Multivariate Statistik

Vorlesungsskriptum

Univ.-Prof. Dipl.-Ing. Dr.techn.
Peter Filzmoser

Institut für Stochastik und Wirtschaftsmathematik

Wien, im Oktober 2018

Vorwort

Unter *multivariater Analyse* versteht man die Untersuchung mehrerer Variablen, die an mehreren Objekten simultan beobachtet wurden. Das notwendige Instrumentarium dafür nennt man *multivariate Methoden*. *Multivariate Statistik* ist somit die Erweiterung der *univariaten* (eindimensionalen) Statistik auf mehrere Dimensionen.

Ein Beispiel von Daten, mit denen eine multivariate Analyse durchgeführt werden kann, sind Prüfungsergebnisse von Studenten in mehreren Gegenständen. Die Objekte oder Beobachtungen sind die Studenten; ihre Leistungen wurden jeweils für die Gegenstände (Variablen) ermittelt. Der gravierende Unterschied zur *univariaten Analyse* ist, dass nicht jede Variable (Prüfungsfach) gesondert untersucht wird, sondern simultan alle Messungen ausgewertet werden. Für unser Beispiel möchte man etwa Aussagen darüber machen, ob die Unterrichtsgegenstände in Gruppen eingeteilt werden können. Die Interpretation der Gruppen könnte dann lauten, dass die Gegenstände, die eine Gruppe bilden, mathematisches Verständnis voraussetzen, während eine andere Gruppe Gegenstände mit Schwerpunkt auf geometrischem Verständnis beinhaltet. Eine interessante Fragestellung wäre auch, in welchen Unterrichtsfächern die Talente eines Studenten liegen und ob gleich talentierte Studenten die selbe schulische Vorbildung haben. Abhängig von der Antwort könnten Rückschlüsse auf verschiedene Typen von Mittelschulen gemacht werden.

Ein umfangreicheres Beispiel wäre, die Gesellschaftsstruktur der europäischen Länder zu ermitteln. Anhand von Wirtschafts-, Sozial-, Umwelt-, Bevölkerungsdaten etc. könnten interessante Fragestellungen mittels multivariater Analysen (hoffentlich) umfassend beantwortet werden. Die Antworten und Rückschlüsse könnten die Grundlage für politische Entscheidungen bilden.

Schon durch diese beiden Beispiele ist erkennbar, dass Fragestellungen, die mehrere Themen oder Gebiete in Betracht ziehen, oft erst mit Hilfe von multivariaten Analysen beantwortet werden können. Notwendige Voraussetzung für sinnvolle Aussagen sowie für eine sinnvolle Anwendung der Methoden ist natürlich eine entsprechende Erhebung der Daten.

Allgemein könnte man sagen, dass die multivariate Analyse folgende Zielsetzungen hat:

- Datenreduktion und Vereinfachung der zu untersuchenden Struktur
- sortieren und gruppieren
- Untersuchung von Abhängigkeiten der Variablen
- Vorhersage
- Testen von Hypothesen

Die Entwicklung der Theorie der multivariaten Analyse begann etwa 1930, wobei es bei den Methoden vorerst noch starke Einschränkungen gab (z.B. multivariate Normalverteilung). Mit der Erfindung des Computers wurden multivariate Verfahren immer attraktiver. Die schnellen Rechner in der heutigen Zeit erlauben es, die Modelle der multivariaten Methoden immer besser der Wirklichkeit anzupassen.

Inhaltsverzeichnis

Kapitel 1

Einführung

1.1 Einfache grafische Techniken

Ein sehr einfaches aber hilfreiches Mittel in der Datenanalyse sind Grafiken. Gerade wegen ihrer Einfachheit werden diese Hilfsmittel oft als „minderwertig“ oder „wenig wissenschaftlich“ eingestuft. Die Erfahrung zeigt aber, dass eine grafische Darstellung viel Einsicht vermittelt über die Qualität und Struktur der Daten.

Eine sehr bekannte Darstellung von Daten im zweidimensionalen Raum ist das *Streudiagramm*. Dabei werden zwei Variablen einander gegenübergestellt und die Beobachtungen in dieses Koordinatensystem eingetragen.

Beispiel 1.1.1 *In einem Artikel des Magazins “Forbes” vom 30. April 1990 wurden von den größten Verlagen Finanzdaten publiziert. Wir betrachten dabei die Variable x, die die Anzahl der Beschäftigten angibt, und die Variable y mit dem Profit pro Beschäftigtem. Abbildung 1.1 zeigt ein Streudiagramm dieser Daten. Es ist allerdings auffällig, dass “Dun & Bradstreet”, die Firma mit den meisten Beschäftigten, eigentlich typisches Verhalten bei Profit pro Beschäftigtem zeigt und somit vom Hauptteil der Daten abweicht. “Time Warner” hat zwar eine typische Anzahl von Beschäftigten, hat aber im Vergleich geringen Profit (sogar negativ) pro Beschäftigtem. Der empirische Korrelationskoeffizient zwischen x und y beträgt -0.39. Wird “Dun & Bradstreet” von der Berechnung weggelassen, erhält man einen Koeffizienten von -0.56. Nimmt man zusätzlich “Time Warner” heraus, wird der Korrelationskoeffizient -0.50. Das Beispiel zeigt also gut, dass Beobachtungen, die von der Hauptstruktur der Daten abweichen, relativ starken Einfluss auf statistische Schätzer haben können. Visuell könnte man diese atypischen Beobachtungen gut durch eine zweidimensionale Erweiterung des Boxplots, den sogenannten “Bagplot” erkennen (Rousseeuw et al., 1999). In Abbildung 1.2 werden die Daten mit Hilfe des Bagplots dargestellt. Analog zum univariaten Boxplot werden (robustes) Mittel, der innere 50%-Bereich der Daten (dunkel schattiert), und ein Bereich (hell schattiert), der die Ausreißer von den regulären Beobachtungen abgrenzt, dargestellt. Man erkennt sehr gut die schon vorher vermuteten Beobachtungen als Ausreißer. Durch die Form des dunkel schattierten “bags” bietet diese Darstellung auch ein Maß für die Korrelation.*

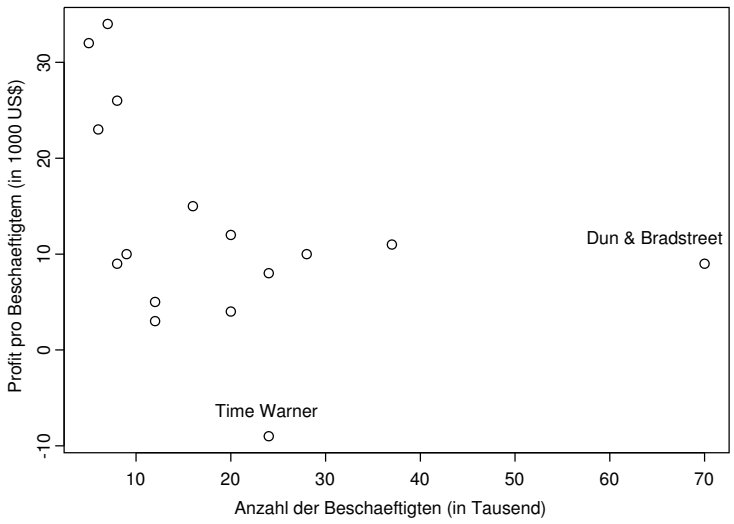


Abbildung 1.1: Streudiagramm für die 2-dimensionalen Finanzdaten

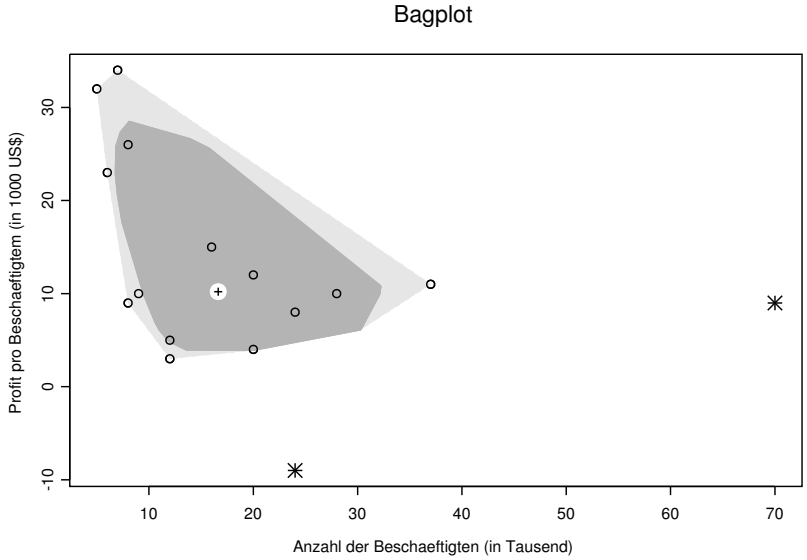


Abbildung 1.2: 2-dimensionale Boxplot-Darstellung der Finanzdaten

Das Streudiagramm kann natürlich auch für mehr als 2-dimensionale Daten eingesetzt werden, wie folgendes Beispiel zeigt:

Beispiel 1.1.2 *Tabelle 1.1 zeigt Daten der Qualität von Papier. Nachdem die Fasern im Papier in einer bestimmten Richtung ausgerichtet sind, werden verschiedene Reißfestigkeiten beobachtet, wenn entlang der Richtung, in der das Papier produziert wird, oder quer dazu gemessen wird. Variable x_1 gibt die Dichte des Papiers in g/cm^3, Variable x_2 bzw. x_3 geben die Reißfestigkeit entlang der Produktion von der Maschine bzw. quer dazu in Pfund an.*

Diese 3-dimensionalen Daten können wiederum durch Streudiagramme visua-

Tabelle 1.1: Papierqualität; x_1 = Dichte, x_2 = Reißfestigkeit längs, x_3 = Reißfestigkeit quer.

x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
0.801	121.41	70.42	0.832	117.51	71.62	0.822	130.50	80.33
0.824	127.70	72.47	0.796	109.81	53.10	0.822	127.90	75.68
0.841	129.20	78.20	0.759	109.10	50.85	0.843	123.90	78.54
0.816	131.80	74.89	0.770	115.10	51.68	0.824	124.10	71.91
0.840	135.10	71.21	0.759	118.31	50.60	0.788	120.80	68.22
0.842	131.50	78.39	0.772	112.60	53.51	0.782	107.40	54.42
0.820	126.70	69.02	0.806	116.20	56.53	0.795	120.70	70.41
0.802	115.10	73.10	0.803	118.00	70.70	0.805	121.91	73.68
0.828	130.80	79.28	0.845	131.00	74.35	0.836	122.31	74.93
0.819	124.60	76.48	0.822	125.70	68.29	0.788	110.60	53.52
0.826	118.31	70.25	0.971	126.10	72.10	0.772	103.51	48.93
0.802	114.20	72.88	0.816	125.80	70.64	0.776	110.71	53.67
0.810	120.30	68.23	0.836	125.50	76.33	0.758	113.80	52.42
0.802	115.70	68.12	0.815	127.80	76.75			

lisiert werden, indem alle Paare von Variablen als Koordinatenachsen genommen werden (siehe Abbildung 1.3).

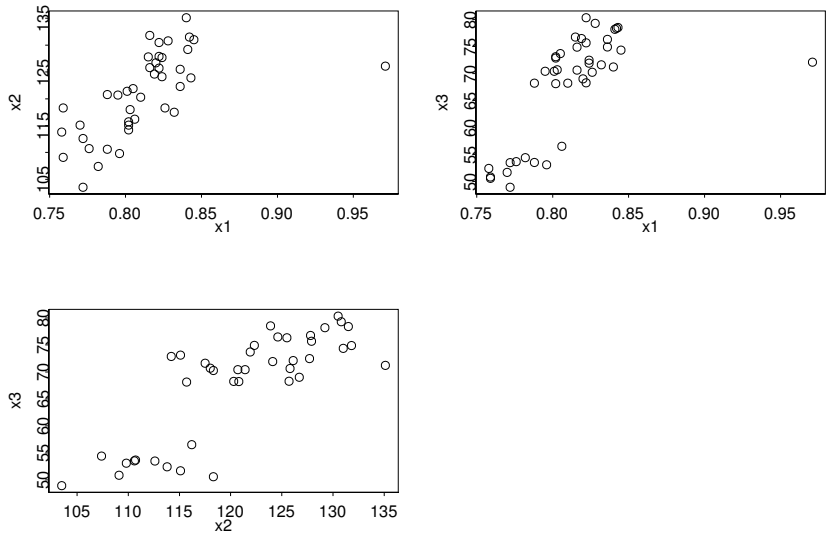


Abbildung 1.3: Streudiagramme der Papierdaten

Die modernere Computersoftware macht es möglich, Daten auch in einem 3-dimensionalen Streudiagramm darzustellen. Durch kontinuierliche Rotation der 3 Koordinatenachsen erhält man visuell einen Eindruck der Datenstruktur. Es ist auch denkbar, mehr als 3 Dimensionen durch ein Streudiagramm darzustellen. Ermöglicht

Tabelle 1.2: Festigkeitsmessung von Platten; x_1: Messung durch Druckwelle, x_2: Messung durch Vibration, x_3, x_4: durch statische Tests.

x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
1889	1651	1561	1778	1954	2149	1180	1281
2403	2048	2087	2197	1325	1170	1002	1176
2119	1700	1815	2222	1419	1371	1252	1308
1645	1627	1110	1533	1828	1634	1602	1755
1976	1916	1614	1883	1725	1594	1313	1646
1712	1712	1439	1546	2276	2189	1547	2111
1943	1685	1271	1671	1899	1614	1422	1477
2104	1820	1717	1874	1633	1513	1290	1516
2983	2794	2412	2581	2061	1867	1646	2037
1745	1600	1384	1508	1856	1493	1356	1533
1710	1591	1518	1667	1727	1412	1238	1469
2046	1907	1627	1898	2168	1896	1701	1834
1840	1841	1595	1741	1655	1675	1414	1597
1867	1685	1493	1678	2326	2301	2065	2234
1859	1649	1389	1714	1490	1382	1214	1284

wird dies durch die *Grand-Tour* (Asimov, 1985), einer Methode, die 2-dimensionale Projektionen des p-dimensionalen Raumes berechnet, wobei man durch die p Dimensionen kontinuierlich „hindurchgleitet“. Ein Softwarepaket mit einer Implementierung der Grand-Tour ist *GGobi* (`http://www.ggobi.org`).

Beispiel 1.1.3 *Tabelle 1.2 zeigt 4 Festigkeitsmessungen von 30 Platten. Die erste Messung wurde erhalten, indem eine Druckwelle auf die Platte ausgesandt wurde, bei der zweiten Messung wurde die Platte einer Vibration ausgesetzt, und die restlichen beiden Messungen wurden durch statische Tests erhalten.*

Abbildung 1.4 zeigt verschiedene Perspektiven der Daten von Tabelle 1.2 in einem 3-dimensionalen Streudiagramm. In Abbildung 1.4a sind die Ausreißer gut erkennbar, Abbildung 1.4b hingegen maskiert die Ausreißer.

Abbildung 1.4: 3-dimensionale Streudiagramme mit verschiedenen Perspektiven von den Plattendaten.

Selbst für diese einfachen Darstellungen von Daten kann man mit guter Software viel Information extrahieren. Z.B. sind in Abbildung 1.3 deutlich zwei Gruppen und ein Ausreißer erkennbar. Wenn man diese Gruppen markiert, ev. mit verschiedenen Farben, kann man durch Verknüpfungen zu anderen Grafiken (*linking*) diese Gruppen deutlich machen. Besonders interessant ist diese Methode bei Regionaldaten: Die verschiedenfärbigen Gruppen können in einer Landkarte visualisiert werden. Man könnte außerdem Objekte aus anderen Gruppen wegnehmen, und mit der verbleibenden homogenen Gruppe neue Statistiken berechnen.

Es gibt neben den Streudiagrammen noch die *multivariaten Grafiken*. Dabei wird versucht, mehrdimensionale Daten durch bestimmte grafische Symbole zu visualisieren. Bei den *Star-Plots* werden die Variablen als sternförmige Achsen dargestellt, und die Werte für jede Variable eines Objektes in diesem sternförmigen Koordinatensystem eingetragen. Danach werden die eingetragenen Werte benachbarter Variablen verbunden. Es entsteht somit für jedes Objekt ein „sternförmiges Gebilde“, die für relativ gleichartige Objekte sehr ähnlich aussehen. Somit können Gruppen in den Daten sichtbar gemacht werden.

Abbildung 1.5 zeigt einen Star-Plot der Wahlergebnisse der republikanischen Partei von 1856-1976 (31 Werte) für 9 der nordöstlichen Bundesstaaten der USA. Die Werte werden, von rechts beginnend, im Uhrzeigersinn aufgetragen. Man erkennt zum Teil sehr ähnliche Strukturen.

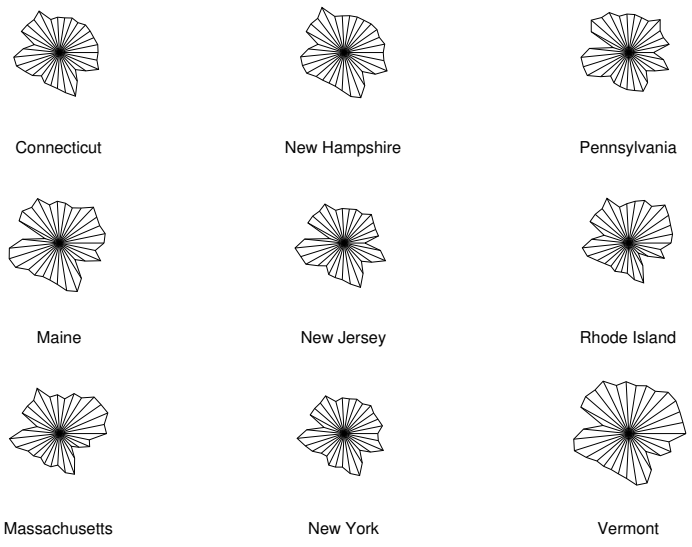


Abbildung 1.5: Star-Plot der Wahlergebnisse 1856-1976 für 9 Bundesstaaten der USA.

Eine andere bekannte, aber manchmal sehr irreführende multivariate Grafik sind die *Chernoff-Faces*. Augen, Ohren, Mund, Nase, Gesichtsform, ..., repräsentieren die verschiedenen Variablen. Die Objekte werden dann als „Gesicht“ dargestellt, wobei der Wert jeder Variable die Form oder Größe des Merkmals angibt.

1.2 Notation

Eine Matrix $\boldsymbol{X}$ der Ordnung $(n \times p)$ wird geschrieben als

$$\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \quad .$$

Generell wird eine Matrix mit fettgedruckten Großbuchstaben bezeichnet. Eine Matrix mit Spaltenordnung 1 wird Spaltenvektor genannt. Daher ist

$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

ein Spaltenvektor mit n Komponenten. Spaltenvektoren werden mit fettgedruckten Kleinbuchstaben bezeichnet. Ein Spaltenvektor kann auch transponiert dargestellt werden:

$$\boldsymbol{x}^\top = (x_1, \dots, x_n) \quad .$$

Die Spalten einer Matrix $\boldsymbol{X}$ werden geschrieben als

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p$$

und die Zeilen (geschrieben als Spaltenvektoren)

$$\boldsymbol{x}_{1.}, \boldsymbol{x}_{2.}, \dots, \boldsymbol{x}_{n.} \quad .$$

Diese Notation ist eher unüblich, und oft wird das selbe Symbol für Zeilen oder Spalten einer Matrix verwendet. Dies kann aber zu Verwirrungen führen, und daher scheint diese Lösung sinnvoll zu sein. Somit ist

$$\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p) = \begin{pmatrix} \boldsymbol{x}_{1.}^\top \\ \boldsymbol{x}_{2.}^\top \\ \vdots \\ \boldsymbol{x}_{n.}^\top \end{pmatrix}$$

$$\text{mit} \quad \boldsymbol{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix} \quad \text{und} \quad \boldsymbol{x}_{i.} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} \quad .$$

1.3 Beschreibende Statistik

Wie im univariaten (eindimensionalen) Fall möchte man auch für multivariate Daten einfache statistische Kenngrößen angeben. Das Stichprobenmittel gibt Aufschluß über die Lage bzw. das Zentrum der betrachteten Größe. Sinnvollerweise wird man dieses Lagemaß für jede betrachtete Variable (d.h. für jede Spalte einer Datenmatrix) gesondert ermitteln.

Seien $x_{11}, x_{21}, \ldots, x_{n1}$ die n Messungen der ersten Variable. Dann ist das *arithmetische Mittel* dieser Messungen für die erste Variable

$$\bar{x}_1 = \frac{1}{n} \sum_{i=1}^{n} x_{i1} \ .$$

Auf analoge Weise kann das arithmetische Mittel für die anderen Variablen bestimmt werden. Man erhält somit $\bar{x}_1, \ldots, \bar{x}_p$ für p Variablen der $(n \times p)$-Datenmatrix $\boldsymbol{X}$. Das *multivariate (arithmetische) Mittel* ist dann der p-dimensionale Vektor $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^\top$.

Eine weitere wichtige Kenngröße der beschreibenden Statistik ist die *Varianz*. Für unsere Matrix $\boldsymbol{X}$ ist die Stichprobenvarianz der ersten Variable

$$s_{11} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \ .$$

Im allgemeinen kann die Stichprobenvarianz für die p Variablen definiert werden durch

$$s_{jj} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

für $j = 1, \ldots, p$. Nachdem auch ein Variationsmaß zwischen den einzelnen Variablen angegeben werden kann (Kovarianz), sollte man diese Streuungsmaße in Matrixform anschreiben. Wir ordnen daher die Stichprobenvarianzen s_{jj} entlang der Hauptdiagonale einer (Varianz-)Kovarianz-Matrix $\boldsymbol{S}$ an. Die *Kovarianz* s_{jk}, die den linearen Zusammenhang zweier Variablen j und k angibt, ist definiert als

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \ .$$

Es ist offensichtlich, dass die Kovarianzmatrix symmetrisch ist, da $s_{jk} = s_{kj}$ gilt.

Ein weiteres wichtiges Maß ist der Korrelationskoeffizient. Er ist ein standardisiertes Maß für den linearen Zusammenhang zweier Variablen und ist definiert durch

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}\sqrt{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}}$$

für $j = 1, \ldots, p$ und $k = 1, \ldots, p$. Klarerweise gilt $r_{jj} = 1$. Die Elemente r_{jk} können in der Stichproben-Korrelationsmatrix $\boldsymbol{R}$ zusammengefasst werden.

1.4 Eigenwerte und Eigenvektoren

Sei $\boldsymbol{\Sigma}$ eine $(p \times p)$-Matrix (quadratisch). Dann ist

$$q(a) = |\boldsymbol{\Sigma} - a\boldsymbol{I}| \tag{1.1}$$

ein Polynom der Ordnung p in a. Die p-ten Wurzeln von $q(a)$, nämlich $a_1, \ldots, a_p$, nennt man Eigenwerte von $\boldsymbol{\Sigma}$.

Es gilt dann für $i = 1, \ldots, p$

$$|\boldsymbol{\Sigma} - a_i\boldsymbol{I}| = 0 \quad , \tag{1.2}$$

sodass $\boldsymbol{\Sigma} - a_i\boldsymbol{I}$ singulär ist. Es gibt daher einen Vektor $\boldsymbol{\gamma}_i \neq \mathbf{0}$ mit

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_i = a_i\boldsymbol{\gamma}_i \quad \text{für } i = 1, \ldots, p. \tag{1.3}$$

$\boldsymbol{\gamma}_i$ wird bezeichnet als (Rechts-) Eigenvektor von $\boldsymbol{\Sigma}$ zum Eigenwert a_i. Ein reeller Eigenvektor $\boldsymbol{\gamma}_i$ heißt standardisiert, wenn gilt: $\boldsymbol{\gamma}_i^\top \boldsymbol{\gamma}_i = 1$.

