Ausreißer identifizierte Punkte enthält	S. 25
$\underline{x}_i^r$	reguläre Beobachtung, Realisation eines nach $N(\underline{\mu}, \Sigma)$ verteilten Zufallsvektors	S. 28
ε^*	Finite–sample Bruchpunkt	S. 29
$\lVert . \rVert$	euklidische Norm des $\mathbb{R}^p$	S. 29
$\lambda_i(A)$	Eigenwert einer Matrix $A \in \mathbb{R}^{p \times p}$ mit $\lambda_1 \geq \ldots \geq \lambda_p$	S. 29
ε_*	additiver Finite–sample Bruchpunkt	S. 30
$\underset{\sim}{x}_n^r = (\underline{x}_1^r, \ldots, \underline{x}_n^r)$	Stichprobe vom Umfang n aus regulären Beobachtungen	S. 31
β^M	Masking–Punkt	S. 31
$\underset{\sim}{x}_k^0 = (\underline{x}_1^0, \ldots, \underline{x}_k^0)$	Stichprobe vom Umfang k aus δ–Ausreißern	S. 31
$\varepsilon^M(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta)$	Masking–Bruchpunkt eines Identifizierers $\underline{\mathrm{OR}}$	S. 31
β^S	Swamping–Punkt	S. 32
$\varepsilon^S(\underline{\mathrm{OR}}, \boldsymbol{\alpha}, \underset{\sim}{x}_n^r, \delta)$	Swamping–Bruchpunkt eines Identifizierers $\underline{\mathrm{OR}}$	S. 32
$\mathrm{vol}(.)$	Volumen eines Körpers	S. 34
$[x]$	Gaußklammer von x, $x \in \mathbb{R}$	S. 38
$\det(.)$	Determinante einer Matrix	S. 41
$\mathcal{I}$	Einheitsmatrix des $\mathbb{R}^{p \times p}$	S. 43
$\mathrm{PDS}(p)$	Menge aller positiv definiten, symmetrischen Matrizen des $\mathbb{R}^{p \times p}$	S. 50
$\underline{\mathrm{OR}}_{\mathrm{S_{MB}}}$	Ausreißer–Identifizierer, basierend auf $\mathrm{S_{MB}}$–Schätzern	S. 51

$\underline{\mathrm{OR}}_{\mathrm{MVE}}$	Ausreißer–Identifizierer, basierend auf MVE–Schätzern	S. 51
$\mathrm{tr}(.)$	Spur einer Matrix	S. 55
$B(\underline{\mathrm{OR}}, \eta, \boldsymbol{\delta})$	maximaler asymptotischer Bias eines Identifizierers $\underline{\mathrm{OR}}$	S. 61
$b(\underline{m}, \eta, \boldsymbol{\delta})$	maximaler asymptotischer Bias eines Lokations-	
	schätzers $\underline{m}$	S. 62
$\Vert.\Vert_2$	Spektralnorm des $\mathbb{R}^{p \times p}$	S. 62
$b(S, \eta, \boldsymbol{\delta})$	maximaler asymptotischer Bias eines Kovarianz-	
	schätzers S	S. 63
$\mathcal{O}$	Landau'sches $\mathcal{O}$–Symbol	S. 67
$\underline{\mathrm{OR}}_{\mathrm{MD}}$	klassischer Ausreißer–Identifizierer, basierend auf $\overline{\underline{x}}_N$	
	und S_N	S. 80
$\overline{\underline{x}}_N$	arithmetisches Mittel einer Stichprobe vom Umfang N	S. 80
S_N	Stichprobenkovarianzmatrix einer Stichprobe vom	
	Umfang N	S. 80
$\underline{\mathrm{OR}}_{\mathrm{BW}}$	robuster Ausreißer–Identifizierer, basierend auf	
	S–Schätzern unter Verwendung von Tukey's Biweight	S. 84
$\underline{\mathrm{OR}}_{\mathrm{TW}}$	Modifikation des Identifizierers $\underline{\mathrm{OR}}_{\mathrm{BW}}$	S. 87

Literatur

Allison, T., Cicchetti, D. V. (1976), Sleep in Mammals: Ecological and Constitutional Correlates, *Science*, **194**, 732–734.

Anderson, E. (1960), A Semi–Graphical Method for the Analysis of Complex Problems, *Technometrics*, **2**, 287–292.

Andrews, D. F. (1972), Plots of High-Dimensional Data, *Biometrics*, **28**, 126–136.

Andrews, D. F. (1974), A Robust Method for Multiple Linear Regression, *Technometrics*, **16**, 523–531.