Sind $\boldsymbol{\xi}$ und $\boldsymbol{\zeta}$ Eigenvektoren zum Eigenwert a_i, dann sind $\boldsymbol{\xi} + \boldsymbol{\zeta}$ und $\alpha\boldsymbol{\xi}$ $(\alpha \in I\!R)$ auch Eigenvektoren zu a_i.

Satz 1.4.1 *Sei $\boldsymbol{C}$ eine nichtsinguläre $(p \times p)$-Matrix. Dann gilt:*

$$|\boldsymbol{\Sigma} - a_i\boldsymbol{I}| = |\boldsymbol{C}|\ |\boldsymbol{\Sigma} - a_i\boldsymbol{C}^{-1}\boldsymbol{C}|\ |\boldsymbol{C}^{-1}| = |\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{-1} - a_i\boldsymbol{I}| \quad . \tag{1.4}$$

D.h. $\boldsymbol{\Sigma}$ und $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{-1}$ haben dieselben Eigenwerte. Ist $\boldsymbol{\gamma}$ ein Eigenvektor von $\boldsymbol{\Sigma}$ zum Eigenwert a_i, dann ist

$$a_i\boldsymbol{C}\boldsymbol{\gamma} = \boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{-1}(\boldsymbol{C}\boldsymbol{\gamma}) \quad . \tag{1.5}$$

Daraus folgt, dass $\boldsymbol{C}\boldsymbol{\gamma}$ ein Eigenvektor von $\boldsymbol{C}\boldsymbol{\Sigma}\boldsymbol{C}^{-1}$ zum Eigenwert a_i ist.

Satz 1.4.2 *Alle Eigenwerte einer symmetrischen $(p \times p)$-Matrix $\boldsymbol{\Sigma}$ sind reell.*

Beweis: Sei $\boldsymbol{\gamma} = \boldsymbol{\xi} + i\boldsymbol{\zeta}$, $a = b + ic$, $\boldsymbol{\gamma} \neq \mathbf{0}$.

Durch Einsetzen in $\boldsymbol{\Sigma}\boldsymbol{\gamma} = a\boldsymbol{\gamma}$, kann Real- und Imaginärteil berechnet werden.

$$\boldsymbol{\Sigma}\boldsymbol{\xi} = b\boldsymbol{\xi} - c\boldsymbol{\zeta} \quad , \quad \boldsymbol{\Sigma}\boldsymbol{\zeta} = c\boldsymbol{\xi} + b\boldsymbol{\zeta} \quad . \tag{1.6}$$

Premultiplikation mit $\boldsymbol{\zeta}^\top$ bzw. $\boldsymbol{\xi}^\top$ und Subtraktion liefert $c = 0$, d.h. a ist reell. $\boldsymbol{\gamma}$ kann auch reell gewählt werden, d.h. $\boldsymbol{\zeta} = \mathbf{0}$. □

Satz 1.4.3 (Spektralzerlegungssatz oder Jordanscher Zerlegungssatz) *Jede symmetrische $(p \times p)$-Matrix $\boldsymbol{\Sigma}$ kann zerlegt werden in*

$$\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^{\top} = \sum_{i=1}^{p} a_i\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^\top \quad , \tag{1.7}$$

wobei $\boldsymbol{A} = Diag(a_1, \ldots, a_p)$ eine Diagonalmatrix mit den Eigenwerten von $\boldsymbol{\Sigma}$ ist und $\boldsymbol{\Gamma}$ eine orthogonale Matrix ist (d.h. $\boldsymbol{\Gamma}^\top = \boldsymbol{\Gamma}^{-1}$), deren Spalten standardisierte Eigenvektoren $\boldsymbol{\gamma}_i$ von $\boldsymbol{\Sigma}$ sind.

Beweis: Seien $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$ orthonormale Vektoren mit

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_i = a_i\boldsymbol{\gamma}_i \tag{1.8}$$

für Zahlen a_i $(i = 1, \ldots, p)$.

Dann ist

$$\boldsymbol{\gamma}_i^\top \boldsymbol{\Sigma}\boldsymbol{\gamma}_j = a_j\boldsymbol{\gamma}_i^\top\boldsymbol{\gamma}_j = \begin{cases} a_i & \text{für } i = j \\ 0 & \text{für } i \neq j \end{cases} , \tag{1.9}$$

oder in Matrixform

$$\boldsymbol{\Gamma}^\top \boldsymbol{\Sigma}\boldsymbol{\Gamma} = \boldsymbol{A} \quad . \tag{1.10}$$

Daraus folgt:

$$\boldsymbol{\Gamma}(\boldsymbol{\Gamma}^\top \boldsymbol{\Sigma}\boldsymbol{\Gamma})\boldsymbol{\Gamma}^{-1} = \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^\top \quad . \tag{1.11}$$

Aus Satz 1.4.1 folgt, dass $\boldsymbol{\Sigma}$ und $\boldsymbol{A}$ die selben Eigenwerte haben, nämlich genau die Elemente von $\boldsymbol{A}$.

Nun muss noch eine Orthonormalbasis aus Eigenvektoren gefunden werden. Sind $a_i \neq a_j$ verschiedene Eigenwerte mit Eigenvektor $\boldsymbol{\xi} + \boldsymbol{\zeta}$, dann gilt:

$$a_i\boldsymbol{\xi}^\top\boldsymbol{\zeta} = \boldsymbol{\xi}^\top a_i\boldsymbol{\zeta} = \boldsymbol{\xi}^\top\boldsymbol{\Sigma}\boldsymbol{\zeta} = \boldsymbol{\zeta}^\top\boldsymbol{\Sigma}\boldsymbol{\xi} = a_j\boldsymbol{\zeta}^\top\boldsymbol{\xi} \quad \Longrightarrow \quad \boldsymbol{\zeta}^\top\boldsymbol{\xi} = 0 \quad . \tag{1.12}$$

D.h. bei einer symmetrischen Matrix sind die Eigenvektoren, die zu verschiedenen Eigenwerten gehören, orthogonal.

Der Eigenraum einer Matrix $\boldsymbol{\Sigma}$ mit Eigenwert a ist definiert durch $kn(\boldsymbol{\Sigma} - \lambda\boldsymbol{I})$. Angenommen $\boldsymbol{\Sigma}$ hat k verschiedene Eigenwerte mit zugehörigen Eigenräumen $H_1, \ldots, H_k$ der Dimensionen $r_1, \ldots, r_k$ und sei $r = \sum_{j=1}^{k} r_j$.

Da verschiedene Eigenräume orthogonal sind, existiert eine orthonormale Menge $\{(\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{r_1}), (\boldsymbol{e}_{r_1+1}, \boldsymbol{e}_{r_1+2}, \ldots, \boldsymbol{e}_{r_1+r_2}), \ldots, (\boldsymbol{e}_{r_1+r_2+\ldots+r_{k-1}+1}, \ldots, \boldsymbol{e}_{r_1+r_2+\ldots+r_k})\}$, bzw. in kompakterer Schreibweise $\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_r\}$, von Vektoren, sodass die Vektoren mit Index

$$\sum_{i=1}^{j-1} r_i + 1, \sum_{i=1}^{j-1} r_i + 2, \ldots, \sum_{i=1}^{j-1} r_i + r_j - 1, \sum_{i=1}^{j} r_i \tag{1.13}$$

eine Basis von H_j bilden. r_j ist kleiner oder gleich der Vielfachheit des entsprechenden Eigenwertes. Durch Umordnung der Eigenwerte λ_i (falls notwendig) gilt

$$\boldsymbol{\Sigma}\boldsymbol{e}_i = a_i\boldsymbol{e}_i \quad \text{für} \quad i = 1, \ldots, r \quad , \tag{1.14}$$

wobei $r \leq p$ ist. Sind alle Eigenwerte verschieden, ist $r = p$ und $\boldsymbol{\gamma}_{(i)} = \boldsymbol{e}_i$.

Es ist noch zu zeigen, dass der Fall $r < p$ zu einem Widerspruch führt.

Seien o.B.d.A. alle Eigenwerte von $\boldsymbol{\Sigma}$ strikt positiv. (Falls nicht, ersetzen wir $\boldsymbol{\Sigma}$ durch $\boldsymbol{\Sigma} + \alpha\boldsymbol{I}$ für positives α, da $\boldsymbol{\Sigma}$ und $\boldsymbol{\Sigma} + \alpha\boldsymbol{I}$ die selben Eigenvektoren haben.) Sei

$$\boldsymbol{B} = \boldsymbol{\Sigma} - \sum_{i=1}^{r} a_i\boldsymbol{e}_i\boldsymbol{e}_i^\top \quad . \tag{1.15}$$

Dann ist

$$tr\ \boldsymbol{B} = tr\ \boldsymbol{\Sigma} - \sum_{i=1}^{r} a_i(\boldsymbol{e}_i^\top\boldsymbol{e}_i) = \sum_{i=r+1}^{p} a_i > 0 \quad , \tag{1.16}$$

da $r < p$. $\boldsymbol{B}$ hat mindestens einen Eigenwert $\theta \neq 0$. Sei $\boldsymbol{x} \neq \boldsymbol{0}$ der zugehörige Eigenvektor, so gilt für $1 \leq j \leq r$:

$$\begin{aligned}\theta \boldsymbol{e}_j^\top \boldsymbol{x} = \boldsymbol{e}_j^\top \boldsymbol{B}\boldsymbol{x} &= \boldsymbol{e}_j^\top \Big(\boldsymbol{\Sigma} - \sum_{i=1}^{r} a_i \boldsymbol{e}_i \boldsymbol{e}_i^\top\Big)\boldsymbol{x} \\ &= \Big(a_j \boldsymbol{e}_j^\top - \sum_{i=1}^{r} a_i (\boldsymbol{e}_j^\top \boldsymbol{e}_i)\boldsymbol{e}_i^\top\Big)\boldsymbol{x} = (a_j \boldsymbol{e}_j^\top - a_j \boldsymbol{e}_j^\top)\boldsymbol{x} = 0 \quad .\end{aligned}$$

D.h. $\boldsymbol{x}$ ist orthogonal zu $\boldsymbol{e}_j$ für $j = 1, \ldots, r$. Daher ist

$$\theta \boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} = \Big(\boldsymbol{\Sigma} - \sum_{i=1}^{r} a_i \boldsymbol{e}_i \boldsymbol{e}_i^\top\Big)\boldsymbol{x} = \boldsymbol{\Sigma}\boldsymbol{x} - \sum_{i=1}^{r} a_i \boldsymbol{e}_i (\boldsymbol{e}_i^\top \boldsymbol{x}) = \boldsymbol{\Sigma}\boldsymbol{x} \quad , \tag{1.17}$$

sodass $\boldsymbol{x}$ ein Eigenvektor von $\boldsymbol{\Sigma}$ ist. D.h. $\theta = a_i$ für ein i und $\boldsymbol{x}$ ist Linearkombination gewisser Vektoren $\boldsymbol{e}_i$, was ein Widerspruch zur Orthogonalität von $\boldsymbol{x}$ und $\boldsymbol{e}_i$ ist. □

1.5 Erwartungswert und Kovarianzmatrix

Sei $\boldsymbol{X}$ eine $(p \times q)$-Matrix von Zufallsgrößen, die mit dem Symbol $\boldsymbol{X} = [(x_{ij})]$ bezeichnet wird. Der Erwartungswert dieser Matrix ist definiert als

$$E(\boldsymbol{X}) = [(E(x_{ij}))] \quad , \tag{1.18}$$

der $(p \times q)$-Matrix der Erwartungswerte $E(x_{ij})$. Aus der Linearität des Erwartungswert-Operators folgt für Datenmatrizen $\boldsymbol{A}$, $\boldsymbol{B}$ und $\boldsymbol{C}$

$$E(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{A}\ E(\boldsymbol{X})\ \boldsymbol{B} + \boldsymbol{C} \quad . \tag{1.19}$$

Auch die Kovarianz kann auf den mehrdimensionalen Fall übertragen werden. Seien $\boldsymbol{x}$ und $\boldsymbol{y}$ zwei Zufallsvektoren, die nicht notwendigerweise die selbe Dimension haben. Dann gilt

$$Cov(\boldsymbol{x}, \boldsymbol{y}) = [(Cov(x_i, y_j))] = E[(\boldsymbol{x} - \boldsymbol{\mu_x})(\boldsymbol{y} - \boldsymbol{\mu_y})^\top] = E(\boldsymbol{x}\boldsymbol{y}^\top) - \boldsymbol{\mu_x}\boldsymbol{\mu_y}^\top \quad , \tag{1.20}$$

wobei $\boldsymbol{\mu_x}$ bzw. $\boldsymbol{\mu_y}$ die Erwartungswerte von $\boldsymbol{x}$ bzw. $\boldsymbol{y}$ sind. Wenn $\boldsymbol{x}$ und $\boldsymbol{y}$ statistisch unabhängig sind, gilt $Cov(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{O}$. Für $\boldsymbol{x} = \boldsymbol{y}$ wird statt $Cov(\boldsymbol{x}, \boldsymbol{x})$ nur $Cov(\boldsymbol{x})$ geschrieben und es gilt

$$Cov(\boldsymbol{x}) = [(Cov(x_i, x_j))] = E[(\boldsymbol{x} - \boldsymbol{\mu_x})(\boldsymbol{x} - \boldsymbol{\mu_x})^\top] \quad . \tag{1.21}$$

Die so erhaltene Matrix enthält in der Diagonale $(i = j)$ die Varianzen und wird deshalb oft mit *Varianz-Kovarianz-Matrix* bezeichnet. Für Datenmatrizen $\boldsymbol{A}$ und $\boldsymbol{B}$ gilt, wie leicht zu zeigen ist,

$$Cov(\boldsymbol{A}\boldsymbol{x}, \boldsymbol{B}\boldsymbol{y}) = \boldsymbol{A}\ Cov(\boldsymbol{x}, \boldsymbol{y})\ \boldsymbol{B}^\top \quad , \tag{1.22}$$

woraus weiters folgt

$$Cov(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}\ Cov(\boldsymbol{x})\ \boldsymbol{A}^\top \quad . \tag{1.23}$$

1.6 Die multivariate Normalverteilung

Die Dichte der multivariaten Normalverteilung wird erhalten, indem von der Dichte der univariaten Normalverteilung

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-[(x-\mu)/\sigma]^2/2} \qquad -\infty < x < \infty \tag{1.24}$$

die Distanz $(x-\mu)/\sigma$ durch eine multivariate Distanz

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) := \mathrm{MD}^2(\boldsymbol{x}) \tag{1.25}$$

ersetzt wird. Ebenso wird die Konstante $\frac{1}{\sqrt{2\pi\sigma^2}}$ ersetzt durch eine allgemeinere Konstante, die das Volumen, das durch die Dichtefunktion beschrieben wird, auf 1 standardisiert. $\mathrm{MD}(\boldsymbol{x})$ wird mit *Mahalanobis-Distanz* (MD) bezeichnet.

Sei also $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ ein p-dimensionaler Zufallsvektor. Der Erwartungswert von $\boldsymbol{x}$ sei $\boldsymbol{\mu}$, die Kovarianzmatrix sei $\boldsymbol{\Sigma} = [(\sigma_{ij})]$, die positiv definit ist (Schreibweise $\boldsymbol{\Sigma} \geq \boldsymbol{O}$).

Definition 1.6.1 *Mit obigen Bezeichnungen ist* $\boldsymbol{x}$ **p-dimensional normalverteilt**, *wenn für die Dichtefunktion von* $\boldsymbol{x}$ *gilt*

$$f(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1.26}$$

mit $-\infty < x_i < \infty$ *für* $i = 1, \ldots, p$. *Als Notation wird dafür* $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *verwendet.*

Eigenschaften: Für $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ gilt:

(a) $\boldsymbol{x} - \boldsymbol{\mu} \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$

(b) Sei $\boldsymbol{A}$ eine $(q \times p)$-Matrix, dann gilt $\boldsymbol{A}\boldsymbol{x} \sim N_q(\boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^\top)$.

(c) Wird die MD gleich einer Konstanten c gesetzt, also

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) := c^2 \quad , \tag{1.27}$$

so wird dadurch ein Ellipsoid im p-dimensionalen Raum mit Zentrum $\boldsymbol{\mu}$ beschrieben (siehe auch Kapitel Hauptkomponentenanalyse). Die Dichte der p-variaten Normalverteilung ist also auf solchen Ellipsoiden konstant. Es gilt außerdem, dass für normalverteiltes $\boldsymbol{x}$ die Größe

$$(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \sim \chi^2_p \,, \tag{1.28}$$

also einer χ^2-Verteilung mit p Freiheitsgraden folgt (p ist die Dimension von $\boldsymbol{x}$).

Die Größe $\sum_{i=1}^n (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^\top$ ist das multivariate Analogon zur univariaten Quadratsumme $\sum_{i=1}^n (y_i - \bar{y})^2$. Dies führt zur **Wishart-Verteilung**, die folgendermaßen definiert ist:

Definition 1.6.2 *Eine symmetrische $(p \times p)$-Matrix $\boldsymbol{W}$ mit Zufallsgrößen folgt einer Wishart-Verteilung, wenn $\boldsymbol{W}$ mit unabhängigen p-dimensionalen identisch nach $N_p(\mathbf{0}, \boldsymbol{\Sigma})$-verteilten Zufallsvektoren $\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.}$ dargestellt werden kann als*

$$\boldsymbol{W} = \sum_{i=1}^{n} \boldsymbol{x}_{i.} \boldsymbol{x}_{i.}^{\top} = \boldsymbol{X}^{\top} \boldsymbol{X} \quad . \tag{1.29}$$

Man schreibt dafür $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$ oder kurz $\boldsymbol{W} \sim W_p$ und spricht von einer Wishart-Verteilung mit n Freiheitsgraden.

Eigenschaften:

(a) Sei $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$ und $\boldsymbol{A}$ eine $(p \times q)$-Matrix, dann ist

$$\boldsymbol{A}^{\top} \boldsymbol{W} \boldsymbol{A} \sim W_q(\boldsymbol{A}^{\top} \boldsymbol{\Sigma} \boldsymbol{A}, n) \quad .$$

(b) $\boldsymbol{\Sigma^{-1/2}} \boldsymbol{W} \boldsymbol{\Sigma^{-1/2}} \sim W_p(\boldsymbol{I}, n)$

(c) Für $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$ und einen festen p-dimensionalen Vektor $\boldsymbol{a}$ mit $\boldsymbol{a}^{\top} \boldsymbol{\Sigma} \boldsymbol{a} \neq 0$ gilt

$$\frac{\boldsymbol{a}^{\top} \boldsymbol{W} \boldsymbol{a}}{\boldsymbol{a}^{\top} \boldsymbol{\Sigma} \boldsymbol{a}} \sim \chi^2_n \quad .$$

(d) Für $p = 1$ sind die Elemente von $\boldsymbol{x} = (x_1, \ldots, x_n)^{\top}$ identisch verteilt nach $N_1(0, \sigma^2)$ und $\boldsymbol{x}^{\top} \boldsymbol{x} \sim W_1(\sigma^2, n)$. D.h. die Verteilung $W_1(\sigma^2, n)$ ist die gleiche Verteilung wie $\sigma^2 \chi^2_n$.

(e) Für die Diagonalelemente von $\boldsymbol{W}$ gilt $w_{ii} \sim \sigma_i^2 \chi^2_n$.

(f) Wenn $\boldsymbol{W_1} \sim W_p(\boldsymbol{\Sigma}, n_1)$ und $\boldsymbol{W_2} \sim W_p(\boldsymbol{\Sigma}, n_2)$ gilt und $\boldsymbol{W_1}$ und $\boldsymbol{W_2}$ unanhängig sind, dann ist $\boldsymbol{W_1} + \boldsymbol{W_2} \sim W_p(\boldsymbol{\Sigma}, n_1 + n_2)$.

Die Dichtefunktion der multivariaten Normalverteilung spielt in der multivariaten Statistik eine besonders große Rolle, weil die meisten Methoden bzw. Tests auf der Annahme basieren, dass die Daten einer multivariaten Normalverteilung entstammen. Eine Approximation der multivariaten Normalverteilung erhält man durch folgenden Satz:

Satz 1.6.1 (Zentraler Grenzverteilungssatz)
Für n voneinander unabhängige Beobachtungen einer Grundgesamtheit mit Mittel $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$ (endlich, nicht singulär), gilt für großes $n-p$ approximativ:

$$\sqrt{n}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}) \tag{1.30}$$

und

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \sim \chi^2_p \; . \tag{1.31}$$

Wie kann man nun überprüfen, ob die Daten einer multivariaten Normalverteilung entstammen? Ähnlich wie im univariaten Fall können Tests oder Grafiken zu einer Antwort führen.

Gängige statistische **Tests für multivariate Normalverteilung** sind der χ^2-Anpassungstest (Conover, 1980), der Kolmogoroff-Smirnoff-Test (Smirnov, 1948; Afifi und Azen, 1979) und der Shapiro-Wilks-Test (Shapiro und Wilk, 1965). Letzterer gibt i.a. die zuverlässigeren Resultate. Diese Tests werden hier nicht näher behandelt, es sei auf die zitierte Literatur verwiesen.

Wir wollen durch ein **grafisches Verfahren** die multivariate Normalverteilung beurteilen. Beim univariaten Q-Q-Plot (Hazen, 1914) werden die Quantile der empirischen Verteilung den Quantilen der Normalverteilung gegenübergestellt. Die Erweiterung zum multivariaten Fall nennt man **χ^2-Plot** oder Gamma-Plot (Easton und McCulloch, 1990), der folgendermaßen konstruiert wird:

- Man berechne für jede Beobachtung $\boldsymbol{x}_{i.}$ die quadrierten Mahalanobis-Distanzen

$$\mathrm{MD}_i^2 = (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^\top \boldsymbol{S}^{-1} (\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}}) \tag{1.32}$$

  $(i = 1, \ldots, n)$ und sortiere diese Werte $\mathrm{MD}_{(1)}^2 \leq \ldots \leq \mathrm{MD}_{(n)}^2$.

- Man berechne $\chi_{p,i/n}^2$, also die Quantile i/n der χ_p^2-Verteilung, und zeichne die Punktepaare $(\chi_{p,i/n}^2, \mathrm{MD}_{(i)}^2)$ in eine Grafik.

Liegen die Punkte in etwa auf einer Geraden, so kann man davon ausgehen, dass die Daten multivariat normalverteilt sind. Dieser grafische Test ist nur dann sinnvoll, wenn sowohl n als auch $n - p$ größer als 30 sind. Obwohl die Distanzen MD_i^2 nicht voneinander unabhängig sind oder exakt χ^2-verteilt sind, ist diese Grafik sehr informativ.

Beispiel 1.6.1 *Wir betrachten die 4-dimensionalen Plattendaten aus Tabelle 1.2. Der entsprechende χ^2-Plot ist in Abbildung 1.6 dargestellt. Man erkennt, dass die Ausreißer (Objekte 9 und 16) mit den größten Distanzen nicht gut ins Bild der multivariaten Normalverteilung passen.*

1.7 Transformation zu Normalverteilung

Wenn festgestellt wird, dass die vorliegenden Daten nicht normalverteilt sind, dann können durch Verfahren, die diese Voraussetzung benötigen, falsche Schlüsse gezogen werden. Es gibt aber noch die Möglichkeit, die Daten so zu transformieren, sodass sie einer Normalverteilung entsprechen. Man kann dann die üblichen Verfahren, die für normalverteilte Daten gültig sind, anwenden.

Eine Transformation der Daten ist nichts anderes als eine Skalenänderung. Sind z.B. die Daten extrem rechtsschief, kann eine Transformation mit dem Logarithmus Symmetrie um das Mittel sowie annähernde Normalverteilung bewirken.

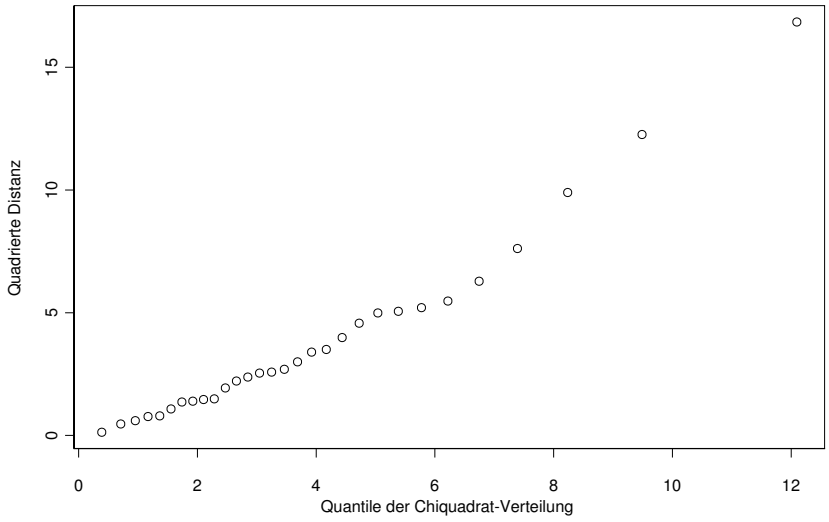


Abbildung 1.6: χ^2-Plot für die Plattendaten aus Tabelle 1.2.

Durch theoretische Überlegungen kann man je nach Datentyp folgende Transformationen empfehlen:

Ursprüngliche Skalierung	*Transformation*
Zähldaten y	$\sqrt{y}$
Verhältnisse, Anteile $\hat{p}$	$\text{logit}(\hat{p}) = \frac{1}{2}\log\left(\frac{\hat{p}}{1-\hat{p}}\right)$
Korrelationen r	Fisher's $z(r) = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right)$

Oft kann man nur direkt anhand der Daten entscheiden, welche Transformation am besten geeignet ist. Eine Vielfalt von Möglichkeiten bietet die *Power-Transformation*: Sei x eine Beobachtung (nur positive Werte) und λ ein Parameter (der auch negativ sein kann). Dann ist die Power-Transformation definiert durch x^λ. Für $\lambda = 0$ wird $x^0 = \ln x$ definiert. Um die entsprechende Potenz (Power) zu finden, können das Histogramm oder ein Q-Q-Plot betrachtet werden. Ergibt sich gute Symmetrie um den Mittelwert und entspricht die Form der Verteilung am ehesten der Normalverteilung, so hat man die optimale Transformation gefunden.

Eine leicht modifizierte Transformation wird von Box und Cox (1964) vorgeschlagen:

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases} \tag{1.33}$$

Diese Transformation ist definiert für reelles λ und $x > 0$. Sind Beobachtungen $x_1, \ldots, x_n$ gegeben, ist die Box-Cox-Lösung für die richtige Wahl der Potenz λ die

Lösung, die den Ausdruck (Likelihood-Funktion)

$$l(\lambda) = -\frac{n}{2} \ln \left[\frac{1}{n} \sum_{i=1}^{n} (x_i^{(\lambda)} - \overline{x^{(\lambda)}})^2 \right] + (\lambda - 1) \sum_{i=1}^{n} \ln x_i \tag{1.34}$$

maximiert. Dabei ist $x_i^{(\lambda)}$ in (1.33) definiert, und

$$\overline{x^{(\lambda)}} = \frac{1}{n} \sum_{i=1}^{n} x_i^{(\lambda)} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i^{\lambda} - 1}{\lambda} \right) \tag{1.35}$$

ist das arithmetische Mittel der transformierten Beobachtungen.

$l(\lambda)$ kann mittels Computer für viele Werte von λ berechnet werden, und mit einer Grafik der beiden Werte kann das Maximum leicht ermittelt werden. Meist entscheidet man sich dann für "schöne" Werte, d.h. Werte, die in der Nähe von $\hat{\lambda}$ liegen, wie z.B. $\lambda = 0$ (Logarithmus) oder $\lambda = \frac{1}{2}$ (Wurzel).

Obige Transformationen waren eigentlich für den univariaten Fall, also für *eine* Variable beschrieben. Im mehrdimensionalen Fall (für p Variablen) wird für jede einzelne Variable eine optimale Transformation durchgeführt. Entschließt man sich z.B. dazu, jede Variable der Box-Cox-Transformation zu unterziehen, so muss man die Potenzen $\lambda_1, \ldots, \lambda_p$ durch maximieren einer Likelihoodfunktion für jedes einzelne λ_k suchen.

Es sei noch erwähnt, dass selbst die "beste" Transformation nicht notwendigerweise bewirken muss , dass die Daten nun tatsächlich normalverteilt sind. Eine solche Transformation bewirkt lediglich eine möglichst gute Annäherung an die Normalverteilung.

1.8 Tests, Konfidenzbereiche

Analog zum eindimensionalen Fall können auch bei multivariaten Daten statistische Tests durchgeführt werden. Bekannte Tests wie z.B. Tests auf den Mittelwert können leicht auf den multivariaten Fall übertragen werden. Der Unterschied ist nun, dass die p korrelierten Variablen *gemeinsam* analysiert werden.

Möchte man im univariaten Fall testen, ob das Mittel μ einer Grundgesamtheit gleich einem bestimmten Wert μ_0 ist, d.h. wir testen die Nullhypothese $H_0 : \mu = \mu_0$ gegen die Alternative $H_1 : \mu \neq \mu_0$, so lautet die entsprechende Teststatistik (bei unbekannter Varianz)

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \ , \tag{1.36}$$

die t-verteilt ist mit $n-1$ Freiheitsgraden (unter der Voraussetzung, dass eine Stichprobe einer normalverteilten Grundgesamtheit vorliegt).

Wir betrachten nun im multivariaten Fall eine Stichprobe einer normalverteilten Grundgesamtheit, d.h. $\boldsymbol{x}_{1.}, \ldots, \boldsymbol{x}_{n.} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Ein Test $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ kann durchgeführt werden mit der Teststatistik

$$T^2 = (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0)^{\top} \left(\frac{1}{n} \boldsymbol{S} \right)^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}_0) \ , \tag{1.37}$$

wobei

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i.} \ , \quad \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^\top \text{ und } \boldsymbol{\mu}_0 = (\mu_{10}, \ldots, \mu_{p0})^\top \ .$$

Diese Teststatistik wird mit *Hotelling's* T^2 bezeichnet (nach dem berühmten Statistiker Harold Hotelling), sie ist verteilt nach

$$T^2 \sim \frac{(n-1)p}{(n-p)} F_{p,n-p} \ , \tag{1.38}$$

also F-verteilt mit p und $n-p$ Freiheitsgraden. Somit gilt für ein Fehlerniveau α:

$$\alpha = P\left[T^2 > \frac{(n-1)p}{(n-p)} F_{p,n-p;1-\alpha}\right] \tag{1.39}$$

Im univariaten Fall kann zu einem Test zum Niveau α unmittelbar ein *Konfidenzintervall* konstruiert werden. Analoges gilt auch im multivariaten Fall, nur dass man hier einen *Konfidenzbereich* erhält. Anhand des obigen Tests kann man auf einfache Weise einen $100(1-\alpha)\%$ Konfidenzbereich für das Mittel einer p-dimensionalen normalverteilten Größe definieren. Es ist dies ein Ellipsoid, das bestimmt ist für alle $\boldsymbol{\mu}$, die

$$n(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^\top \boldsymbol{S}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \leq \frac{p(n-1)}{(n-p)} F_{p,n-p;1-\alpha} \tag{1.40}$$

erfüllen.

Literatur

A.A. Afifi and S.P. Azen. *Statistical Analysis. A Computer Oriented Approach.* Acad. Press, New York, 1979.

T.W. Anderson. Asymptotic theory for principal component analysis. *Ann. Math. Statist.*, 34:122–148, 1963.

R.A. Becker, W.S. Cleveland, and A.R. Wilks. Dynamic graphics for data analysis. *Statistical Science*, 2(4):355–395, 1987.

G.E.P. Box and D.R. Cox. An analysis of transformations (with discussion). *J. Roy. Statist. Soc. B*, 26(2):211–252, 1964.

W.R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications.* John Wiley & Sons, New York, 1984.

G.S. Easton and R.E. McCulloch. A multivariate generalization of quantile-quantile plots. *Journal of the American Statistical Association*, 85(410):376–386, 1990.

B. Everitt. *Graphical Techniques for Multivariate Data.* North-Holland, New York, 1978.

D.M. Hawkins, editor. *Topics in Applied Multivariate Analysis.* Cambridge University Press, Cambridge, 1982.

A. Hazen. Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers*, 77:1529–1669, 1914.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, London, 4th edition, 1998.

A.M. Kshirsagar. *Multivariate Analysis*. M. Dekker, New York, 1972.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Acad. Press, London, 1979.

P.J. Rousseeuw, I. Ruts, and J.W. Tukey. The bagplot: A bivariate boxplot. *The American Statistician*, 53(4):382–387, 1999.

G.A.F. Seber. *Multivariate observations*. John Wiley & Sons, New York, 1984.

S.S. Shapiro and M.B. Wilk. An analysis of variance test for normality. *Biometrika*, 52(3):591–611, 1965.

N.V. Smirnov. Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics*, 19:279–281, 1948.

J.W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, Mass., 1977.

Kapitel 2

Clusteranalyse

2.1 Einleitung

Die Clusteranalyse bietet die Möglichkeit, Strukturen (Klassen) in einer Menge von Objekten (ev. auch Variablen) zu erkennen. Nimmt man an, dass eine Menge von n Objekten derart strukturiert ist, dass sie in mehrere Gruppen (Cluster, Klassen) zerfällt, so lassen sich mittels der Clusteranalyse diese Gruppen festlegen. Die Gruppenzugehörigkeiten der Objekte sollen dabei wesentlich durch den Grad der „Ähnlichkeit“ der Objekte festgelegt werden, d.h. Objekte eines Clusters sollen einander ähnlich sein (Homogenität innerhalb der Klassen), Objekte verschiedener Cluster sollen einander unähnlich sein (Heterogenität zwischen den Klassen).

Als *Distanz- oder Unähnlichkeitsmaß* zwischen dem i-ten und j-ten Objekt kann z.B. die *Euklidische Distanz*

$$d(i,j) = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} = \|\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\|_2 \tag{2.1}$$

oder die *Manhattan-Distanz*

$$d(i,j) = \sum_{k=1}^{p}|x_{ik} - x_{jk}| = \|\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\|_1 \tag{2.2}$$

gewählt werden. Ausgehend von einer Distanzmatrix können verschiedene Clustermethoden angewandt werden.

2.2 Klassifikationstyp

Es kommen im wesentlichen vier Klassifikationstypen in Frage: Überdeckung, Partition, Quasihierarchie und Hierarchie. Bei allen vier Typen, die in Folge beschrieben werden, kann man zwischen exhaustiver und nicht-exhaustiver Klassifikation unterscheiden. Bei einer *exhaustiven Klassifikation* wird jedes Objekt klassifiziert, bei einer *nicht-exhaustiven Klassifikation* können auch einzelne Objekte nicht den Klassen zugeteilt werden.

Eine **Überdeckung** ist eine Klassifikation, bei der sich einzelne Klassen überschneiden dürfen. Es darf dabei aber nicht eine Klasse vollständig in einer anderen enthalten sein.

Eine **Partition** ist eine spezielle Überdeckung, bei der verlangt wird, dass kein Objekt zu mehr als einer Klasse gehört. D.h. die Klassen dürfen sich nicht überschneiden und sind somit disjunkt.

Nachdem Überdeckungen und Partitionen die Struktur einer Objektmenge oft nur sehr grob wiedergeben, wählt man manchmal die feineren Klassifikationstypen Quasihierarchie und Hierarchie.

Eine **Quasihierarchie** ist eine Klassifikation, die durch eine Folge von Überdeckungen gebildet wird. Man erhält somit eine Art „Stammbaum“ mit einander überschneidenden Klassen, dessen unterste Stufe die gröbste und dessen oberste Stufe die feinste Überdeckung der Quasihierarchie ist.

Eine **Hierarchie** ist analog zur Quasihierarchie definiert, nur müssen hier die Klassen disjunkt sein. Die Hierarchie wird somit durch eine Folge von Partitionen bestimmt. Mit Hilfe eines „Stammbaums“ können die einzelnen Stufen der Partitionierung dargestellt werden. Eine andere Möglichkeit zur grafischen Veranschaulichung bietet das *Dendrogramm* (siehe später).

2.3 Bewertungskriterien für Klassifikationen

Bei den verschiedenen Verfahren der Clusteranalyse möchte man hohe Homogenität (Gleichartigkeit) innerhalb der Klassen, gleichzeitig aber hohe Heterogenität (Verschiedenheit) zwischen den Klassen erreichen. Aus diesem Grund werden verschiedene Gütemaße für Homogenität und Heterogenität angegeben.

2.3.1 Maße für die Homogenität einer Klasse

Gegeben sei eine Klasse K_l $(l = 1, \ldots, k)$ aus einer Klassifikation κ. Die Homogenität dieser Klasse wird mit einem Homogenitätsmaß $h(K_l) \geq 0$ gemessen. Die resultierende Zahl wird umso kleiner, je homogener die Klasse K_l ist.

Die Homogenität einer Klasse kann mit Hilfe einer *Distanzmatrix* gemessen werden. Eine Distanzmatrix enthält die Distanzen oder Unähnlichkeiten (engl. *dissimilarities*) aller Paare von Objekten (siehe oben).

- Summe der Distanzen:

$$h(K_l) = \frac{1}{|K_l|} \sum_{\substack{i<j \\ i,j \in K_l}} d(i,j)$$

  Die Normierungskonstante $|K_l|$ bezeichnet die Anzahl der Objekte in der Klasse K_l.

- Maximum der Distanzen:

$$h(K_l) = \max_{i,j \in K_l} d(i,j)$$

Dieses Maß entspricht den beiden unähnlichsten Objekten einer Klasse und ist somit relativ streng. Große Klassen werden schlecht beurteilt.

- Minimum der Distanzen:

$$h(K_l) = \min_{i,j \in K_l} d(i,j)$$

Große heterogene Klassen haben mit diesem Maß oft einen sehr kleinen Wert.

- Summe der Varianzen der Variablen:

Hier geht man nicht von der Distanzmatrix, sondern direkt von der $(n \times p)$-Datenmatrix $\boldsymbol{X} = [(x_{ij})]$ aus.

Das Mittel der j-ten Variable $(j = 1, \ldots, p)$ in der Klasse K_l ist definiert als

$$\bar{x}_j = \frac{1}{|K_l|} \sum_{i \in K_l} x_{ij} \ .$$

Somit ist die Varianz der j-ten Variable in der Klasse K_l

$$s_j^2(K_l) = \frac{1}{|K_l| - 1} \sum_{i \in K_l} (x_{ij} - \bar{x}_j)^2 \ .$$

Als Homogenitätsmaß nimmt man die Summe dieser Varianzen

$$h(K_l) = \sum_{j=1}^{p} s_j^2(K_j) \ .$$

Für einelementige Mengen ist dieses Maß nicht verwendbar.

2.3.2 Maße für die Heterogenität zwischen den Klassen

Gegeben seien zwei Klassen K_{l_1} und K_{l_2} aus einer Klassifikation. Ein Heterogenitätsmaß $v(K_{l_1}, K_{l_2})$ soll nichtnegative Werte annehmen, die kleiner sind, je ähnlicher sich die beiden Klassen sind. Man verlangt außerdem Symmetrie, also $v(K_{l_1}, K_{l_2}) = v(K_{l_2}, K_{l_1})$.

Bei den Heterogenitätsmaßen muss man unterscheiden, ob die Klassen disjunkt sind oder ob sie sich überschneiden. Das Maß hängt also vom Klassifikationstyp ab.

Disjunkte Klassen:

Wir gehen wiederum von einer Distanzmatrix aus. Bekannte Maße für die Heterogenität zweier disjunkter Klassen K_{l_1} und K_{l_2} sind:

- Complete Linkage:

$$v(K_{l_1}, K_{l_2}) = \max_{i \in K_{l_1}, j \in K_{l_2}} d(i,j)$$

Die Verschiedenheit wird durch das unähnlichste Objektpaar bestimmt. Man erhält dadurch oft zu wenig Cluster.

- Single Linkage:
$$v(K_{l_1}, K_{l_2}) = \min_{i \in K_{l_1}, j \in K_{l_2}} d(i,j)$$
Die Verschiedenheit wird durch das ähnlichste Objektpaar bestimmt. Man erhält dadurch oft zu viele Cluster.

- Average Linkage:
$$v(K_{l_1}, K_{l_2}) = \frac{1}{|K_{l_1}|\ |K_{l_2}|} \sum_{i \in K_{l_1}} \sum_{j \in K_{l_2}} d(i,j)$$
Die Verschiedenheit wird durch die durchschnittliche Ähnlichkeit der beiden Klassen bestimmt. Dieses Verfahren ist somit ein Kompromiss zwischen den beiden oben beschriebenen Methoden.

- Centroid Methode:
Hier geht man nicht von der Distanzmatrix, sondern direkt von der $(n \times p)$-Datenmatrix $\boldsymbol{X} = [(x_{ij})]$ aus.
Man berechnet den euklidischen Abstand zwischen den Zentren (centroid), die definiert sind als
$$\bar{\boldsymbol{x}}(K_{l_m}) = \frac{1}{|K_{l_m}|} \left(\sum_{i \in K_{l_m}} x_{i1}, \ldots, \sum_{i \in K_{l_m}} x_{ip} \right)^\top \quad \text{für } m = 1, 2\ ,$$
das entspricht den Mittelwertvektoren für die p Variablen in den Klassen K_{l_1} und K_{l_2}. Das Heterogenitätsmaß ist dann
$$v(K_{l_1}, K_{l_2}) = \|\bar{\boldsymbol{x}}(K_{l_1}) - \bar{\boldsymbol{x}}(K_{l_2})\|_2\ .$$

Überschneidende Klassen:

Obige Heterogenitätsmaße werden so modifiziert, dass die Verschiedenheit der um die identischen Objekte reduzierten Klassen berechnet wird.

Ist eine Klasse vollständig in einer anderen enthalten (Hierarchie, Quasihierarchie), so wird die größere Klasse um die in ihr enthaltenen reduziert.
Beispiel: $K_{l_1} = \{8, 10\}$, $K_{l_2} = \{2, 7, 8, 10, 12\} \Longrightarrow K^*_{l_2} = K_{l_2} - K_{l_1} \cap K_{l_2} = \{2, 7, 12\} \Longrightarrow v(K_{l_1}, K^*_{l_2})$

2.3.3 Maße für die Güte einer Klassifikation

Durch Maße für Homogenität und Heterogenität werden die einzelnen Klassen bzw. die Klassenpaare einer Klassifikation bewertet. Mit einem Gütemaß möchte man die Klassifikation selbst bewerten.

Das Gütemaß wird umso kleiner, je „besser“ die Klassifikation ist. Dies hängt aber wiederum vom Klassifikationstyp ab.

Als Gütemaß für **Überdeckungen** könnte man die Summe aller Klassenhomogenitäten minus der Summe aller Klassenhomogenitäten aller Teilmengen von Objekten, die gleichzeitig mehreren Klassen angehören, heranziehen. Nachdem dieses Maß nur von den Homogenitäten abhängt, könnte man diese Größe noch durch ein Maß für die Heterogenitäten dividieren.

Bei **Partitionen** könnte man die Summe der Klassenhomogenitäten als Gütemaß verwenden. Eine andere Möglichkeit ist ein normierter Kehrwert der Heterogenitäten oder der Quotient von Homogenitäten und Heterogenitäten.

Gütemaße für **Hierarchien** und **Quasihierarchien** werden für jede Stufe in der (Quasi-)Hierarchie berechnet, wobei obige Maße verwendet werden können.

2.4 Konstruktionsverfahren

Die Verfahren zur Konstruktion einer Clusterung hängen natürlich vom Klassifikationstyp ab. Aus Platzgründen wollen wir in Folge nur Verfahren für die gängigeren Typen Partition und Hierarchie betrachten.

2.4.1 Partitionen (Partitionierungmethoden)

Ein iteratives Verfahren für große Objektmengen

Die Klassenzahl k muss vorgegeben werden, wobei man dieses Verfahren ev. mit verschiedener Klassenzahl wiederholen kann. Man wählt eine Anfangspartition, d.h. k Objekte werden zufällig ausgewählt (Zentralobjekte). Die restlichen $n-k$ Objekte werden jeweils dem Zentralobjekt zugeordnet, dem sie am ähnlichsten sind (kleinste Distanz). Die Anfangspartition ist jetzt komplett.

Dann wird für jedes Objekt, das nicht zu einer einelementigen Klasse gehört, geprüft, ob die Güte verkleinert wird, wenn das Objekt in eine der anderen Klassen transferiert wird. Ist dies möglich, so wird jenes Objekt transferiert, das die größte Abnahme der Güte bewirkt. Es wird so lange iteriert, bis keine wesentliche Verbesserung der Güte möglich ist. Diese Methode kann ev. mit verschiedenen Anfangspartitionen durchgeführt werden.

Bei der **k-means-Methode** werden, ausgehend von der Anfangspartition, so lange die Objekte in andere Cluster transferiert, bis ein „Fehlermaß“ nicht mehr reduziert werden kann. Ein solches Fehlermaß ist z.B. die Summe über alle Objekte von der quadrierten euklidischen Distanz jedes Punktes zu seinem Clustercentroid, das jedesmal neu berechnet werden muss.

Ein rekursives Verfahren für große Objektmengen

Hier werden die Klassen nacheinander bestimmt, die Klassenanzahl braucht somit nicht vorgegeben werden.

Als Zentralobjekt der ersten Klasse wird eines jener beiden Objekte mit minimalem Abstand zueinander gewählt. Alternativ kann jenes Objekt i mit größter „Punktdichte“ gewählt werden, d.h. die Anzahl der Objekte j mit $d(i,j) \leq \tilde{d}$ soll maximiert werden. Ausgehend von diesem Kernobjekt wird jenes Objekt in die Klasse

genommen, das minimalen Abstand (größte Ähnlichkeit) zum Kernobjekt besitzt. Dies wird solange fortgeführt, bis eine Homogenitätsschranke überschritten wird oder die Homogenität von einem Schritt zum nächsten radikal vergrößert wird. Aus der Restmenge wird, analog zu oben, wieder ein Kernobjekt gewählt und Objekte adjungiert. Das Verfahren wird abgebrochen, wenn alle Objekte klassifiziert sind (exhaustiv) oder wenn nur noch sehr kleine Gruppen auftreten.

2.4.2 Hierarchie

Hierarchische Verfahren werden durch eine Folge von Partitionen (disjunkte Klassen) gebildet. Die Konstruktion kann divisiv (von einer groben Partition schrittweise immer feinere erzeugen) oder agglomerativ erfolgen. Divisive Verfahren sind sehr rechenaufwendig und daher wenig gebräuchlich.

Agglomerative Methoden:

Ausgehend von n einelementigen Klassen (Anfangspartition) werden in jedem Schritt jene beiden Klassen zusammengefasst, die minimal verschieden sind, bis nur noch eine Klasse übrigbleibt. Zur Feststellung, welche Klassen größte Ähnlichkeit (kleinste Verschiedenheit) haben, wählt man ein Distanzmaß (siehe Maße für Heterogenität disjunkter Klassen) wie z.B. **single linkage**, **complete linkage**, **average linkage**, **centroid method**. Jene Klassen mit minimalem Distanzmaß werden vereint.