Andrews, D. F., Pregibon, D. (1978), Finding the Outliers that Matter, *Journal of the Royal Statistical Society, Ser. B*, **44**, 1–36.

Atkinson, A. C., Mulira, H.–M. (1993), The Stalactite Plot for the Detection of Multivariate Outliers, *Statistics and Computing*, **3**, 27–35.

Bacon–Shone, J., Fung, W. K. (1987), A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data, *Applied Statistics*, **36**, 153–162.

Barnett, V. (1979), Some Outlier Tests for Multivariate Samples, *South African Statistical Journal*, **13**, 29–52.

Barnett, V. (1983), Marginal Outliers in the Bivariate Normal Distribution, *Bulletin of the International Statistical Institute*, **50 (4)**, 579–583.

Barnett, V., Lewis, T. (1994), *Outliers in Statistical Data*, 3^{rd} ed., Wiley, New York.

Beaton, A. E., Tukey, J. W. (1974), The Fitting of Power Series, Meaning Polynomials, Illustrated on Band–Spectroscopic Data, *Technometrics*, **16**, 147–185.

Becker, C. (1992), *Multivariate Ausreißer–Identifizierer mit hohem Bruchpunkt*, Diplomarbeit, Universität Dortmund.

Bellman, R. E. (1961), *Adaptive Control Processes*, Princeton University Press.

Bendre, S., Kale, B. K. (1985), Masking Effect on Tests for Outlier in Exponential Samples, *Journal of the American Statistical Association*, **80**, 1020–1025.

Bendre, S., Kale, B. K. (1987), Masking Effect on Tests for Outliers in Normal Samples, *Biometrika*, **74**, 891–896.

Bertin, T. (1967), *Semiologie Graphique*, Gauthier–Villars, Paris.

Bhandary, M. (1992), Detection of the Numbers of Outliers Present in a Data Set Using an Information Theoretic Criterion, *Communications in Statistics – Theory and Methods*, **21**, 3263–3274.

Bickel, P. J. (1976), Another Look at Robustness: A Review of Reviews and Some New Developments, *Scandinavian Journal of Statistics*, **3**, 145–168.

Blatter, C. (1979), *Analysis II*, 2. Aufl., Springer, Berlin.

Boscher, H. (1992), *Behandlung von Ausreißern in linearen Regressionsmodellen*, Dissertation, Fachbereich Statistik, Universität Dortmund.

Brownlee, K. A. (1965), *Statistical Theory and Methodology in Science and Engineering*, 2nd ed., Wiley, New York.

Caroni, C., Prescott, P. (1992), Sequential Application of Wilks's Multivariate Outlier Test, *Applied Statistics*, **41**, 355–364.

Carroll, R. J., Ruppert, D. (1985), Transformations in Regression: A Robust Analysis, *Technometrics*, **27**, 1–12.

Chernoff, H. (1973), Using Faces to Represent Points in k–dimensional Space Graphically, *Journal of the American Statistical Association*, **68**, 361–368.

Choudhury, D. R., Das, M. N. (1992), Use of Combinatorics for Unique Detection of Unknown Numbers of Outliers Using Group Tests, *Sankhyā B*, **54**, 92–99.

Christmann, A., Gather, U., Scholz, G. (1994), Some Properties of the Length of the Shortest Half, *Statistica Neerlandica*, **48**, 209–213.

Daniel, C., Wood, F. S. (1971), *Fitting Equations to Data*, Wiley, New York.

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

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

Davies, P. L., Gather, U. (1993), The Identification of Multiple Outliers, *Journal of the American Statistical Association*, **88**, 782–792.

Dempster, A. P., Gasko–Green, M. (1981), New Tools for Residual Analysis, *The Annals of Statistics*, **9**, 945–959.

Donoho, D. L., Huber, P. J. (1983), The Notion of Breakdown Point, in: Bickel, P. J., Doksum, K. A., Hodges, J. L. (eds.), *A Festschrift for Erich Lehmann*, Wadsworth, Belmont, CA, 157–184.

Easton, G. S., McCulloch, R. E. (1990), A Multivariate Generalization of Quantile–Quantile Plots, *Journal of the American Statistical Association*, **85**, 376–386.

Edgeworth, F. Y. (1887), The Choice of Means, *Philosophical Magazine*, **24**, Ser. 5, 268–271.

Fahrmeir, L., Hamerle, A., Tutz, G. (Hrsg.) (1996), *Multivariate statistische Verfahren*, 2. Aufl., de Gruyter, Berlin.

Ferguson, T. S. (1961), On the Rejection of Outliers, *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 253–287.