Single linkage erzeugt viele (kleine) Klassen (Ketteneffekt) und ermöglicht eine Erkennung stark verzweigter Klassen. Nachteil ist, dass oft zwei unterschiedliche Klassen nur dadurch vereint werden, weil zwei Objekte dieser Klassen geringen Abstand haben. Große Klassen werden früh vereint.

Complete linkage führt bei inhomogenen Klassen zu schlechten Gruppierungen. Man erhält oft wenige relativ große Klassen.

Agglomerative Verfahren können beendet werden, wenn nur noch eine Gruppe übrig bleibt, oder wenn eine Heterogenitätsschranke überschritten wird.

Da die Distanzen der Gruppen in jedem Schritt größer werden (Monotonie), können die Partitionen der Hierarchie grafisch in einem **Dendrogramm** dargestellt werden. Bei Fusionierung zweier Klassen kann das Distanzmaß (Heterogenitätsmaß) angegeben werden, das als Gütemaß dient.

Beispiel 2.4.1 *Von 88 Studenten wurden die Prüfungsergebnisse in den Fächern Mechanik, Analytische Geometrie, Lineare Algebra, Analysis und Elementare Statistik aufgezeichnet. Von den 100 erreichbaren Punkten gab es die in Tabelle 2.1 angeführten Ergebnisse. Die 88 Studenten sollen anhand ihrer Leistungen in den fünf Prüfungsfächern geclustert werden. Die Clusterung wird mit complete linkage sowie mit single linkage durchgeführt. Die Resultate sind mit Hilfe von* Dendrogrammen *in Abbildung 2.1 und 2.2 dargestellt.*

Tabelle 2.1: Prüfungsergebnisse von 88 Studenten in den Fächern Mechanik (ME), Analytische Geometrie (AG), Lineare Algebra (LA), Analysis (AN) und Elementare Statistik (ES); 100 Punkte waren erreichbar.

Student	ME	AG	LA	AN	ES	Student	ME	AG	LA	AN	ES
1	77	82	67	67	81	45	46	61	46	38	41
2	63	78	80	70	81	46	40	57	51	52	31
3	75	73	71	66	81	47	49	49	45	48	39
4	55	72	63	70	68	48	22	58	53	56	41
5	63	63	65	70	63	49	35	60	47	54	33
6	53	61	72	64	73	50	48	56	49	42	32
7	51	67	65	65	68	51	31	57	50	54	34
8	59	70	68	62	56	52	17	53	57	43	51
9	62	60	58	62	70	53	49	57	47	39	26
10	64	72	60	62	45	54	59	50	47	15	46
11	52	64	60	63	54	55	37	56	49	28	45
12	55	67	59	62	44	56	40	43	48	21	61
13	50	50	64	55	63	57	35	35	41	51	50
14	65	63	58	56	37	58	38	44	54	47	24
15	31	55	60	57	73	59	43	43	38	34	49
16	60	64	56	54	40	60	39	46	46	32	43
17	44	69	53	53	53	61	62	44	36	22	42
18	42	69	61	55	45	62	48	38	41	44	33
19	62	46	61	57	45	63	34	42	50	47	29
20	31	49	62	63	62	64	18	51	40	56	30
21	44	61	52	62	46	65	35	36	46	48	29
22	49	41	61	49	64	66	59	53	37	22	19
23	12	58	61	63	67	67	41	41	43	30	33
24	49	53	49	62	47	68	31	52	37	27	40
25	54	49	56	47	53	69	17	51	52	35	31
26	54	53	46	59	44	70	34	30	50	47	36
27	44	56	55	61	36	71	46	40	47	29	17
28	18	44	50	57	81	72	10	46	36	47	39
29	46	52	65	50	35	73	46	37	45	15	30
30	32	45	49	57	64	74	30	34	43	46	18
31	30	69	50	52	45	75	13	51	50	25	31
32	46	49	53	59	37	76	49	50	38	23	9
33	40	27	54	61	61	77	18	32	31	45	40
34	31	42	48	54	68	78	8	42	48	26	40
35	36	59	51	45	51	79	23	38	36	48	15
36	56	40	56	54	35	80	30	24	43	33	25
37	46	56	57	49	32	81	3	9	51	47	40
38	45	42	55	56	40	82	7	51	43	17	22
39	42	60	54	49	33	83	15	40	43	23	18
40	40	63	53	54	25	84	15	38	39	28	17
41	23	55	59	53	44	85	5	30	44	36	18
42	48	48	49	51	37	86	12	30	32	35	21
43	41	63	49	46	34	87	5	26	15	20	20
44	46	52	53	41	40	88	0	40	21	9	14

Abbildung 2.1: Clusterung der Prüfungsdaten (*Complete Linkage*)

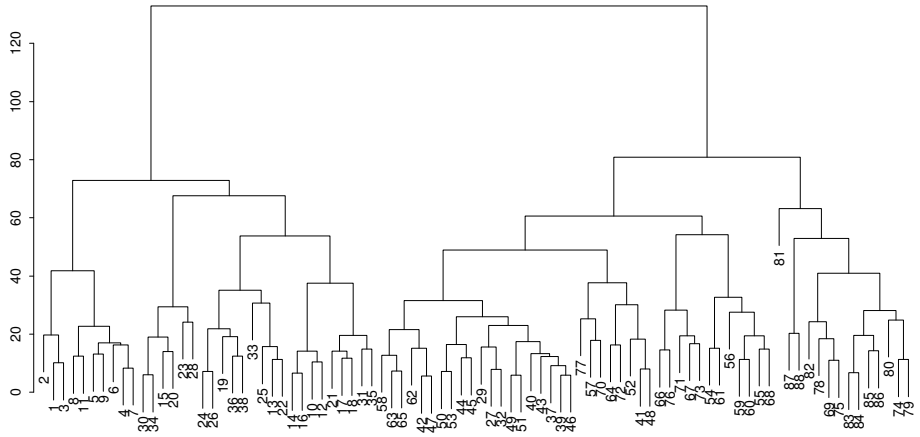
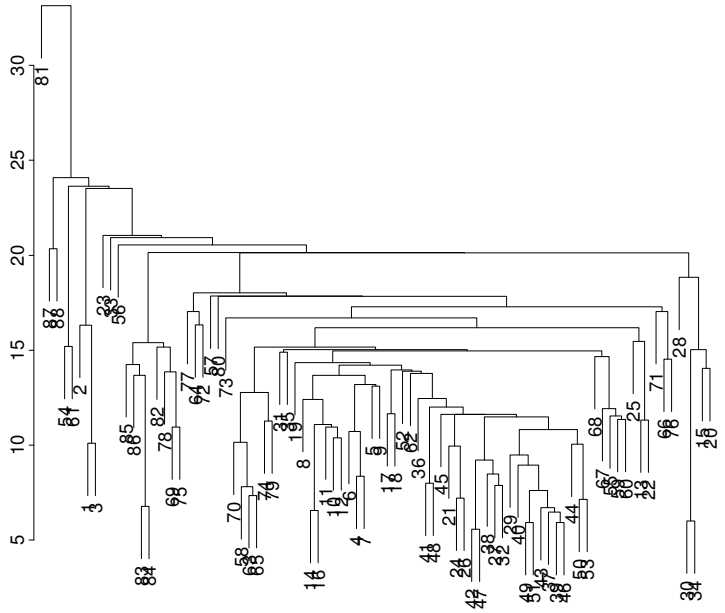


Abbildung 2.2: Clusterung der Prüfungsdaten (*Single Linkage*)



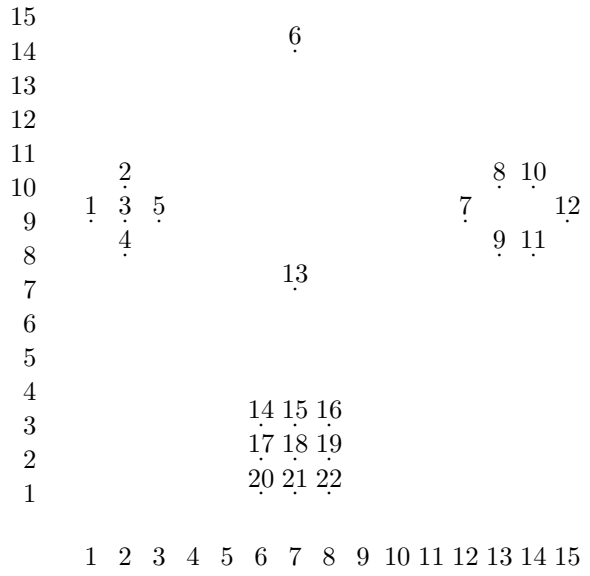
2.5 Fuzzy Clusterung

Bei den bisher besprochenen Methoden wurde jedes Objekt (exhaustiv) genau einem Cluster zugeordnet (man spricht von *harter* Clusterung). Bei der *Fuzzy*-Clusterung wird jedes Objekt auf alle Cluster „prozentuell“ aufgeteilt, wobei diese Aufteilung je nach Ähnlichkeit zum Cluster erfolgt. Die Anzahl der Cluster muss hier vorgegeben

werden.

Daten wie in Abbildung 2.3 können durch die Fuzzy-Clusterung wesentlich besser beschrieben werden als durch Partitionierungsmethoden.

Abbildung 2.3: Datensatz mit zwei Objekten, die nicht eindeutig einem bestimmten Cluster zugeordnet werden können



Dieses Beispiel lässt deutlich drei Cluster erkennen sowie die Objekte 6 und 13, die jedoch nicht eindeutig einem Cluster zugeordnet werden können.

Diese Methode kann aussagen, dass z.B. Objekt 1 fast sicher zu Cluster 1 gehört, wohingegen Objekt 13 mit praktisch gleicher Chance einem der drei Cluster zugeteilt werden sollte.

Jener Koeffizient, der angibt, zu welchem Teil ein Objekt zu einem bestimmten Cluster gehört, hat einen Wert zwischen 0 und 1 (0-100%) und heißt *Zugehörigkeitskoeffizient*.

Die Zugehörigkeitskoeffizienten zu den Daten aus Abbildung 2.3 sind in Tabelle 2.2 dargestellt. Wie aus diesem Beispiel auch ersichtlich ist, beträgt die Summe der Zugehörigkeitskoeffizienten eines Objekts 1 (100%). Der wesentliche Unterschied dieser Methode zur harten Clusterung liegt darin, dass sie detailliertere Information über die Struktur der Daten liefert. Der Nachteil dabei ist, dass mit wachsender Anzahl der Objekte der Ausdruck der Ergebnisse unüberschaubar wird und so die erhaltene Information oft nicht verarbeitet werden kann.

Ein möglicher Algorithmus wäre folgender: Es gibt keine repräsentativen Objekte, stattdessen wird versucht, die Objektfunktion

$$\sum_{v=1}^{k} \frac{\sum_{i,j=1}^{n} u_{iv}^2 u_{jv}^2 d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^2} \tag{2.3}$$

Tabelle 2.2: Zugehörigkeitskoeffizienten zu den Daten in Abbildung 6.2

Objekt	Cluster 1	Cluster 2	Cluster 3
1	0.87	0.06	0.07
2	0.88	0.05	0.07
3	0.93	0.03	0.04
4	0.86	0.06	0.08
5	0.87	0.06	0.07
6	0.42	0.35	0.23
7	0.08	0.82	0.10
8	0.06	0.87	0.07
9	0.06	0.86	0.08
10	0.06	0.87	0.07
11	0.06	0.86	0.08
12	0.07	0.84	0.09
13	0.36	0.27	0.37
14	0.12	0.08	0.80
15	0.08	0.07	0.85
16	0.10	0.10	0.80
17	0.08	0.06	0.86
18	0.04	0.04	0.92
19	0.07	0.07	0.86
20	0.10	0.08	0.82
21	0.07	0.06	0.87
22	0.09	0.09	0.82

zu minimieren. Sie enthält nur das Ähnlichkeitsmaß $d(i, j)$ und den gesuchten Zugehörigkeitskoeffizienten u_{iv} des Objekts i zum Cluster v. Die Summe im Zähler geht über alle Objektpaare $\{i, j\}$ (anstelle der Distanzsumme der Objekte zu den Clustermittelpunkten bei anderen Verfahren). Da doppelt aufsummiert wird ($\{i, j\} = \{j, i\}$), wird durch 2 dividiert. Die äußere Summe geht über alle Cluster v, wodurch die zu minimierende Objektfunktion ein Maß für die Streuung ist.

Verwandte Methoden

Als eine der ersten Fuzzy-Clustermethoden wurde die *fuzzy k-means*-Methode entwickelt (siehe Bezdek, 1974, Dunn, 1974). Diese Methode minimiert die Objektfunktion

$$\sum_i \sum_v u_{iv}^2 \|\boldsymbol{x}_{i.} - \boldsymbol{m}_v\|_2^2 = \sum_i \sum_v u_{iv}^2 \sum_{j=1}^{p} (x_{ij} - m_{vj})^2 \quad . \tag{2.4}$$

Dabei ist $\boldsymbol{m}_v$ das Zentrum des Clusters v, und $\boldsymbol{x}_{i.}$ bezeichnet das i-te Objekt. Für jede Variable (Index j) wird das Zentrum folgendermaßen berechnet:

$$m_{vj} = \frac{\sum_i u_{iv}^2 x_{ij}}{\sum_i u_{iv}^2} \quad . \tag{2.5}$$

Die Norm $\|\boldsymbol{x}_{i.} - \boldsymbol{m}_v\|_2$ ist die euklidische Distanz zwischen Objekt $\boldsymbol{x}_{i.}$ und dem Zentrum des Clusters v. Wesentlich bei dieser Methode ist die Annahme, dass die verschiedenen Objekte durch ihre Koordinaten im p-dimensionalen Raum gegeben sind. Das ist eine Einschränkung im Vergleich zu der oben beschriebenen Methode, die eine solche Forderung nicht benötigt, sondern nur Distanzen oder Unähnlichkeiten zwischen den Objekten voraussetzt. Die beiden Methoden können, falls die Daten aus Messungen bestehen, direkt verglichen werden. Durch Umformungen lässt sich zeigen, dass folgende Gleichheit gilt:

$$\sum_i \sum_v u_{iv}^2 \sum_j (x_{ij} - m_{vj})^2 = \sum_v \frac{\sum_i \sum_j u_{iv}^2 u_{jv}^2 \|\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\|_2^2}{2 \sum_j u_{jv}^2} \quad . \tag{2.6}$$

Der letzte Ausdruck ist genau die Objektfunktion der oben beschriebenen Methode, mit dem Unterschied, dass die Distanzen quadriert sind. Das veranschaulicht die Äquivalenzen zwischen der oben beschriebenen und der *fuzzy k-means*-Methode, wenn die Objekte durch Messungen beschrieben werden. (Als Eingabe für die erste Methode wird eine Distanzmatrix, in der die Quadrate der Euklidischen Abstände zwischen den Objekten eingetragen werden, benötigt.)

Eine andere Methode ist der *MND2-Algorithmus* (siehe Roubens, 1978). Dabei wird folgende Objektfunktion minimiert:

$$\sum_v \sum_{i,j} u_{iv}^2 u_{jv}^2 d(i,j) \quad . \tag{2.7}$$

Auch dieser Ausdruck ist der Objektfunktion der ersten Methode ähnlich. Im Nenner der Objektfunktion der ersten Methode steht der Ausdruck $\sum_j u_{jv}^2$. Dieser Term ist der wesentliche Unterschied zwischen den beiden Objektfunktionen.

Literatur

J.C. Bezdek. Cluster validity with fuzzy sets. *Cybernetics*, 3:58–72, 1974. Scripta Publ. Comp., Washington, D.C.

H.H. Bock. Automatische Klassifikation. In K.P. Grotemeyer, D. Morgenstern, and H. Tietz, editors, *Studia Mathematica/Mathematische Lehrbücher*, volume XXIV. Vandenhoeck & Ruprecht, Göttingen, 1974.

J.C. Dunn. A fuzzy relative of the isodata progress and its use in detecting compact well-separated clusters. *Cybernetics*, 3:32–57, 1974. Scripta Publ. Comp., Washington, D.C.

J.A. Hartigan. *Clustering algorithms*. Wiley & Sons, New York, 1975.

L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data.* Wiley & Sons, New York, 1990.

M. Roubens. Pattern classification problems and fuzzy sets. *Fuzzy Sets and Systems*, 1:239–253, 1978.

Kapitel 3

Multivariate lineare Regression

3.1 Einführung

Die multivariate Regressionsanalyse ist ein Instrumentarium zur Vorhersage von Werten einer oder mehrerer quantitativer Größen, den abhängigen Variablen, durch bekannte (unabhängige) Größen. Es wird versucht, einen funktionellen Zusammenhang (der hier als linear vorausgesetzt ist) zwischen den abhängigen und den unabhängigen Größen zu finden.

Wir werden zunächst den Fall betrachten, wo nur eine einzige abhängige Variable vorliegt. Später wird dieses Modell erweitert für mehrere abhängige Variablen.

3.2 Lineare multiple Regression

Seien $x_1, \ldots, x_q$ die q Variablen, die zur Vorhersage einer abhängigen Variable Y vorliegen. Z.B. könnte die abhängige Variable der Preis eines neuen Autos sein, die unabhängigen Größen könnten Hubraum, Leistung, Gewicht, Luxuskategorie etc. sein. Das lineare multiple Regressionsmodell sagt aus, dass Y dargestellt werden kann durch ein Mittel, das kontinuierlich von x_j abhängt, sowie von einem zufälligen Fehler ε. Dieser Fehlerterm beinhaltet sowohl Messfehler als auch den Einfluss anderer Variablen, die in diesem Modell nicht berücksichtigt wurden. Die Werte der Vorhersagevariablen $x_1, \ldots, x_q$ werden als *fest* angesehen, der Fehlerterm und damit auch die abhängige Größe sind *zufällig*.

Das lineare Regressionsmodell mit *einer abhängigen Größe* hat somit die Form

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_q x_q + \varepsilon \; . \tag{3.1}$$

Die unbekannten *Regressionsparameter* $\beta_0, \beta_1, \ldots, \beta_q$ geben den funktionellen Zusammenhang an, der bei diesem Modell *linear* ist. Diese Regressionskoeffizienten können aufgrund von konkreten Beobachtungen in weiterer Folge geschätzt werden.

Sind nun n voneinander unabhängige Beobachtungen sowohl von Y als auch von den zugehörigen unabhängigen Größen $x_1, \ldots, x_q$ gegeben, so ist das gesamte Modell (**multiples Regressionsmodell**)

$$Y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \ldots + \beta_q x_{1q} + \varepsilon_1$$

$$\begin{aligned} Y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \ldots + \beta_q x_{2q} + \varepsilon_2 \\ &\vdots \quad \vdots \\ Y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \ldots + \beta_q x_{nq} + \varepsilon_n \end{aligned}$$

oder in Matrixschreibweise:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ . \tag{3.2}$$

$\boldsymbol{Y}$ ist ein n-dimensionaler Zufallsvektor. Um den Unterschied zu einem Vektor mit konkreten Beobachtungen (später mit $\boldsymbol{y}$ bezeichnet) besser hervorzuheben, ist der Zufallsvektor $\boldsymbol{Y}$ ausnahmsweise (und nur in diesem Kapitel) mit großem fettgedrucktem Symbol bezeichnet.

Die erste Spalte der *Design-Matrix* $\boldsymbol{X}$ besteht aus Einsen, da sie Multiplikator der Konstanten β_0 ist. Somit ist $\boldsymbol{X}$ eine $(n \times (q+1))$-Matrix.

Von den Fehlertermen wird folgendes vorausgesetzt:

1. $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ und
2. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top) = \sigma^2 \boldsymbol{I}_n$.

σ^2 beschreibt hier die Varianz des Fehlerterms, die für alle Komponenten gleich ist. Außerdem sind verschiedene Komponenten des Fehlerterms unkorreliert.

Zum Testen von Hypothesen bzw. für die Konstruktion von Konfidenzintervallen wird man außerdem noch gemeinsame Normalverteilung voraussetzen müssen.

3.3 Der Kleinste-Quadratsummen-Schätzer

In der Regressionsanalyse wird nun versucht, aufgrund gegebener Werte der Vorhersage-Variablen den genauen funktionellen Zusammenhang zur abhängigen Variable festzustellen. Man versucht also, aufgrund von bekannten Werten $1, x_{i1}, \ldots, x_{iq}$ das Regressionsmodell (3.2) den beobachteten Werten y_i anzupassen. Genauer gesagt, es sollen die Regressionskoeffizienten $\boldsymbol{\beta}$ und die Fehlervarianz σ^2 aufgrund der vorliegenden Daten ermittelt werden.

Die Differenz $y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_q x_{iq}$, die die Abweichung der Anpassung vom beobachteten Wert angibt, wird auch als *Residuum* bezeichnet. Bei der *Methode der Kleinsten Quadrate* soll die Summe der quadrierten Residuen minimiert werden:

$$\begin{aligned} S(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_q x_{iq})^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned} \tag{3.3}$$

Jener Koeffizient, der dieses Kriterium minimiert, wird mit $\widehat{\boldsymbol{\beta}}$ bezeichnet. Er ist der Kleinste-Quadratsummen-Schätzer des Regressionsparameters $\boldsymbol{\beta}$.

Gleichung (3.3) kann ausmultipliziert werden:

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.4}$$

Da es sich bei den Summanden um Skalare handelt, gilt

$$\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y})^\top = \boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.5}$$

Die Residuen-Quadratsumme (QS) ist somit

$$S(\boldsymbol{\beta}) = \boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{y} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.6}$$

Partielles Differenzieren nach dem Parametervektor $\boldsymbol{\beta}$ ergibt

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 - 2\boldsymbol{X}^\top \boldsymbol{y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta} \quad . \tag{3.7}$$

Das Minimum der Residuen-QS erhält man durch Nullsetzen obiger Gleichung. Der Kleinste-QS-Schätzer für $\boldsymbol{\beta}$ ist dann

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \quad . \tag{3.8}$$

Bemerkung: Hat $\boldsymbol{X}$ nicht vollen Rang $q+1 \leq n$, so muss $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$ ersetzt werden durch eine Verallgemeinerte Inverse (z.B. Moore-Penrose) $(\boldsymbol{X}^\top \boldsymbol{X})^-$.

Mit Hilfe von $\widehat{\boldsymbol{\beta}}$ können die *geschätzten Werte* $\widehat{\boldsymbol{y}}$ von $\boldsymbol{y}$ berechnet werden durch

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \ , \tag{3.9}$$

wobei $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$ als "Hat"-Matrix bezeichnet wird. Die *geschätzten Residuen* sind

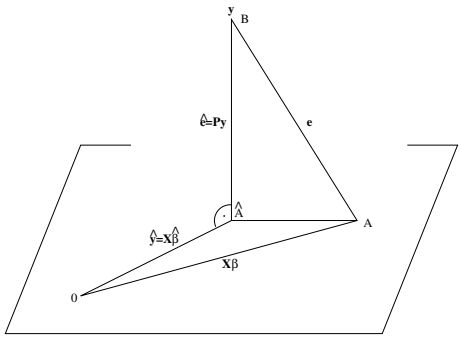
$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \ . \tag{3.10}$$

Es gilt: $\boldsymbol{X}^\top \widehat{\boldsymbol{\varepsilon}} = \boldsymbol{X}^\top (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{0}$ und $\widehat{\boldsymbol{y}}^\top \widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\beta}}^\top \boldsymbol{X}^\top \widehat{\boldsymbol{\varepsilon}} = 0$. D.h., die geschätzten Residuen sind orthogonal zu den Spalten von $\boldsymbol{X}$ und zu den geschätzten Werten $\widehat{\boldsymbol{y}}$.

Eine **geometrische Veranschaulichung** dieser Methode lässt die Hintergründe klarer erkennen: Der Erwartungswert $E(\boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{\beta}$ ist eine Linearkombination der Spalten von $\boldsymbol{X}$. $\boldsymbol{X}\boldsymbol{\beta}$ spannt also für variierendes $\boldsymbol{\beta}$ den Raum aller möglichen Linearkombinationen auf. I.a. wird der Beobachtungsvektor $\boldsymbol{y}$ nicht exakt in diesem Raum liegen, da ein zufälliger Fehler $\boldsymbol{\varepsilon}$ additiv eingeht. Sind konkrete Beobachtungen verfügbar, ist die Summe der quadrierten Residuen $S(\boldsymbol{\beta})$ so klein wie möglich, wenn $\boldsymbol{\beta}$ so gewählt wird, dass $\boldsymbol{X}\boldsymbol{\beta}$ jener Punkt im aufgespannten Raum ist, der am nähesten zu $\boldsymbol{y}$ ist. Das ist jener Punkt, der bei orthogonaler Projektion von $\boldsymbol{y}$ auf den Raum entsteht. In Abbildung 3.1 ist der „wahre" Wert von $\boldsymbol{X}\boldsymbol{\beta} = E(\boldsymbol{Y})$ mit Punkt A bezeichnet, der beobachtete Wert $\boldsymbol{y}$ von $\boldsymbol{Y}$ mit Punkt B. Den Schätzwert $\widehat{A}$ für A erhält man, indem Punkt B auf die durch die Spalten von $\boldsymbol{X}$ aufgespannte Ebene orthogonal projiziert wird. Der Residuenvektor $\widehat{\boldsymbol{\varepsilon}}$ (in der Abbildung mit $\widehat{\boldsymbol{e}}$ bezeichnet) ist orthogonal zu dieser Ebene (vgl. Resultate vorher).

Satz 3.3.1 *Für das lineare Regressionsmodell (3.2) gilt:*
Der Kleinste-QS-Schätzer $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ *ist unverzerrt, d.h.* $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, *und* $Cov(\widehat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$.
Die Residuen $\widehat{\boldsymbol{\varepsilon}}$ *haben folgende Eigenschaften:*
$E(\widehat{\boldsymbol{\varepsilon}}) = \boldsymbol{0}$ *und* $Cov(\widehat{\boldsymbol{\varepsilon}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$. $\widehat{\boldsymbol{\beta}}$ *und* $\widehat{\boldsymbol{\varepsilon}}$ *sind unkorreliert.*

Abbildung 3.1: Geometrische Darstellung des multiplen Regressionsmodells



Beweis: siehe z.B. Johnson und Wichern (1998), S. 388.

Damit hat dieser Schätzer folgende optimale Eigenschaft:

Satz 3.3.2 (Gauss-Markov)
Beim multiplen Regressionsmodell (3.2) sei vorausgesetzt, dass die Elemente des Fehlervektors unkorreliert sind, $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}_n$ (Homoskedastizität). Dann gilt:

(a) $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ *ist ein eindeutig bestimmter, effizienter, linearer Schätzer für* $\boldsymbol{\beta}$*,*

(b) $s^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^\top \widehat{\boldsymbol{\varepsilon}}}{n-q-1} = \frac{1}{n-q-1}(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})^\top (\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$ *ist ein unverzerrter Schätzer für die Residuenvarianz* σ^2*.*

Beweis: siehe z.B. Mardia et al. (1979), S. 172.

Ein effizienter Schätzer hat eine Kovarianzmatrix, die kleiner ist als die jedes anderen linearen unverzerrten Schätzers. Beim Kleinsten-QS-Schätzer spricht man auch vom *besten* linearen unverzerrten Schätzer (BLUE).

Satz 3.3.3 *Gilt zu den Voraussetzungen des Gauss-Markov-Theorems (Satz 3.3.2) zusätzlich*

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n) \quad , \tag{3.11}$$

so können mit der Bezeichnung $(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = [(g_{ij})]$ *für* $i, j = 1, \ldots, q$ *folgende Konfidenzintervalle für die Parameter* $\beta_j \quad (j = 1, \ldots, q)$ *und* σ^2 *ermittelt werden:*

(a) $\left[\widehat{\beta}_j - t_{n-q-1;1-\frac{\alpha}{2}} \sqrt{s^2 g_{jj}},\ \widehat{\beta}_j + t_{n-q-1;1-\frac{\alpha}{2}} \sqrt{s^2 g_{jj}}\ \right]$
ist ein Konfidenzintervall für $\beta_j \quad (j = 1, \ldots, q)$ *mit Überdeckungswahrscheinlichkeit* $1 - \alpha$*.*

(b) $\left[\frac{(n-q-1)s^2}{\chi^2_{n-q-1;1-\frac{\alpha}{2}}},\ \frac{(n-q-1)s^2}{\chi^2_{n-q-1;\frac{\alpha}{2}}}\right]$
ist ein Konfidenzintervall für σ^2 *mit Überdeckungswahrscheinlichkeit* $1 - \alpha$*.*

Satz 3.3.4 (Maximum-Likelihood Schätzer)
Sei $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, *wobei* $\boldsymbol{X}$ *vollen Rang* $q+1$ *habe und* $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. *Der Maximum-Likelihood (Plausible) Schätzer von* $\boldsymbol{\beta}$ *ist gleich dem Kleinsten-QS-Schätzer* $\widehat{\boldsymbol{\beta}}$. *Außerdem gilt:*

(a) $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \sim N_{q+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1})$,

(b) $\widehat{\boldsymbol{\beta}}$ *und die Residuen* $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ *sind voneinander unabhängig,*

(c) $n\widehat{\sigma}^2 = \widehat{\boldsymbol{\varepsilon}}^\top\widehat{\boldsymbol{\varepsilon}} \sim \sigma^2\chi^2_{n-q-1}$, *wobei* $\widehat{\sigma}^2$ *der Maximum-Likelihood Schätzer von* σ^2 *ist.*

Beweis: siehe z.B. Johnson & Wichern (1998), S. 390.

Beispiel 3.3.1 *Gegeben seien die Daten mit den Prüfungsergebnissen der 88 Studenten aus Tabelle 2.1. Wir wollen eine multiple Regressionsanalyse durchführen, bei der die abhängige Größe die erste Variable Mechanik (ME) ist und die restlichen Fächer die unabhängigen Größen darstellen. Als Modell wählen wir den linearen multiplen Regressionsansatz mit einem konstanten Term.*

Man erhält folgendes Resultat:

```
Residual Standard Error = 14.1372,  Multiple R-Square = 0.3764
N = 88,  F-statistic = 12.5253 on 4 and 83 df, p-value = 0

             coef std.err  t.stat p.value
Intercept -12.0171  7.7542 -1.5498  0.1250
       AG   0.4643  0.1461  3.1771  0.0021
       LA   0.5223  0.2421  2.1572  0.0339
       AN  -0.0022  0.1507 -0.0147  0.9883
       ES   0.0273  0.1217  0.2241  0.8233
```

Die letzten beiden Fächer (AN, ES) leisten keinen Beitrag zur Erklärung der Variable ME.

Das Regressionsmodell und dessen Voraussetzungen kann mit der Residuenanalyse *überprüft werden. In Abbildung 3.2 sind die geschätzen Werte* $\widehat{\boldsymbol{y}}$ *den beobachteten Werten* $\boldsymbol{y}$ *gegenübergestellt. Abbildung 3.3 zeigt* $\widehat{\boldsymbol{y}}$ *gegen die standardisierten Residuen* $\hat{\boldsymbol{\varepsilon}}^{norm}$ *aufgetragen, wobei*

$$\hat{\boldsymbol{\varepsilon}}^{norm} = \frac{\hat{\boldsymbol{\varepsilon}}}{s}$$

gilt. Zur Überprüfung auf Normalverteilung der Residuen wird in Abbildung 3.4 ein Q-Q-Plot der $\frac{i}{n+1}$*-Quantile* $u_{\frac{i}{n+1}}$ *der Standardnormalverteilung gegen die geordneten standardisierten Residuen* $\hat{\boldsymbol{\varepsilon}}^{norm}_{(1)} \leq \hat{\boldsymbol{\varepsilon}}^{norm}_{(2)} \leq \ldots \leq \hat{\boldsymbol{\varepsilon}}^{norm}_{(n)}$ *dargestellt.*

Abbildung 3.2: Residuenplot der geschätzten gegen die beobachteten y-Werte

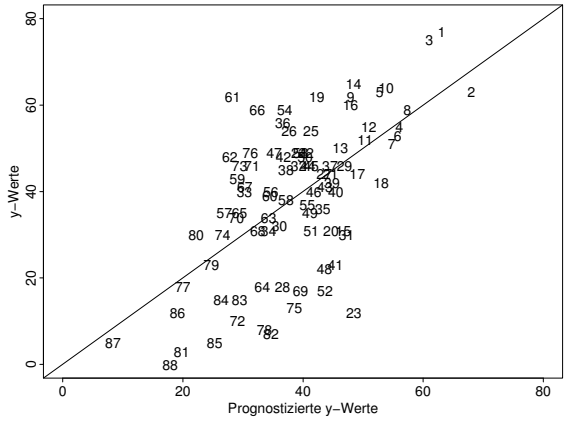
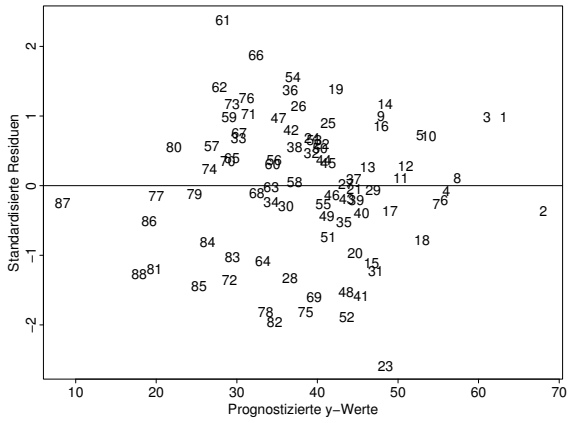


Abbildung 3.3: Residuenplot der geschätzten y-Werte gegen die standardisierten geschätzten Residuen

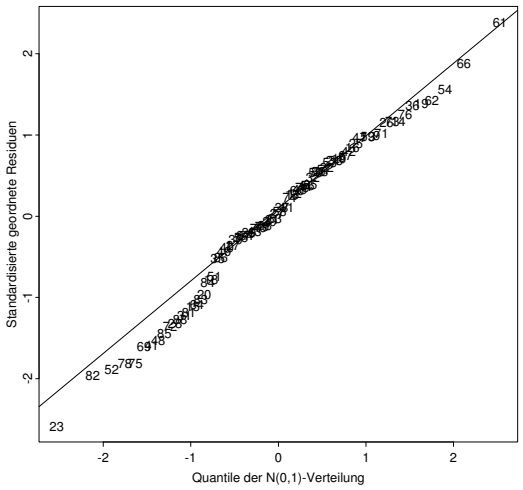


3.4 Multivarate lineare Regression

Haben wir bei der multiplen Regression den Fall *einer* abhängigen Größe betrachtet, so werden bei der multivariaten Regression m abhängige Größen $Y_1, \ldots, Y_m$ berücksichtigt. Somit kann für jede einzelne abhängige Größe ein Regressionsmodell für die Vorhersagevariablen $x_1, \ldots, x_q$ angesetzt werden:

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}x_1 + \ldots + \beta_{q1}x_q + \varepsilon_1 \\ &\vdots \quad\quad \vdots \\ Y_j &= \beta_{0j} + \beta_{1j}x_1 + \ldots + \beta_{qj}x_q + \varepsilon_j \\ &\vdots \quad\quad \vdots \\ Y_m &= \beta_{0m} + \beta_{1m}x_1 + \ldots + \beta_{qm}x_q + \varepsilon_m \end{aligned} \tag{3.12}$$

Abbildung 3.4: Q-Q-Plot der Quantile der Standardnormalverteilung gegen die Quantile der standardisierten geschätzten Residuen



Der Fehlerterm $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_m)^\top$ hat als Voraussetzung $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ und $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$, also die zu verschiedenen abhängigen Größen gehörigen Fehlerterme können miteinander korreliert sein.

Für eine Stichprobe von Umfang n kann nun für jede abhängige Größe $\boldsymbol{Y}_j = (Y_{1j}, \ldots, Y_{nj})^\top$ $(j = 1 \ldots, m)$ das Regressionsmodell analog zum vorigen Abschnitt angeschrieben werden:

$$\begin{aligned} Y_{1j} &= \beta_{0j} + \beta_{1j}x_{11} + \ldots + \beta_{qj}x_{1q} + \varepsilon_{1j} \\ Y_{2j} &= \beta_{0j} + \beta_{1j}x_{21} + \ldots + \beta_{qj}x_{2q} + \varepsilon_{2j} \\ &\vdots \quad \vdots \\ Y_{nj} &= \beta_{0j} + \beta_{1j}x_{n1} + \ldots + \beta_{qj}x_{nq} + \varepsilon_{nj} \end{aligned}$$

Diese Gleichungen können mit Matrixschreibweise auch zusammengefasst werden in

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.13}$$

mit den $(n \times m)$-Matrizen $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m)$ und $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_m)$. Aufgrund des konstanten Terms (Eins in der ersten Spalte) hat $\boldsymbol{X}$ Dimension $(n \times (q+1))$. $\boldsymbol{\beta}$ ist die $((q+1) \times m)$-Matrix der Regressionkoeffizienten mit den Spalten $\boldsymbol{\beta}_j$. Für den Fehlerterm gilt somit $E(\boldsymbol{\varepsilon}_j) = \mathbf{0}$ und $\mathrm{Cov}(\boldsymbol{\varepsilon}_j, \boldsymbol{\varepsilon}_k) = \sigma_{jk}\boldsymbol{I}$ $(j, k = 1, \ldots, m)$. D.h. die m Beobachtungen des i-ten Versuches $(i = 1, \ldots, n)$ haben Kovarianzmatrix $\boldsymbol{\Sigma} = [(\sigma_{jk})]$, aber Beobachtungen von verschiedenen Versuchen sind unkorreliert.

Der Kleinste-QS-Schätzer bzw. der Maximum-Likelihood Schätzer für $\boldsymbol{\beta}_j$ ist

$$\widehat{\boldsymbol{\beta}}_j = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y}_j \tag{3.14}$$

bzw. zusammengefasst

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \ . \tag{3.15}$$

Nachdem sich die Form des Regressionsschätzers zum vorigen Abschnitt nicht geändert hat, gelten auch alle Aussagen in analoger Weise, insbesondere auch der Satz von Gauss-Markov.

Literatur

N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley & Sons, New York, 1981.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, London, 4th edition, 1998.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Acad. Press, London, 1979.

G.A.F. Seber. *Multivariate observations*. John Wiley & Sons, New York, 1984.

Kapitel 4

Auszug aus Robuster Statistik

4.1 Einleitung

Statistische Entscheidungen basieren auf Beobachtungen oder werden aufgrund spezieller Voraussetzungen (Zufälligkeit, Unabhängigkeit, Normalverteilung, usw.) getroffen. Gerade diese Voraussetzungen sind es, unter denen einerseits die klassischen Methoden funktionieren und andererseits die mathematisch einfache Handhabung ermöglicht wird. Geringfügige Abweichungen führen schon meist zu Fehlschlüssen. In der Praxis sind aber diese idealen Voraussetzungen eher selten erfüllt. Dies kann z.B. durch das Auftreten von Fehlern (Tipp- oder Übertragungsfehler) oder anderen Einflüssen zustande kommen. Auch die zugrundeliegende Verteilung wird nicht exakt mit der angenommenen übereinstimmen, sondern viel mehr in einer Umgebung dieser liegen.

Ein Ziel der robusten Statistik ist es nun Prozeduren zu finden, die gegenüber solchen Abweichungen resistent sind. Es sollen also kleine Abweichungen vom Modell auch nur geringe Auswirkungen haben.

Wir wollen uns hier auf zwei wichtige Themen konzentrieren: robuste lineare Regression und robuste Schätzung von Lokation und Kovarianz.

4.2 Robuste lineare Regression – Teil 1: Methoden

4.2.1 LS-Regression

In der Regressionsanalyse wird versucht, aufgrund gegebener Werte der Vorhersage-Variablen den genauen funktionellen Zusammenhang zur abhängigen Variable festzustellen. Man versucht also, aufgrund von bekannten Werten $1, x_{i1}, \ldots, x_{iq}$ das Regressionsmodell (3.2) den beobachteten Werten y_i anzupassen. Genauer gesagt, es sollen die Regressionskoeffizienten $\boldsymbol{\beta}$ und die Fehlervarianz σ^2 aufgrund der vorliegenden Daten ermittelt werden.

Die Differenz

$$r_i(\boldsymbol{\beta}) = y_i - (\beta_0 + \beta_1 x_{i1} + \ldots + \beta_q x_{iq}) \ ,$$

die die Abweichung der Anpassung vom beobachteten Wert angibt, wird auch als *Residuum* bezeichnet. Bei der *Methode der Kleinsten Quadrate* (*Least Squares*, oder kurz LS) soll die Summe der quadrierten Residuen minimiert werden:

$$\begin{aligned}\sum_{i=1}^{n} r_i^2(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_q x_{iq})^2 \\ &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned} \tag{4.1}$$

Jener Koeffizient, der dieses Kriterium minimiert, wird mit $\widehat{\boldsymbol{\beta}}_{LS}$ bezeichnet. Er ist der Kleinste-Quadratsummen-Schätzer des Regressionsparameters $\boldsymbol{\beta}$. Es ergibt sich

$$\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \quad . \tag{4.2}$$

Mit Hilfe von $\widehat{\boldsymbol{\beta}}_{LS}$ können die *geschätzten Werte* $\widehat{\boldsymbol{y}}_{LS}$ von $\boldsymbol{y}$ berechnet werden durch

$$\widehat{\boldsymbol{y}}_{LS} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{LS} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y} \ , \tag{4.3}$$

wobei $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ als "Hat"-Matrix bezeichnet wird. Die *geschätzten Residuen* sind

$$\widehat{\boldsymbol{\varepsilon}}_{LS} = \boldsymbol{y} - \widehat{\boldsymbol{y}}_{LS} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \ . \tag{4.4}$$

Beispiel 1: Bei der LS- oder Kleinsten-Quadrate Regression wird zur Schätzung der Koeffizienten *jeder* Wert der Residuen quadriert berücksichtigt. Das kann allerdings dazu führen, dass einzelne Werte, die nicht dem linearen Trend folgen (z.B. Ausreißer), starken Einfluss auf diese Schätzung haben. In Abbildung 4.1 wird dies illustriert. In der linken Grafik sieht man 5 Punkte, die etwa einem linearen Trend folgen. Die LS-Regressionsgerade ist eingezeichnet, die offensichtlich diese lineare Beziehung gut modelliert. Nun wird in der rechten Grafik der 4. Punkt in y-Richtung verschoben. Man erkennt, dass die LS-Regressionsparameter davon stark beeinflusst werden, weil die Gerade "nach oben gezogen" wird.

In Abbildung 4.2 wird bei einer ähnlichen Konfiguration ein Wert (Nr. 1) in x-Richtung verschoben. Die Auswirkung auf LS-Regression ist noch gravierender als vorher, weil die Regressionsgerade nicht nur abgelenkt sondern sogar "gekippt" wird. Man nennt solche x-Ausreißer *Hebelpunkte*, weil sie eine "Hebelwirkung" auf die LS-Regressionsgerade haben.

Bei der (multiplen) Regression sind *Diagnostik-Plots* üblich, um Modellfehler bzw. Ausreißer sichtbar zu machen. Der Plot der *standardisierten Residuen* gegen den Index i der Beobachtung verfolgt diesen Zweck. Die standardisierten Residuen sind definiert als

$$\frac{\hat{\varepsilon}_i(\widehat{\boldsymbol{\beta}}_{LS})}{s_{LS}} \ ,$$

also geschätzte LS-Residuen dividiert durch die geschätzte Standardabweichung der Residuen. Wenn somit obige Modellvoraussetzung $\varepsilon_i \sim N(0, \sigma^2)$ zutrifft, sollten die standardisierten Residuen etwa $N(0,1)$-verteilt sein. Werte, die außerhalb der in der rechten Grafik von Abbildung 4.3 eingezeichneten Grenzen ± 2.5 liegen sind somit

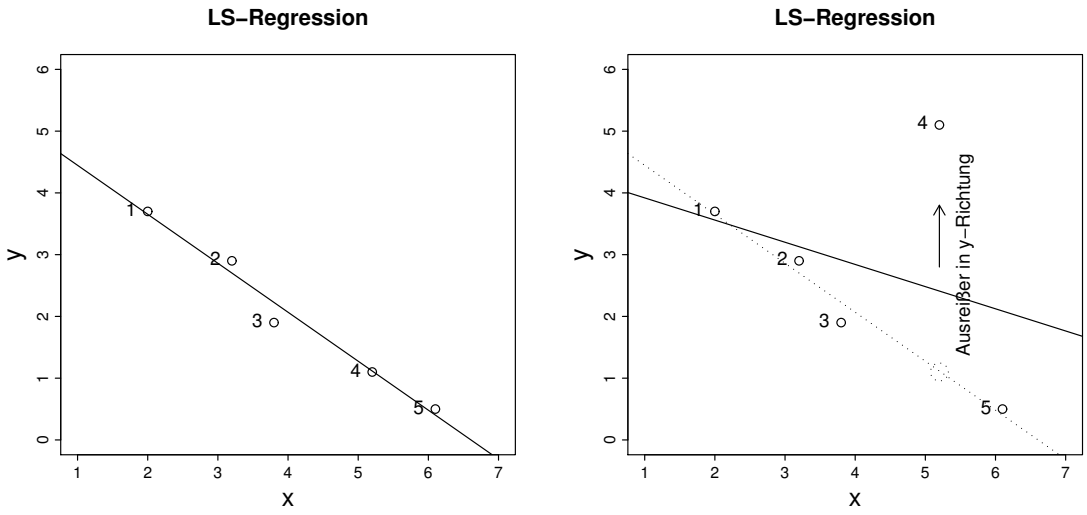


Abbildung 4.1: Auswirkung von y-Ausreißern bei LS-Regression.

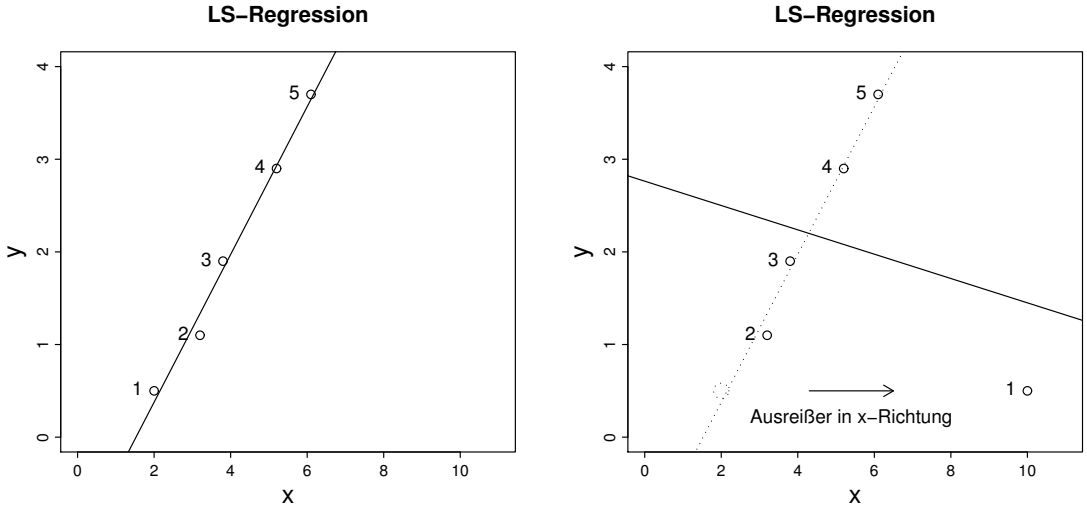


Abbildung 4.2: Auswirkung von x-Ausreißern bei LS-Regression.

potentielle Ausreißer. Abbildung 4.3 zeigt diese Plots für die LS-Regression bei den Daten mit den y-Ausreißern (links) und den x-Ausreißern (rechts). Man beachte, dass in beiden Grafiken kein Wert als Ausreißer identifiziert wird. Der Grund dafür ist die Verzerrung der Schätzung der LS-Regressionsparameter.