Flury, B., Riedwyl, H. (1981), Graphical Representation of Multivariate Data by Means of Asymmetrical Faces, *Journal of the American Statistical Association*, **76**, 757–765.

Friedman, J. H. (1994), An Overview of Predictive Learning and Function Approximation, in: Cherkassky, V., Friedman, J. H., Wechsler, H. (eds.), *From Statistics to Neural Networks*, Springer, Berlin, 1–61.

Fung, W. K. (1988), Critical Values for Testing in Multivariate Statistical Outliers, *Journal of Statistical Computation and Simulation*, **30**, 195–212.

Galambos, J. (1987), *The Asymptotic Theory of Extreme Order Statistics*, 2nd ed., Krieger Publishing Company, Malabar, Florida.

Gather, U., Pigeot, I. (1994), Identifikation von Ausreißern als multiples Testproblem, in: Pöppl, S. J., Lipinski, H.-G., Mansky, T. (Hrsg.), *Medizinische Informatik: Ein integrierender Teil arztunterstützender Technologien* (38. Jahrestagung der GMDS, Lübeck, September 1993), MMV Medizin Verlag, München, 474–477.

Gnanadesikan, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley, New York.

Gnanadesikan, R., Kettenring, J. R. (1972), Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data, *Biometrics*, **28**, 81–124.

Goldwyn, R. M., Friedman, H. P., Siegel, T. H. (1971), Iteration and Interaction in Computed Data Bank Analysis; Case Study in Physiological Classification and Assessment of the Critically Ill, *Computers in Biomedical Research*, **4**, 607–622.

Gordaliza, A.(1991), On the Breakdown Point of Multivariate Location Estimators Based on Trimming Procedures, *Statistics & Probability Letters*, **11**, 387–394.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W. A. (1986), *Robust Statistics. The Approach Based on Influence Functions*, Wiley, New York.

Hara, T. (1988), Detection of Multivariate Outliers with Location Slippage or Scale Inflation in Left Orthogonally Invariant or Elliptically Contoured Distributions, *Annals of the Institute of Statistical Mathematics*, **40**, 395–406.

Harter, H. L. (1964), Criteria for Best Substitute Interval Estimation, *Journal of the American Statistical Association*, **59**, 1133–1140.

Hawkins, D. M. (1973), Repeated Testing for Outliers, *Statistica Neerlandica*, **27**, 1–10.

Hawkins, D. M. (1974), The Detection of Errors in Multivariate Data Using Principal Components, *Journal of the American Statistical Association*, **69**, 340–344.

Hawkins, D. M. (1980), *Identification of Outliers*, Chapman and Hall, London.

Hawkins, D. M., Bradu, D., Kass, G. V. (1984), Location of Several Outliers in Multiple Regression Data Using Elemental Sets, *Technometrics*, **26**, 197–208.

Healy, M. J. R. (1968), Multivariate Normal Plotting, *Applied Statistics*, **17**, 157–161.

Huber, P. J. (1972), Robust Statistics: A Review (The 1972 Wald Lecture), *The Annals of Mathematical Statistics*, **43**, 1041–1067.

Huber, P. J. (1981), *Robust Statistics*, Wiley, New York.

Jennings, L. W., Young, D. M. (1988), Extended Critical Values of the Multivariate Extreme Deviate for Detecting a Single Spurious Observation, *Communications in Statistics – Simulation and Computation*, **17**, 1359–1373.

Jerison, H. J. (1973), *Evolution of the Brain and Intelligence*, Academic Press, New York.

Johnson, A. J., Wichern, D. W. (1982), *Applied Multivariate Statistical Analysis*, Prentice–Hall, Englewood Cliffs, New Jersey.

Karlin, S., Truax, D. (1960), Slippage Problems, *The Annals of Mathematical Statistics*, **31**, 296–324.

Kimber, A. C. (1982), Tests for Many Outliers in an Exponential Sample, *Applied Statistics*, **31**, 263–271.

Kleiner, B., Hartigan, J. A. (1981), Representing Points in Many Dimensions by Trees and Castles (with Discussion), *Journal of the American Statistical Association*, **76**, 260–276.

Launer, R. L., Wilkinson, G. N. (eds.) (1979), *Robustness in Statistics*, Academic Press, New York.

Lehmann, E. L. (1983), *Theory of Point Estimation*, Wiley, New York.

Li, G. (1985), Robust Regression, in: Hoaglin, D., Mosteller, F., Tukey, J. (eds.), *Exploring Data Tables, Trends, and Shapes*, Wiley, New York, 281–343.

Lopuhaä, H. P. (1989), On the Relation Between S–Estimators and M–Estimators of Multivariate Location and Covariance, *The Annals of Statistics*, **17**, 1662–1683.