4.2.2 L_1-Regression

Die Empfindlichkeit bei LS-Regression gegenüber Ausreißern ist bedingt durch das quadratische Eingehen von großen Residuen bei der zu minimierenden Funktion (4.1). Dies kann leicht vermieden werden, indem anstelle des Quadrats der Absolut-

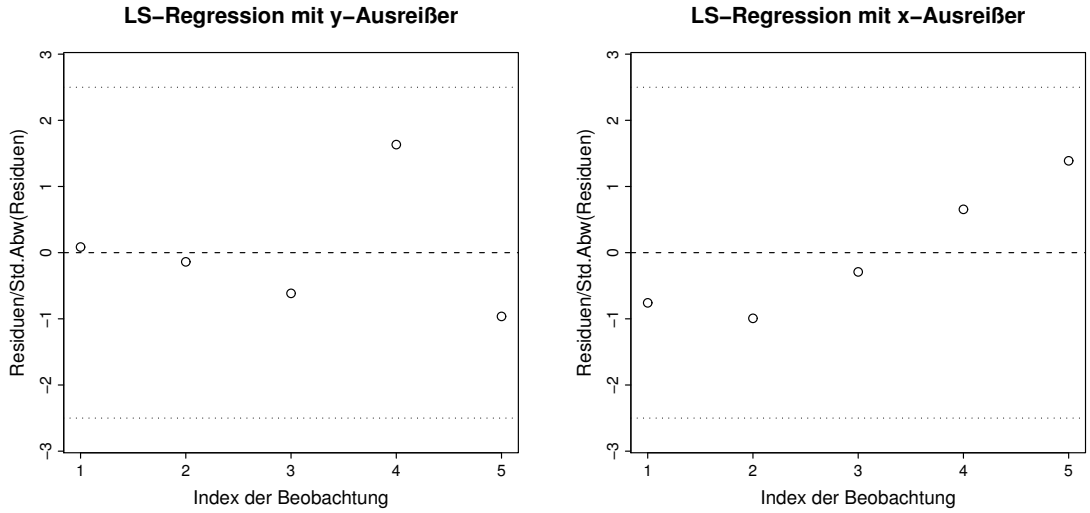


Abbildung 4.3: Plot der standardisierten Residuen auf die Daten mit y- bzw. x-Ausreißer.

betrag genommen wird. Bei L_1-Regression wird somit

$$\sum_{i=1}^{n} |r_i(\boldsymbol{\beta})| \tag{4.5}$$

(L_1-Norm) minimiert. Man erhält die geschätzten Regressionskoeffizienten $\widehat{\boldsymbol{\beta}}_{L1}$ nicht mehr mit einer expliziten Formel, sondern muss einen Algorithmus anwenden, der die Lösung approximiert.

Beispiel 2: Wir betrachten dieselben Daten wie in Beispiel 1 für y- bzw. x-Ausreißer und schätzen die Regressionsparameter mit L_1-Regression. Die Ergebnisse sind in Abbildung 4.4 dargestellt. Die punktierten Linien sind die jeweiligen Resultate ohne Verschiebung der Punkte, die durchgezogenen Linien beziehen sich auf die L_1-Regressionskoeffizienten der Daten mit Ausreißern. Man erkennt, dass y-Ausreißer keinen Einfluss auf L_1-Regression haben, während x-Ausreißer die Regressionsgerade zum Kippen bringen. Analog wie bei LS-Regression genügt also auch hier ein einziger Datenpunkt, um die Modellanpassung beliebig unsinnig zu machen. Man spricht dann auch von *Bruchpunkt* 0%, weil für ein sinnvolles Ergebnis 0% der Daten (beliebig "böswillige") Ausreißer sein dürfen.

4.2.3 LMS-Regression und LTS-Regression

Offensichtlich hat das L_1-Kriterium nicht zum gewünschten Erfolg geführt. Ein Grundprinzip der robusten Statistik ist es, ein Modell nicht *allen* Datenpunkten, sondern nur der Mehrheit der Daten anzupassen. Sowohl bei LS- als auch bei L_1-Regression wurden alle Daten $i = 1, \ldots, n$ berücksichtigt. Bei *Least Median of Squares* (LMS-) Regression wird die Funktion

$$\underset{i}{\text{med}}\ r_i^2(\boldsymbol{\beta}) \tag{4.6}$$

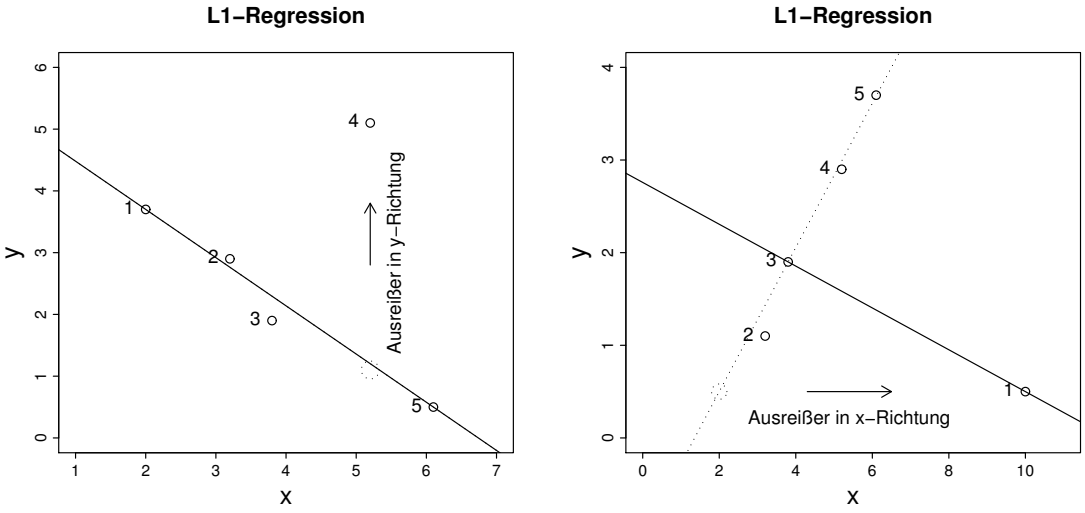


Abbildung 4.4: Auswirkung von y- bzw. x-Ausreißern bei L_1-Regression.

minimiert. Es wird also einfach die Summe in (4.1) durch den Median ersetzt. Auch hier gibt es keine explizite Lösung für die Regressionskoeffizienten $\widehat{\boldsymbol{\beta}}_{LMS}$, sondern nur approximative Algorithmen.

Mit Hilfe von theoretischen Überlegungen erkennt man, dass der Bruchpunkt von LMS-Regression 50% ist, d.h. bis zu 50% der Daten könnten beliebig verschoben werden ohne die Regressionskoeffizienten wesentlich zu verändern.

Eine andere sehr robuste Regressionsmethode ist *Least Trimmed Sum of Squares* (LTS-) Regression, bei der

$$\sum_{i=1}^{h} (r^2(\boldsymbol{\beta}))_{(i)} \tag{4.7}$$

minimiert wird. Dabei sind $(r^2(\boldsymbol{\beta}))_{(1)} \leq (r^2(\boldsymbol{\beta}))_{(2)} \leq \ldots \leq (r^2(\boldsymbol{\beta}))_{(h)} \leq \ldots \leq (r^2(\boldsymbol{\beta}))_{(n)}$ die der Größe nach geordneten *quadrierten* Residuen. Bei einer Wahl von $h \approx n/2$ hat die Methode einen Bruchpunkt von etwa 50%, und bei größerem h sinkt der Bruchpunkt auf $(n-h)/n$.

Beispiel 3: Sowohl LMS-Regression als auch LTS-Regression (mit $3 \leq h \leq 4$) liefern stabile Resultate auf die Daten von Beispiel 2 (siehe Abbildung 4.5).

Ähnlich wie in Abbildung 4.3 bei LS-Regression kann auch hier der Plot der standardisierten Residuen angefertigt werden. Es müssen also die LMS- oder LTS-Residuen durch deren Standardabweichung dividiert werden. Die Standardabweichung σ der Residuen wird bei LMS-Regression geschätzt durch

$$\hat{\sigma}_{LMS} = c_1 \cdot \sqrt{\underset{i}{\text{med}}\ \hat{\varepsilon}_i^2(\widehat{\boldsymbol{\beta}}_{LMS})}$$

und bei LTS-Regression mittels

$$\hat{\sigma}_{LTS} = c_2 \cdot \sqrt{\frac{1}{h} \sum_{i=1}^{h} (\hat{\varepsilon}^2(\widehat{\boldsymbol{\beta}}_{LTS}))_{(i)}}\ .$$

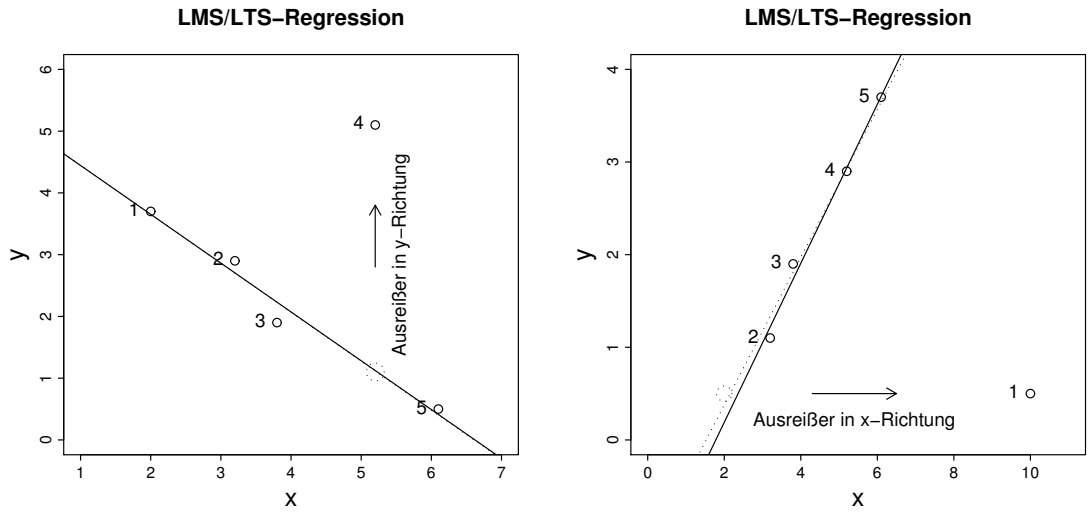


Abbildung 4.5: Auswirkung von y- bzw. x-Ausreißern bei LMS- bzw. LTS-Regression.

c_1 und c_2 sind Konstante für Konsistenz bei Normalverteilung. Man beachte, dass bei LMS- bzw. LTS-Regression nicht nur die Residuen, sondern auch die Standardabweichung der Residuen robust geschätzt werden. Somit kann der Plot der standardisierten Residuen zur Ausreißererkennung herangezogen werden. In Abbildung 4.6 sind diese Plots für die Regressionen in Abbildung 4.5 dargestellt. Man erkennt sehr deutlich, dass die Ausreißer weit außerhalb des Intervalls ± 2.5 liegen.

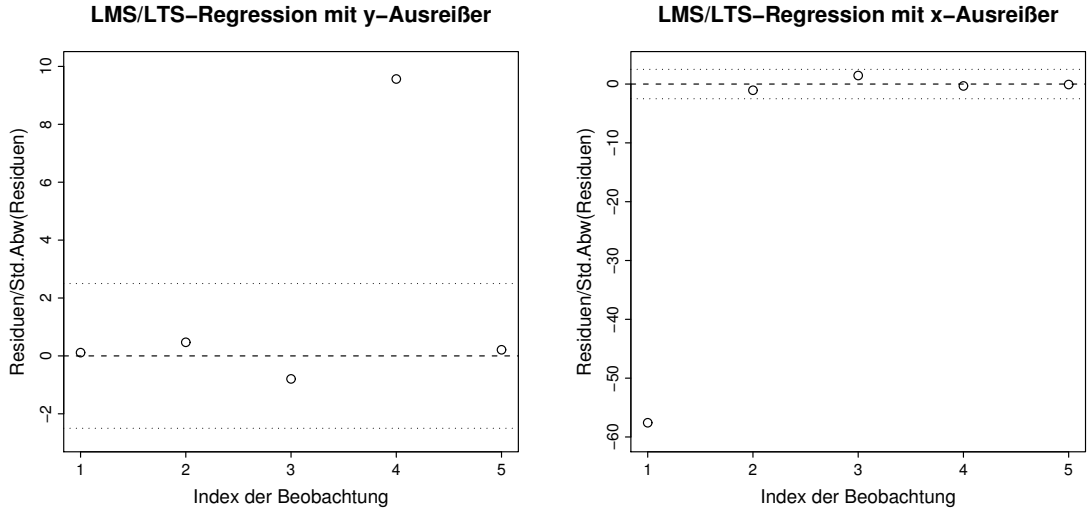


Abbildung 4.6: Plot der standardisierten Residuen auf die Daten mit y- bzw. x-Ausreißern bei LMS- bzw. LTS-Regression.

Beispiel 4: Wir betrachten die sogenannten "Stars"-Daten, die am sternförmigen Cluster CYG-OB1 von 47 Himmelskörpern in Richtung von Cygnus erhoben wurden.

Die x-Variable beschreibt das Spektrum der Himmelskörper und die y-Variable die logarithmierte Lichtintensität. Abbildung 4.7 zeigt in der linken Grafik die Daten, es wurde die LS-Regressionsgerade eingezeichnet. Man bemerkt, dass 4 Ausreißer die Schätzung so stark beeinflussen, dass das Regressionsmodell völlig unsinnig wird. In der rechten Grafik ist der Plot der standardisierten Residuen dargestellt, der keine Ausreißer erkennen lässt, weil alle Werte im Toleranzband ± 2.5 liegen.

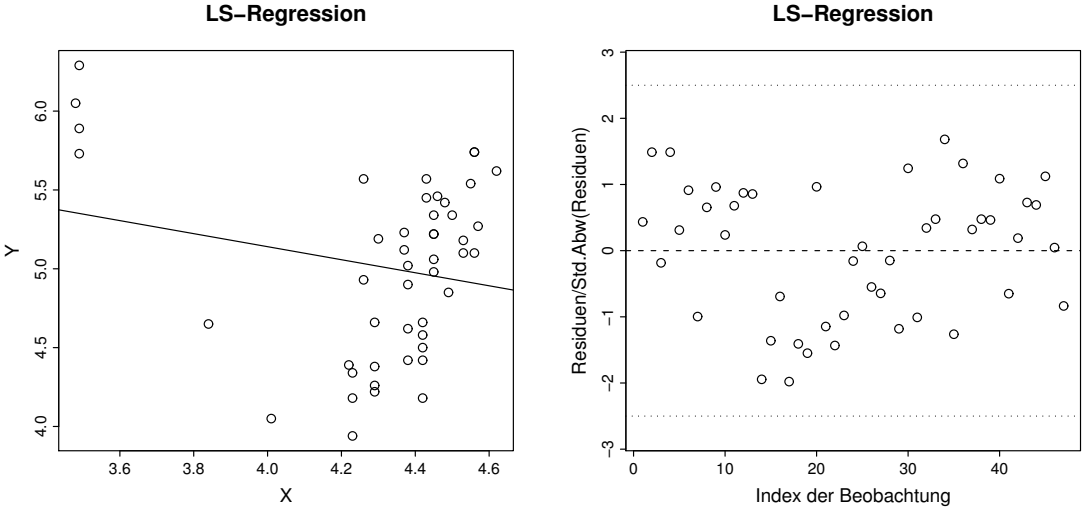


Abbildung 4.7: Stars-Daten: LS-Regression (links) und Diagnostik-Plot (rechts), der keine Ausreißer erkennen lässt.

Anders sieht dies bei LMS- oder LTS-Regression aus (Abbildung 4.8). Die Regressionsgerade folgt dem Haupttrend der Daten, aber auch der Diagnostik-Plot zeigt sehr schön die 4 extremen Ausreißer. Bei näherer Untersuchung hat sich herausgestellt, dass diese 4 Ausreißer keine Sterne, sondern Galaxien sind.

Beispiel 5: Die Daten von Hawkins, Bradu und Kass (1984) sind künstliche Daten, die zum Testen von robuster Regression konstruiert wurden. Man findet die Daten z.B. in `R` im Paket `rrcov` unter `hbk`. Die Daten bestehen aus einer abhängigen Größe y und drei unabhängigen Variablen x_1, x_2 und x_3. Bei diesem multiplen Regressionsproblem können also die Daten selbst nicht mehr wie in den vorigen Beispielen geplottet werden. Von den 75 Beobachtungen sind die ersten 14 Werte Hebelpunkte (also Ausreißer im Raum der x-Variablen), wobei aber die Werte 11-14 entlang der Regressionshyperebene liegen und somit "gute" Hebelpunkte darstellen, die die Präzision der Schätzung sogar verbessern. In Abbildung 4.9 sind die Diagnostik-Plots bezüglich LS-Regression (links) und LMS- bzw. LTS-Regression (rechts) dargestellt. Während mit LS-Regression ein Datenwert als Ausreißer erkannt wird (der aber in Richtung der Regressionshyperebene liegt), werden bei der robusten Regression genau die ersten 10 Werte als Ausreißer erkannt.

Bemerkung: Im Anschluss an LMS- oder LTS-Regression könnte zur Erhöhung der Effizienz der Schätzer eine gewichtete LS-Regression durchgeführt werden. Man

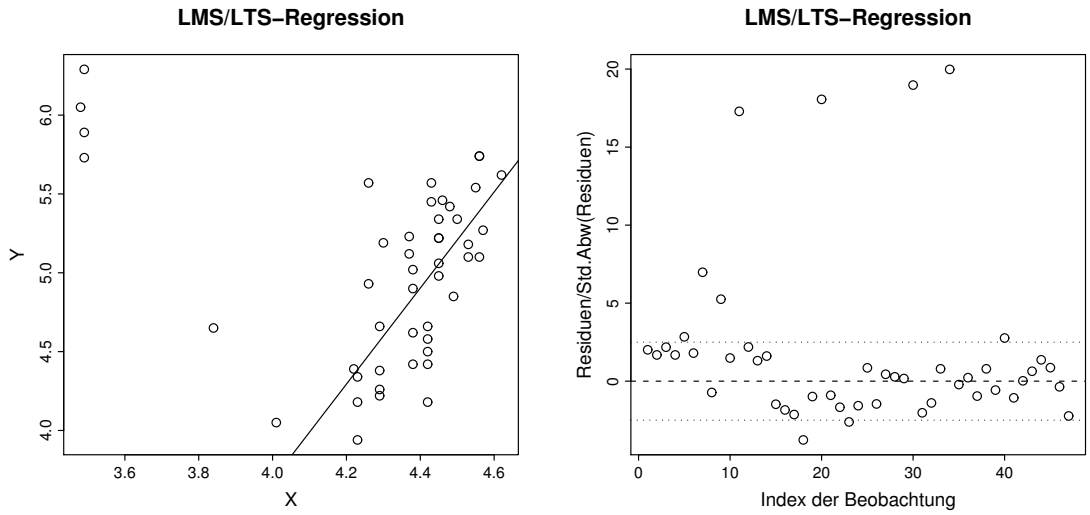


Abbildung 4.8: Stars-Daten: LTS-Regression (links) und Diagnostik-Plot (rechts), der die Ausreißer klar erkennen lässt.

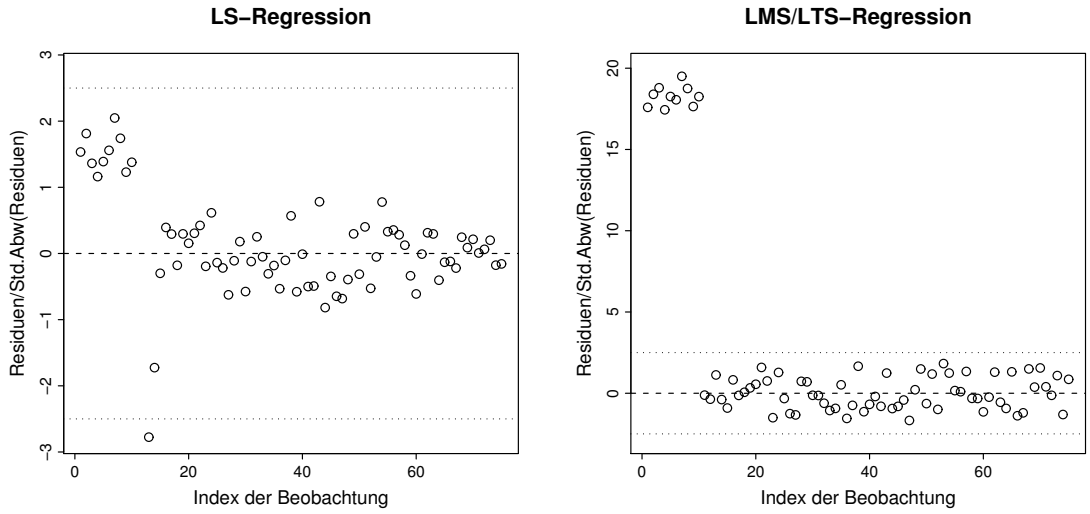


Abbildung 4.9: Daten von Hawkins, Bradu und Kass: LTS-Regression (links) und Diagnostik-Plot (rechts), der die Ausreißer klar erkennen lässt.

kann dazu jede Beobachtung mit Gewicht 0 oder 1 gewichten, jenachdem, ob diese Beobachtung im Diagnostik-Plot der robusten Regression außerhalb oder innerhalb des Toleranzbandes ± 2.5 liegt. Die gewichtete LS-Regression ist damit robust, und es können die bei LS-Regression üblichen Konfidenzintervalle bzw. Tests für die Regressionskoeffizienten verwendet werden.

4.3 Robuste Schätzung von multivariater Lokation und Kovarianz

4.3.1 Allgemeines

Die multivariate Lokation und insbesondere die Kovarianz spielen eine zentrale Rolle in der multivariaten Statistik, weil viele Methoden direkt darauf aufbauen. Bei Hauptkomponentenanalyse wird direkt die Kovarianzmatrix zur Berechnung der Hauptkomponenten herangezogen, ähnlich ist es bei der Faktorenanalyse. Bei Diskiminanzanalyse werden die Kovarianzmatrizen der einzelnen Gruppen verwendet, bei kanonischer Korrelationsanalyse gehen zusätzlich die gemischten Kovarianzen ein, usw.

Im "klassischen" Fall wird der Erwartungswert $\boldsymbol{\mu}$ mit dem arithmetischen Mittel $\bar{\boldsymbol{x}}$, und die Kovarianzmatrix $\boldsymbol{\Sigma}$ mit der Stichprobenkovarianzmatrix $\boldsymbol{S}$ geschätzt. Natürlich sind diese Schätzer empfindlich gegenüber Ausreißern: Das arithmetische Mittel kann bereits durch einen einzigen Ausreißer verzerrt werden, weil bei der Berechnung jeder Datenpunkt mit gleichen Gewicht eingeht (Bruchpunkt 0%). Das Gleiche gilt bei der Berechnung von $\boldsymbol{S}$; hier geht außerdem auch $\bar{\boldsymbol{x}}$ ein (Bruchpunkt 0%).

Es gibt viele Ansätze zur Robustifizierung der Schätzer. Wir werden uns hier auf zwei Methoden beschränken, deren Konzept sehr ähnlich ist. Wichtig ist dabei die Forderung, dass die Schätzer *affin äquivariant* sein sollen. D.h., für einen Lokationsschätzer $\boldsymbol{T}$ soll gelten

$$\boldsymbol{T}(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}) = \boldsymbol{A} \cdot \boldsymbol{T}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) + \boldsymbol{b}$$

und für einen Kovarianz-Schätzer $\boldsymbol{C}$

$$\boldsymbol{C}(\boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{b}, \ldots, \boldsymbol{A}\boldsymbol{x}_n + \boldsymbol{b}) = \boldsymbol{A} \cdot \boldsymbol{C}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \cdot \boldsymbol{A}^\top$$

für eine nicht-singuläre $(p \times p)$-Matrix $\boldsymbol{A}$ und ein $\boldsymbol{b} \in \mathbb{R}$. Das bedeutet also, dass die Schätzer auch bei Transformationen (wie z.B. Standardisierung) richtig mittransformieren.

4.3.2 MVE und MCD Schätzer

Sowohl der *Minimum Volume Ellisoid* (MVE) Schätzer als auch der *Minimum Covariance Determinant* (MCD) Schätzer liefern robuste Schätzungen von Mittel und Kovarianz mit maximalem Bruchpunkt von 50%. Die Definition der Lokationsschätzung $\boldsymbol{T}$ und der Kovarianzschätzung $\boldsymbol{C}$ für eine Datenmatrix $\boldsymbol{X}$ ist folgendermaßen:

- MVE-Schätzer:

  $\boldsymbol{T}(\boldsymbol{X})$: Zentrum jenes Ellipsoids mit minimalem Volumen, das *mindestens* die Hälfte der Datenpunkte von $\boldsymbol{X}$ enthält.

  $\boldsymbol{C}(\boldsymbol{X})$: Gegeben durch die Form dieses Ellipsoids, multipliziert mit einem Faktor für Konsistenz bei Normalverteilung.

Formal werden die Schätzer definiert als Minimierung der Determinante von $\boldsymbol{C}$ unter der Bedingung

$$\#\{i : (\boldsymbol{x}_i - \boldsymbol{T})^\top \boldsymbol{C}^{-1}(\boldsymbol{x}_i - \boldsymbol{T}) \leq c^2\} \geq \left\lfloor \frac{n+p+1}{2} \right\rfloor ,$$

wobei # die Kardinalität ist, und die Konstante c als $\chi^2_{p,0.5}$ gewählt wird. Die approximative Lösung wird über Ellipsoide gemacht, die durch die Kovarianzmatrix von $p+1$ Beobachtungen bestimmt werden.

- MCD-Schätzer:

  Es wird nach jenen h Punkten gesucht, deren empirische Kovarianzmatrix kleinstmögliche Determinante hat.

  $\boldsymbol{T}(\boldsymbol{X})$: arithmetisches Mittel der gesuchten h Punkte.

  $\boldsymbol{C}(\boldsymbol{X})$: Empirische Kovarianzmatrix der gesuchten h Punkte, multipliziert mit einem Faktor für Konsistenz bei Normalverteilung.

Wenn beim MCD-Schätzer der Parameter $h \approx n/2$ gewählt wird, erhält man den maximalen Bruchpunkt. Für größeres h sinkt der Bruchpunkt auf $(n-h)/n$.
Man kann zeigen, dass die h Punkte, die zur MCD Lösung führen, innerhalb eines Ellipsoids liegen. Dies ist aber nicht notwendigerweise das Ellipsoid mit kleinstem Volumen, womit die MCD von der MVE Lösung i.A. verschieden ist.

Beispiel 6: Wir betrachten 2-dimensionale Daten mit 18 Messungen von anorganischem und organischem Phosphor im Boden. Abbildung 4.10 zeigt die Daten, gemeinsam mit 97.5% Toleranzellipsen. Eine $(1-\alpha)$% Toleranzellipse enthält im Falle von normalverteilten Daten $(1-\alpha)$% der Datenpunkte. In der linken Grafik wurde die Toleranzellipse mit der Lokation und Kovarianz des MCD-Schätzers ermittelt (MVE würde ein sehr ähnliches Resultat liefern), in der rechten Grafik wurden die klassischen Schätzungen $\bar{\boldsymbol{x}}$ und $\boldsymbol{S}$ verwendet. Datenpunkte, die außerhalb der Toleranzellipse liegen, sind potentielle Ausreißer. Man erkennt, dass nur mit der robusten Methode die offensichtlichen Ausreißer identifiziert werden.

Beispiel 6 hat gezeigt, dass mit Hilfe von robuster Lokations- und Kovarianzschätzung auch Ausreißer identifiziert werden können. Allerdings bleibt diese Methode mit der Toleranzellipse auf $p = 2$ Dimensionen beschränkt, in höherer Dimension ist das nicht mehr sinnvoll durchführbar.

Stattdessen kann aber ein äquivalenter Ansatz gewählt werden, nämlich die Berechnung der **Mahalanobis-Distanz**, deren Quadrat definiert ist als

$$\mathrm{MD}_i^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^\top \boldsymbol{S}^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \ . \tag{4.8}$$

$\bar{\boldsymbol{x}}$ ist das arithmetische Mittel und $\boldsymbol{S}$ die empirische Kovarianzmatrix. Die Mahalanobis-Distanz wird für jeden Datenpunkt $\boldsymbol{x}_i \in \mathbb{R}^p$ berechnet, sie gibt an, wie weit ein Wert vom Zentrum der Punktwolke der Daten entfernt ist, relativ zur Größe der Punktwolke. Im Falle von p-variat normalverteilten Daten ist die quadrierte Mahalanobis-Distanz approximativ verteilt nach χ^2_p. Nachdem große Werte der Mahalanobis-Distanz auf Ausreißer schließen lassen, kann ein Quantil, z.B. $\chi^2_{p;0.975}$ verwendet

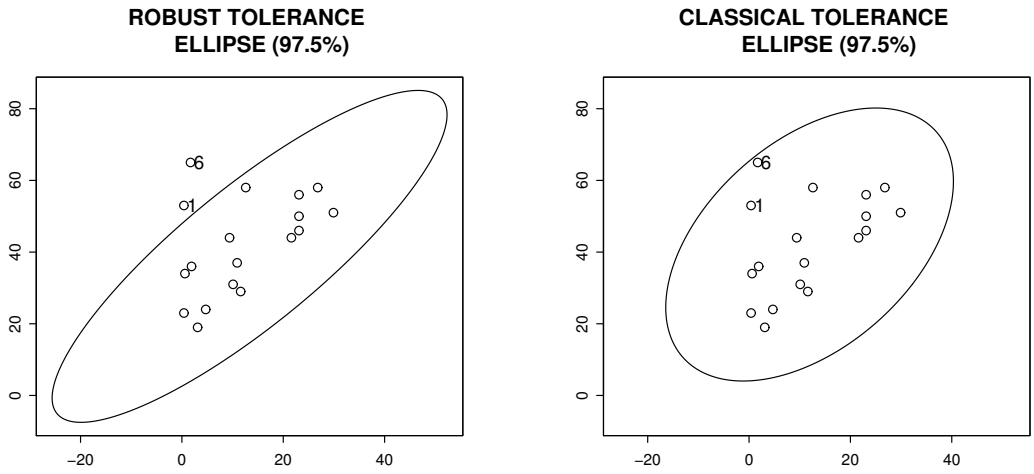


Abbildung 4.10: Phosphor-Daten: MCD-Schätzer (links) und klassische Schätzung (rechts).

werden, um Datenpunkte als Ausreißer zu deklarieren. Es kann gezeigt werden, dass diese Methode tatsächlich äquivalent ist zur Ausreißerbestimmung mittels Toleranzellipse. Der Vorteil ist aber, dass die Mahalanobis-Distanz ein 1-dimensionales Maß ist, und daher für Daten beliebiger Dimension p grafisch veranschaulicht werden kann.

Natürlich werden wir nicht die Mahalanobis-Distanz laut Definition (4.8) verwenden, weil dieses Maß, das auf klassischen Schätzern basiert, nicht robust ist. Stattdessen ersetzen wir die Schätzung von Mittel und Kovarianz durch robuste Gegenstücke, wie sie z.B. der MVE- oder MCD-Schätzer liefert. Diese robustifizierte Mahalanobis-Distanz nennen wir dann **Robuste Distanz** (RD).

Beispiel 7: Wir betrachten nochmals die Phosphor-Daten von Beispiel 6 und berechnen sowohl die Robuste Distanz als auch die Mahalanobis-Distanz (siehe Abbildung 4.11). Wie bei Beispiel 6 werden mit der Robusten Distanz die Beobachtungen 1 und 6 als Ausreißer identifiziert, da sie über dem Wert $\sqrt{\chi^2_{2;0.975}} = 2.72$ liegen. Die Mahalanobis-Distanz wird von den Ausreißern verzerrt und daher werden die Ausreißer nicht als solche angezeigt. Man spricht in diesem Fall auch von *maskierten* Ausreißern bzw. vom *Maskierungseffekt*. Dieser Effekt tritt speziell dann auf, wenn nicht einzelne sondern eine ganze Gruppe von Ausreißern auftritt.

Beide Grafiken können zu einer einzigen zusammengefasst werden, da der Index der Beobachtung als eigene Achse oft nicht so relevant ist. Der resultierende **Distance-Distance Plot** ist in Abbildung 4.12 gezeigt. Mit dem horizontal und vertikal eingezeichneten Wert $\sqrt{\chi^2_{2;0.975}} = 2.72$ werden 4 Quadranten definiert: Im Quadranten links unten befinden sich die regulären Beobachtungen, die keine Ausreißer sind. Der Quadrant rechts unten sollte leer bleiben. Rechts oben befinden sich die Ausreißer, die sowohl mit der klassischen als auch mit der robusten Methode als solche erkannt werden. Im Quadranten links oben sind schließlich die maskierten

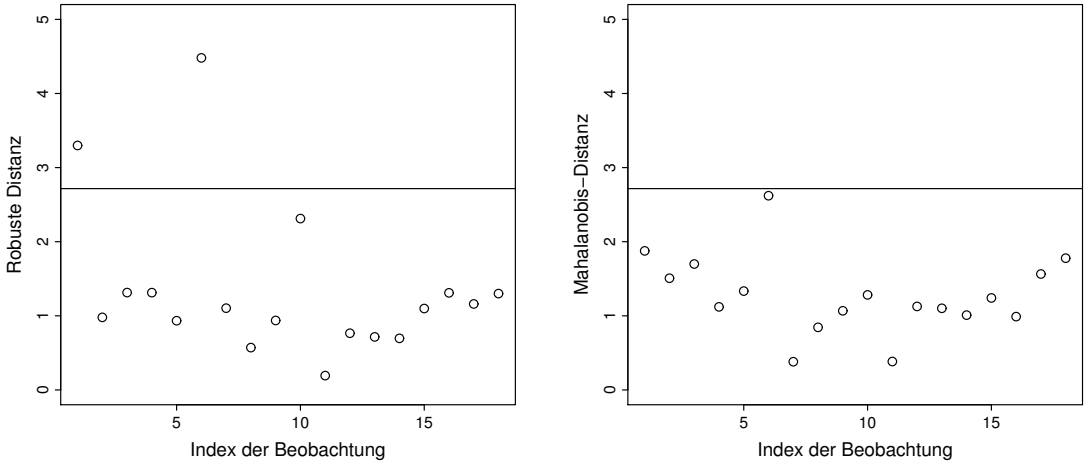


Abbildung 4.11: Phosphor-Daten: Robuste Distanzen (links) und Mahalanobis-Distanzen (rechts).

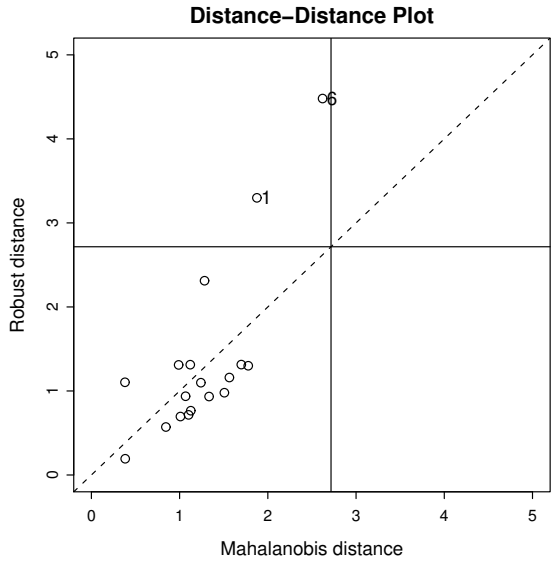


Abbildung 4.12: Phosphor-Daten: Robuste Distanzen versus Mahalanobis-Distanzen.

Ausreißer, die nur mit der Robusten Distanz erkannt werden. Wenn alle Datenpunkte auf der strichlierten Linie liegen würden, kämen die Daten genau aus einer multivariaten Normalverteilung.

Beispiel 8: Wir verwenden die sogenannten *Stackloss* Daten (in `R` im Package *rrcov*), die in einem kontrollierten Prozess die Pflanzenoxidation beschreiben. Es wurden an 21 aufeinanderfolgenden Tagen die Flussgeschwindigkeit der Kühlluft, die Kühlwassertemperatur und die Säurekonzentration gemessen. Der Distance-Distance Plot ist in Abbildung 4.13 links dargestellt. Mit der klassischen Mahalanobis-

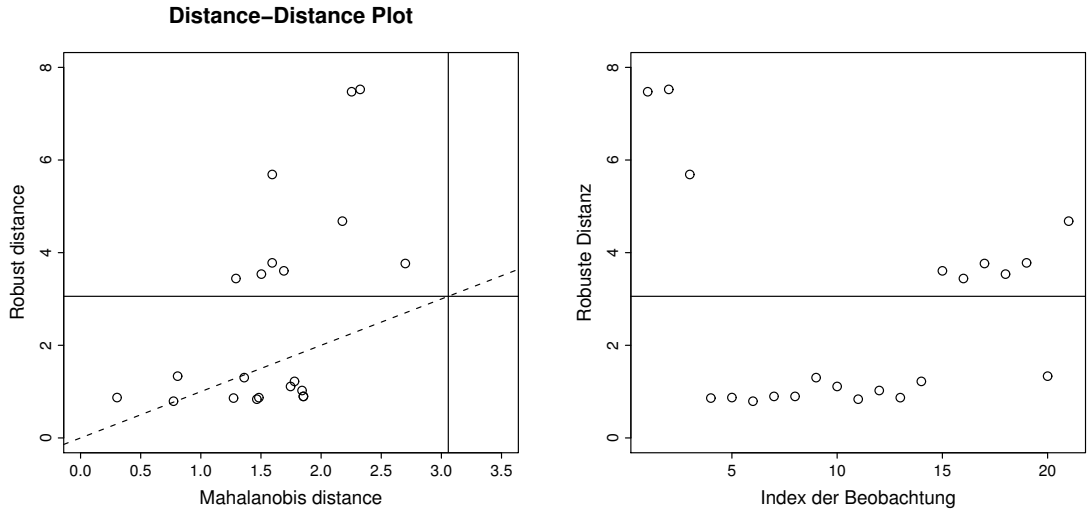


Abbildung 4.13: Stackloss Daten: Robuste Distanzen versus Mahalanobis-Distanzen.

Distanz wird kein einziger Wert als Ausreißer deklariert. Nur die Robuste Distanz zeigt eine ganze Gruppe von maskierten Ausreißern. Zur besseren Übersicht ist in der rechten Grafik die Robuste Distanz gegen den Index der Beobachtung aufgetragen. Die Ausreißer sind an den ersten 3 Tagen und in den letzten Tagen des Experiments zustande gekommen. Offenbar sind die Vorgänge hier sehr verschieden vom stabilen Zustand in der Mitte dieser Periode gewesen.

4.4 Robuste lineare Regression – Teil 2: Diagnostik

Wie bei anderen statistischen Modellanpassungen ist man auch bei der Regression daran interessiert, das Modell und die Voraussetzungen für die Anpassung zu überprüfen. Bei der Regression macht man daher Plots der Residuen, um die Voraussetzung der Normalverteilung und der Gleichheit der Residuenvarianz zu überprüfen. Von besonderem Interesse ist außerdem, ob die Daten dem Modell folgen, oder ob es etwa einzelne Beobachtungen gibt, die davon abweichen. Solche "Ausreißer", die die klassische LS-Regression stark beeinflussen können, sind interessant, weil die Struktur dieser Daten völlig anders ist – aus welchen Gründen auch immer.

4.4.1 Hat-Matrix zur Identifikation von Hebelpunkten

Hebelpunkte (*Leverage Points*) sind Datenpunkte, die eine "Hebelwirkung" auf nicht robuste Regressionsmethoden ausüben können, also die Schätzung der Regressionsparameter stark beeinflussen können. Ein beliebtes Werkzeug für das Auffinden von Hebelpunkten sind die Diagonalelemente der Hat-Matrix aus Abbildung 4.14, definiert als $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Es gilt, dass $\widehat{\boldsymbol{y}}_{LS} = \boldsymbol{H}\boldsymbol{y}$ ist, und insbesondere

ist

$$\widehat{y}_{i_{LS}} = (h_{i1} \cdots \underline{h_{ii}} \cdots h_{in}) \begin{pmatrix} y_1 \\ \vdots \\ \underline{y_i} \\ \vdots \\ y_n \end{pmatrix} .$$

h_{ij} ist somit interpretierbar als der Effekt der j-ten Beobachtung y_j auf $\widehat{y}_{i_{LS}}$. Speziell gibt h_{ii} den Einfluss der i-ten Beobachtung auf die eigene Vorhersage $\widehat{y}_{i_{LS}}$ an.

Eigenschaften der (Diagonalelemente der) Hat-Matrix:

- $\boldsymbol{H}^\top = \boldsymbol{H}$, $\boldsymbol{H} = \boldsymbol{H}\boldsymbol{H}$, $\text{trace}(\boldsymbol{H}) = \text{trace}[\boldsymbol{X}^\top \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1}] = \text{trace}(\boldsymbol{I}_{q+1}) = q+1$
  Daraus folgt, dass das Mittel von h_{ii} gleich $\frac{q+1}{n}$ ist.
- $h_{ii} = (\boldsymbol{H}\boldsymbol{H})_{ii} = \sum_{j=1}^n h_{ij}h_{ji} = \sum_{j=1}^n h_{ij}^2 \quad \Longleftrightarrow \quad h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$
  Daraus folgt, dass $0 \leq h_{ii} \leq 1$ ist.
  Ist $h_{ii} = 0$, dann ist auch $h_{ij} = 0$ für alle j, und es gibt dann keinen Beitrag von $\boldsymbol{y}$ auf $\widehat{y}_{i_{LS}}$.
  Ist $h_{ii} = 1$, dann ist $h_{ij} = 0$ für alle $j \neq i$, und man erhält eine exakte Anpassung $\widehat{y}_{i_{LS}} = y_i$.

Bei einer exakten Anpassung kann man davon ausgehen, dass es sich um einen Hebelpunkt handelt, weil ein extrem starker Einfluss auf die eigene Schätzung ausgeübt wird. I.A. sollte Vorsicht geboten werden bei großen Werten h_{ii}. Als **Faustregel** könnte eine Schranke von $h_{ii} > 2 \cdot \frac{q+1}{n}$ zur Identifizierung von Hebelpunkten genommen werden.

Beispiel 9: Wir betrachten die Daten von Hawkins, Bradu und Kass aus Beispiel 5, die so konstruiert wurden, dass die ersten 14 Werte Hebelpunkte sind. Laut unserem obigen Kriterium sind Werte dann verdächtig auf Hebelpunkte, wenn die Diagonalelemente der Hat-Matrix den Wert $2 \cdot \frac{q+1}{n} = 2 \cdot \frac{4}{75} = 0.107$ überschreiten. In Abbildung 4.14 sind die Werte h_{ii} gegen den Index $i = 1, \ldots, 75$ aufgetragen. Die Schranke 0.107 ist als strichlierte Linie eingezeichnet. Man erkennt, dass nur die Werte 12, 13 und 14 als potentielle Hebelpunkte erkannt werden, die aber "gute" Hebelpunkte darstellen (d.h. deren $\boldsymbol{x}_i$ Werte sind Ausreißer, aber die entsprechenden y_i Werte folgen dem Modell).

4.4.2 Robuste Distanz zur Identifikation von Hebelpunkten

In obigem Beispiel wurden durch die Verwendung der Hat-Matrix die "schlechten" Hebelpunkte maskiert, da das Kriterium offenbar nicht robust ist. Der Grund dafür ist leicht einzusehen. Man kann nämlich zeigen, dass bei Regression mit konstantem Term gilt:

$$h_{ii} = \frac{1}{n-1}\text{MD}_i^2 + \frac{1}{n} \tag{4.9}$$

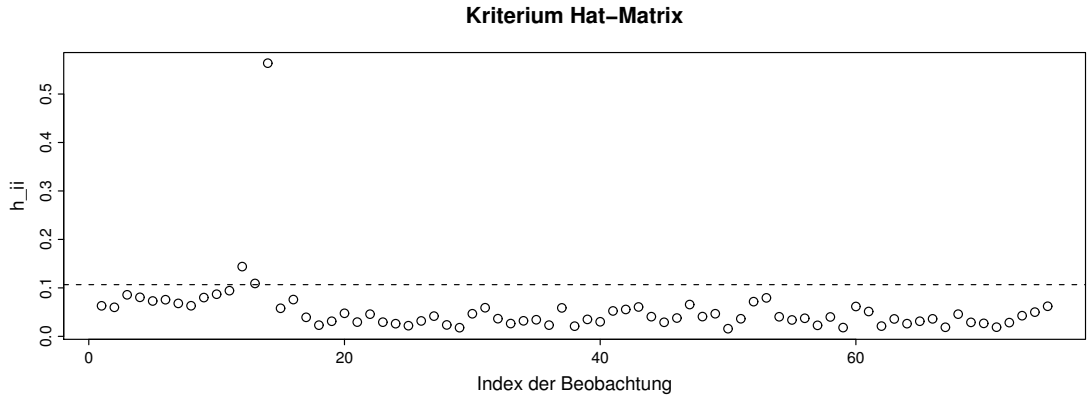


Abbildung 4.14: Daten von Hawkins-Bradu-Kass: Werte über der strichlierten Linie sind laut Hat-Matrix Kriterium potentielle Hebelpunkte.

Die Diagonalelemente der Hat-Matrix sind somit direkt proportional zur quadrierten Mahalanobis-Distanz. Von den Eigenschaften der Mahalanobis-Distanz folgt somit, dass die Hat-Matrix nur dann nicht durch Ausreißer maskiert wird, wenn es einen einzelnen Ausreißer in x-Richtung gibt, wenn aber keine Gruppe von x-Ausreißern vorliegt.

Bemerkung: Die Werte h_{ii} ziehen nur die x-Werte, nicht aber die y-Werte in Betracht. Es wird also nicht zwischen guten und schlechten Hebelpunkten unterschieden.

Beispiel 10: Um die Beziehung (4.9) plausibel zu machen, werden für die Hawkins-Bradu-Kass Daten die Mahalanobis-Distanzen berechnet und analog zu Abbildung 4.14 in Abbildung 4.15 visualisiert. Man erkennt das gleiche Bild wie bei Verwendung der Hat-Matrix, und erhält somit die gleiche verzerrte Aussage bezüglich Hebelpunkte.