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

Mathar, R. (1981), *Ausreißer bei ein- und mehrdimensionalen Wahrscheinlichkeitsverteilungen*, Dissertation, Mathematisch–Naturwissenschaftliche Fakultät, RWTH Aachen.

Morgenthaler, S., Ronchetti, E. M., Stahel, W. A. (eds.) (1993), *New Directions in Statistical Data Analysis and Robustness*, Birkhäuser, Basel.

Naik, D. N. (1990), On Detection of Outliers in Symmetric Normal Models, *Communications in Statistics – Theory and Methods*, **19**, 2315–2321.

Neffe, J. (1993), „Auf der Seite der Sieger“, *DER SPIEGEL*, **23/1993**, 200–214.

Pigeot, I. (1993), *Multiple Testtheorie in der Ausreißererkennung*, Habilitationsschrift, Fachbereich Statistik, Universität Dortmund.

Prescott, P. (1979), Critical Values for a Sequential Test for Many Outliers, *Applied Statistics*, **28**, 36–39.

Radhakrishnan, R., Kshirsagar, A. M. (1981), Influence Functions for Certain Parameters in Multivariate Analysis, *Communications in Statistics A*, **10**, 515–529.

Rao, C. R. (1964), The Use and Interpretation of Principal Component Analysis in Applied Research, *Sankhyā A*, **26**, 329–358.

Rocke, D. M. (1993), Robustness Properties of S–Estimators of Multivariate Location and Shape in High Dimension, *Preprint*, Graduate School of Management, University of California at Davis.

Rocke, D. M., Woodruff, D. L. (1993), Identification of Outliers in Multivariate Data, *Preprint*, University of California at Davis.

Rosner, B. (1975), On the Detection of Many Outliers, *Technometrics*, **17**, 221–227.

Rosner, B. (1983), Percentage Points for a Generalized ESD Many–Outlier Procedure, *Technometrics*, **25**, 165–172.

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

Rousseeuw, P. J., Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, Wiley, New York.

Rousseeuw, P. J., Yohai, V. (1984), Robust Regression by Means of S-Estimators, in: *Robust and Nonlinear Time Series Analysis, Lecture Notes in Statistics*, **26**, Springer, New York, 256–272.

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

Schwager, S. J., Margolin, B. H. (1982), Detection of Multivariate Normal Outliers, *The Annals of Statistics*, **10**, 943–954.

Seaman, J. W. Jr., Turner, D. W., Young, D. M. (1987), Polyhedron Graphs for Displaying Multivariate Data, *Computers and Mathematics with Applications A*, **14**, 269–277.

Simonoff, J. S. (1984), A Comparison of Robust Methods and Detection of Outliers Techniques When Estimating a Location Parameter, *Communications in Statistics – Theory and Methods*, **13**, 813–842.

Sinha, B. K. (1984), Detection of Multivariate Outliers in Elliptically Symmetric Distributions, *The Annals of Statistics*, **12**, 1558–1565.

Stahel, W., Weisberg, S. (eds.) (1991a), *Directions in Robust Statistics and Diagnostics. Part I, Vol. 33*, Springer, New York.

Stahel, W., Weisberg, S. (eds.) (1991b), *Directions in Robust Statistics and Diagnostics. Part II, Vol. 34*, Springer, New York.

Staudte, R. G., Sheater, S. J. (1990), *Robust Estimates and Testing*, Wiley, New York.

Sweeting, T. J. (1983), Independent Scale–Free Spacings for the Exponential and Uniform Distributions, *Statistics & Probability Letters*, **1**, 115–119.

Theobald, C. M. (1975), An Inequality with Application to Multivariate Analysis, *Biometrika*, **62**, 461–466.

Tyler, D. E. (1994), Finite Sample Breakdown Points of Projection Based Multivariate Location and Scatter Statistics, *The Annals of Statistics*, **22**, 1024–1044.

Wainer, H. (1981), Comment on Kleiner and Hartigan, *Journal of the American Statistical Association*, **76**, 272–275.

Wilks, S. S. (1963), Multivariate Statistical Outliers, *Sankhyā A*, **25**, 407–426.

Witting, H., Nölle, G. (1970), *Angewandte mathematische Statistik*, Teubner, Stuttgart.

Woodruff, D. L., Rocke, D. M. (1993), Heuristic Search Algorithms for the Minimum Volume Ellipsoid, *Journal of Computational and Graphical Statistics*, **2**, 69–95.

Zurmühl, R., Falk, S. (1986), *Matrizen und ihre Anwendungen. Teil 2: Numerische Methoden*, 5. Aufl., Springer, Berlin.