Um dem Maskierungseffekt bei der Hat-Matrix bzw. bei der Mahalanobis-Distanz zu entgehen, muss das Kriterium robustifiziert werden. Es liegt aber nun auf der Hand, wie dies gemacht wird, nämlich einfach mit der Robusten Distanz.

Beispiel 11: Wir berechnen für die Hawkins-Bradu-Kass Daten die Robusten Distanzen RD_i, indem Lokation und Kovarianz durch den MCD-Schätzer berechnet werden. Das Resultat ist in Abbildung 4.16 dargestellt. Alle Hebelpunkte 1-14 werden nun richtig durch hohe Werte RD_i angezeigt.

4.4.3 Diagnostik bei Regression

In den vorigen Abschnitten haben wir nun verschiedene Fälle von Ausreißern und Hebelpunkten bei Regression kennengelernt. Wir können daher 4 Typen von Datenwerten unterscheiden:

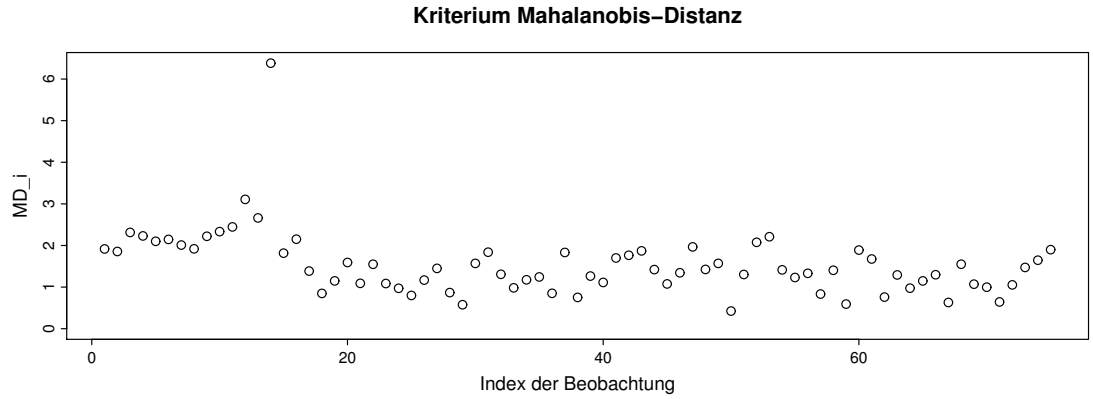


Abbildung 4.15: Daten von Hawkins-Bradu-Kass: Mahalanobis-Distanzen

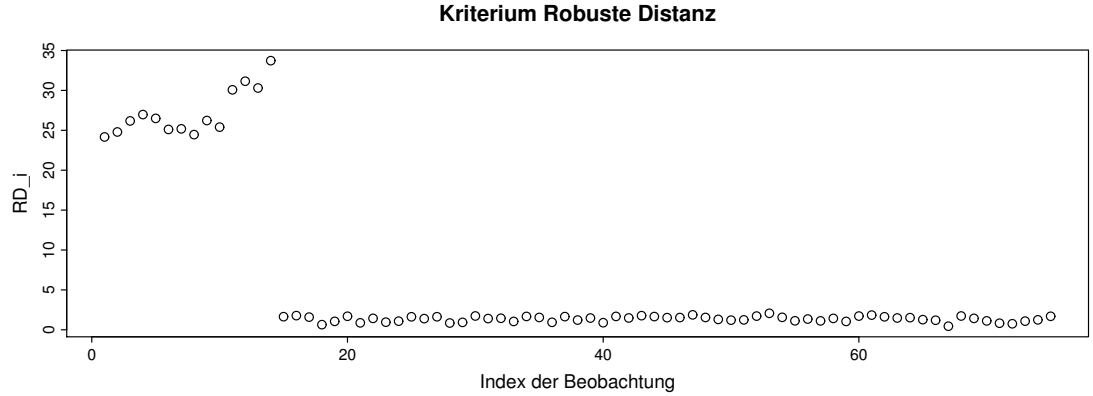


Abbildung 4.16: Daten von Hawkins-Bradu-Kass: Robuste Distanzen

- Reguläre Beobachtungen: $\boldsymbol{x}_i$ im Datenbereich und y_i zum Modell passend
- Vertikale Ausreißer: $\boldsymbol{x}_i$ im Datenbereich aber y_i nicht zum Modell passend
- Gute Hebelpunkte: Ausreißer bei $\boldsymbol{x}_i$ aber y_i zum Modell passend
- Schlechte Hebelpunkte: Ausreißer bei $\boldsymbol{x}_i$ und y_i nicht zum Modell passend

Diese 4 Typen werden im Fall von einfacher linearer Regression in Abbildung 4.17 illustriert. Im Allgemeinen bringen gute Hebelpunkte einen Vorteil, weil sie in Richtung der Regressionsgeraden liegen und damit die Schätzung der Regressionsparameter sogar noch präziser machen. Schlechte Hebelpunkte können starken Einfluss auf die Schätzung der Parameter haben, weil sie die Regressionsgerade zum Kippen bringen können.

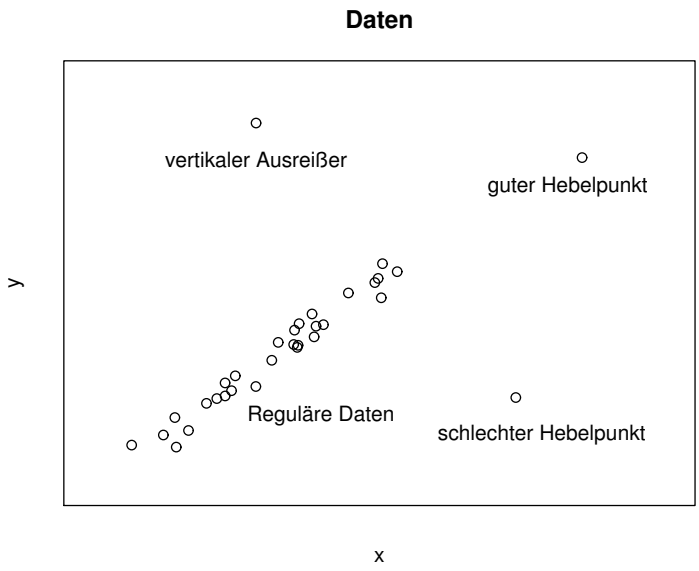


Abbildung 4.17: 4 Typen von Beobachtungen bei einfacher linearer Regression.

Wir haben in den vorhergehenden Abschnitten den Diagnostik-Plot kennengelernt, bei dem die standardisierten LMS- oder LTS-Residuen $\varepsilon_i/\hat{\sigma}$ gegen den Index der Beobachtung aufgetragen werden. Wir haben aber auch den Distance-Distance Plot kennengelernt, bei dem die Robuste Distanz der Mahalanobis-Distanz gegenübergestellt wird. Zur Unterscheidung der erwähnten 4 Typen von Beobachtungen können beide Plots kombiniert werden.

Regression Diagnostic Plot: Ausreißer sind bei Regression gekennzeichnet durch einen hohen Wert der standardisierten robusten Residuen (außerhalb der Schranken ± 2.5). Hebelpunkte sind x-Ausreißer, die somit große Robuste Distanz haben. Der *Regression Diagnostic Plot* macht eine Gegenüberstellung der standardisierten robusten Residuen und der Robusten Distanz.

Abbildung 4.18 zeigt den Regression Diagnostic Plot von den Daten aus Abbildung 4.17. Die regulären Beobachtungen, aber auch die guten Hebelpunkte, findet man im Band ± 2.5 der standardisierten robusten Residuen. Die Unterscheidung, ob eine Beobachtung Hebelpunkt oder kein Hebelpunkt ist, wird aufgrund des Wertes $\sqrt{\chi^2_{q;0.975}}$ bei der Robusten Distanz gemacht.

Beispiel 12: Zu den Phosphor-Daten von Beispiel 6 wurden auch y-Werte ermittelt, nämlich der Phosphor-Gehalt in Maispflanzen, die in den Böden mit den entsprechenden Werten (x_{i1}, x_{i2}) von Abbildung 4.10 gewachsen sind. Der Regression Diagnostik Plot ist in Abbildung 4.19 dargestellt. In der linken Grafik sieht man den Plot mit den robusten Schätzungen, in der rechten Grafik ist zum Vergleich der Plot mit den klassischen Schätzungen gemacht worden. Analog zu Abbildung 4.11 bzw. 4.12 werden x-Ausreißer aufgrund eines Wertes der Robusten Distanz größer als $\sqrt{\chi^2_{2;0.975}} = 2.72$ erkannt. Die Beobachtungen 1 und 6 sind somit Hebelpunkte.

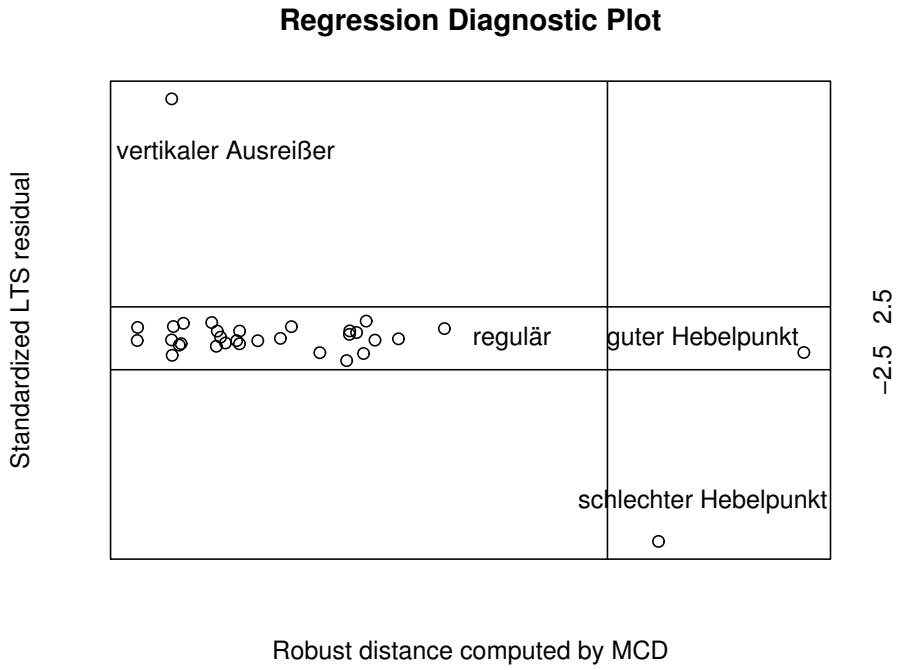


Abbildung 4.18: Identifikation der 4 Typen von Beobachtungen im Regression Diagnostic Plot.

Sie sind allerdings gute Hebelpunkte, weil sie innerhalb des Bandes ± 2.5 liegen. Der Wert 17 wird als vertikaler Ausreißer identifiziert. Auch im Plot mit den klassischen Schätzungen wird Wert 17 als vertikaler Ausreißer erkannt, es werden aber keine Hebelpunkte gefunden, diese werden maskiert in diesem "Diagnostik"-Plot.

Beispiel 13: Auch zu den *Stackloss*-Daten aus Beispiel 8 gibt es y-Werte. Der Regression Diagnostic Plot ist in Abbildung 4.20 gezeigt, wiederum mit den robusten (links) und den klassischen (rechts) Schätzungen. Wir erkennen natürlich die auch in Abbildung 4.13 mit der Schranke $\sqrt{\chi^2_{3;0.975}} = 3.06$ identifizierten x-Ausreißer als Hebelpunkte. Die Werte 1, 3 und 21 sind dabei schlechte Hebelpunkte. Ein Wert (Nr. 4) wird als vertikaler Ausreißer erkannt. Man bemerke, dass im "klassischen Plot" rechts alle Werte als reguläre Beobachtungen eingestuft werden.

Literatur

N.R. Draper and H. Smith. *Applied Regression Analysis*. Wiley & Sons, New York, 1981.

D.M. Hawkins, D. Bradu, and G.V. Kass. Location of several outliers in multiple regression data using elemental sets. *Technometrics*, 26, 197-208, 1984.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, London, 4th edition, 1998.

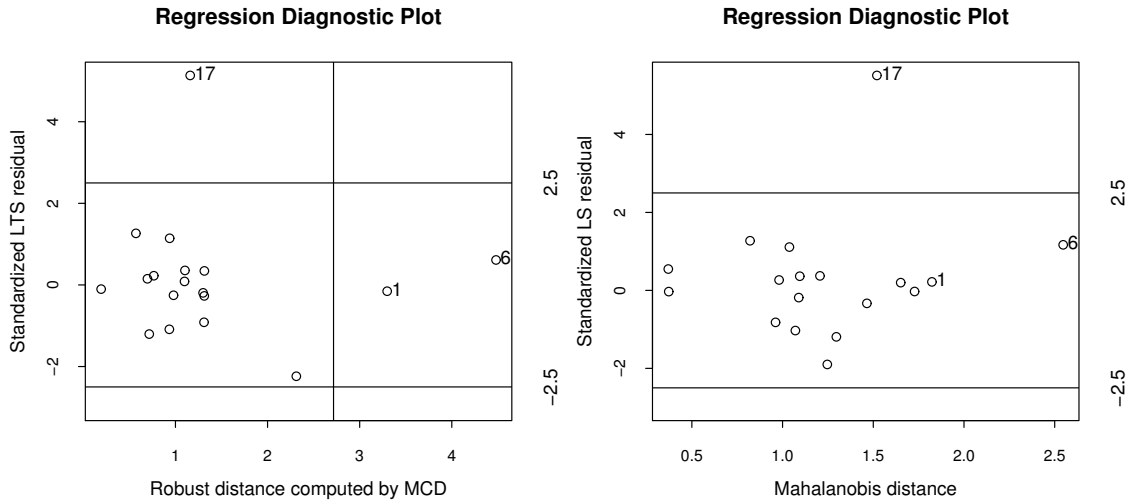


Abbildung 4.19: Regression Diagnostik Plot der Phosphor-Daten: links mit den robusten Schätzungen, rechts mit den klassischen Schätzungen.

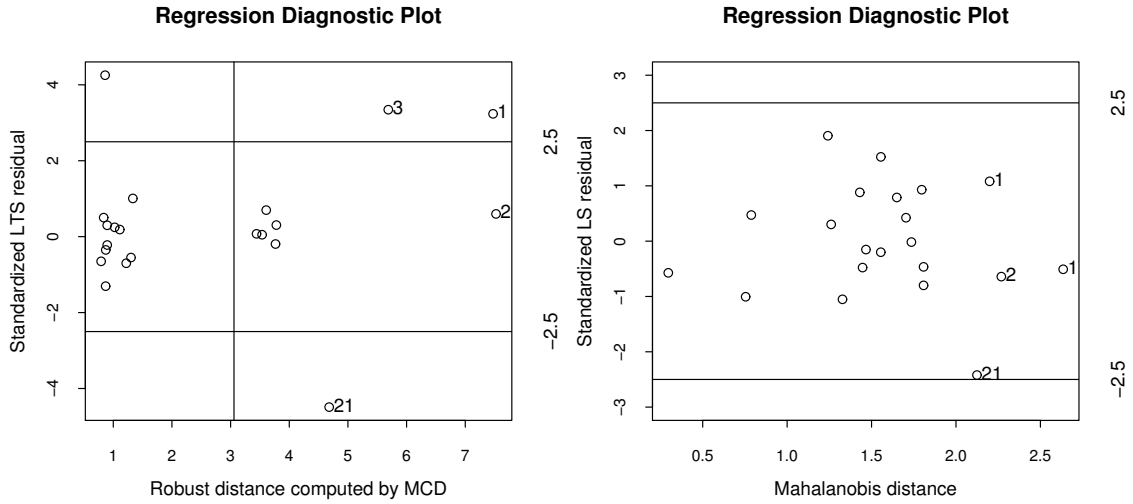


Abbildung 4.20: Regression Diagnostik Plot der Stackloss-Daten: links mit den robusten Schätzungen, rechts mit den klassischen Schätzungen.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Acad. Press, London, 1979.

P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 871-880, 1984.

P.J. Rousseeuw. Multivariate estimation with high breakdown point. In: W. Grossmann, G. Pflug, I. Vincze, and W. Wertz (eds.), *Mathematical Statistics and Applications*, Volume B. Akadémiai Kiadó, Budapest, pp. 283-297, 1985.

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

G.A.F. Seber. *Multivariate Observations.* John Wiley & Sons, New York, 1984.

Kapitel 5

Hauptkomponentenanalyse

5.1 Einleitung

Die Hauptkomponentenanalyse (Principal Component Analysis, *PCA*) hat ihren Ursprung bei Karl Pearson (1901) und wurde von Harrold Hotelling (1933) entwickelt.

Ziel dieser Methode ist es, komplizierte Beziehungen in beobachteten Daten auf eine einfache Form zu reduzieren. Die Daten werden durch Linearkombinationen von gewissen Komponenten so dargestellt, dass möglichst wenig Information verloren geht. Die Dimension wird somit reduziert auf die Anzahl dieser Komponenten.

Die Hauptkomponentenanalyse dient auch zur Ausreißererkennung. Dazu werden alle verschiedenen Paare von Komponenten einander gegenübergestellt und diese Projektionen untersucht.

Die Hauptkomponentenanalyse gehört zu den wichtigsten Methoden in der multivariaten Statistik. Sie ist einerseits relativ einfach in der Handhabung und Interpretation und bildet auf der anderen Seite den Grundstein für viele multivariate Methoden.

5.2 Bestimmung von Hauptkomponenten in der Population

Sei $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ ein p-dimensionaler Zufallsvektor mit Mittelvektor $\boldsymbol{\mu}$ und Kovarianzmatrix

$$\boldsymbol{\Sigma} = E\Big[(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top\Big] \ . \tag{5.1}$$

Sei weiters $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$ eine $(p \times p)$-Matrix mit festen Koeffizienten mit der Bedingung, dass $\boldsymbol{\gamma}_i$ Einheitsvektoren sind, d.h. $\boldsymbol{\gamma}_i^\top \boldsymbol{\gamma}_i = 1$ für $i = 1, \ldots, p$. Außerdem sei $\boldsymbol{\Gamma}$ orthogonal, also $\boldsymbol{\Gamma}^\top = \boldsymbol{\Gamma}^{-1}$. Dann betrachten wir die Lineartransformation

$$\boldsymbol{z} = \boldsymbol{\Gamma}^\top (\boldsymbol{x} - \boldsymbol{\mu}) \tag{5.2}$$

oder, ausgedrückt in Komponenten,

$$z_i = \boldsymbol{\gamma}_i^\top (\boldsymbol{x} - \boldsymbol{\mu}) \qquad \text{für} \qquad i = 1, \ldots, p \ . \tag{5.3}$$

Das Resultat der obigen Transformation ist ein neuer Zufallsvektor $\boldsymbol{z}$ der Dimension p. Die Varianz von z_i $(i = 1, \ldots, p)$ ist

$$\mathrm{Var}(z_i) = E\Big[\boldsymbol{\gamma}_i^\top (\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\gamma}_i\Big] = \boldsymbol{\gamma}_i^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_i \ . \tag{5.4}$$

Wir möchten eine Transformation erhalten, sodass die Varianz jeweils maximiert wird mit der Nebenbedingung von Einheitsvektoren $\boldsymbol{\gamma}_i$. Mathematisch kann man diese Forderung als Maximierung eines Lagrangeschen Ausdrucks mit entsprechender Randbedingung formulieren.

1. Hauptkomponente:

Für $i = 1$ möchte man also eine Komponente z_1 erhalten, sodass $\mathrm{Var}(z_1)$ maximiert wird, und $\boldsymbol{\gamma}_1^\top \boldsymbol{\gamma}_1 = 1$ ist. Das entsprechende Lagrange-Problem lautet somit:

$$\phi_1 = \boldsymbol{\gamma}_1^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_1 - a_1(\boldsymbol{\gamma}_1^\top \boldsymbol{\gamma}_1 - 1) \tag{5.5}$$

Die partiellen Ableitungen nach den Unbekannten $\boldsymbol{\gamma}_1$ werden Null gesetzt. Man erhält

$$\frac{\partial \phi_1}{\partial \boldsymbol{\gamma}_1} = 2\boldsymbol{\Sigma}\boldsymbol{\gamma}_1 - 2a_1\boldsymbol{\gamma}_1 = \mathbf{0} \tag{5.6}$$

bzw.

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_1 = a_1\boldsymbol{\gamma}_1 \ . \tag{5.7}$$

Die Lösung des Maximierungsproblems ergibt somit, dass $\boldsymbol{\gamma}_1$ Eigenvektor von $\boldsymbol{\Sigma}$ zum Eigenwert a_1 ist. Allerdings hat $\boldsymbol{\Sigma}$ insgesamt p Eigenvektoren. Welchen sollte man nun nehmen?

Mit (5.7) sieht man rasch, dass

$$\mathrm{Var}(z_1) = \boldsymbol{\gamma}_1^\top(\boldsymbol{\Sigma}\boldsymbol{\gamma}_1) = \boldsymbol{\gamma}_1^\top(a_1\boldsymbol{\gamma}_1) = a_1\boldsymbol{\gamma}_1^\top\boldsymbol{\gamma}_1 = a_1$$

ist, und nachdem diese Varianz maximiert werden sollte, nimmt man jenen Eigenvektor $\boldsymbol{\gamma}_1$ zum größten Eigenwert a_1. Die Komponente z_1 heißt dann *erste Hauptkomponente*, und $\boldsymbol{\gamma}_1$ ist die Richtung dieser Komponente.

2. Hauptkomponente:

Im nächsten Schritt, für $i = 2$, möchte man wiederum $\mathrm{Var}(z_2)$ maximieren, unter der Bedingung $\boldsymbol{\gamma}_2^\top \boldsymbol{\gamma}_2 = 1$. Zusätzlich fordert man, dass z_1 und z_2 unkorreliert sind. Letztere Bedingung bedeutet:

$$Cov(z_1, z_2) = Cov(\boldsymbol{\gamma}_1^\top(\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{\gamma}_2^\top(\boldsymbol{x} - \boldsymbol{\mu})) = Cov(\boldsymbol{\gamma}_1^\top\boldsymbol{x}, \boldsymbol{\gamma}_2^\top\boldsymbol{x}) = \boldsymbol{\gamma}_1^\top\boldsymbol{\Sigma}\boldsymbol{\gamma}_2 =$$

$$= \boldsymbol{\gamma}_2^\top\boldsymbol{\Sigma}\boldsymbol{\gamma}_1 = \boldsymbol{\gamma}_2^\top a_1\boldsymbol{\gamma}_1 = a_1\boldsymbol{\gamma}_2^\top\boldsymbol{\gamma}_1 = 0$$

Nachdem $a_1 \neq 0$ ist, ist Unkorreliertheit gleichbedeutend mit Orthogonalität von $\boldsymbol{\gamma}_1$ und $\boldsymbol{\gamma}_2$.

Man kann nun wieder das Lagrange Problem formulieren,

$$\phi_2 = \boldsymbol{\gamma}_2^\top \boldsymbol{\Sigma} \boldsymbol{\gamma}_2 - a_2(\boldsymbol{\gamma}_2^\top \boldsymbol{\gamma}_2 - 1) - b\boldsymbol{\gamma}_2^\top \boldsymbol{\gamma}_1 \tag{5.8}$$

mit den Lagrange Koeffizienten a_2 und b. Die partiellen Ableitungen nach den Unbekannten $\boldsymbol{\gamma}_2$ werden Null gesetzt. Man erhält

$$\frac{\partial \phi_2}{\partial \boldsymbol{\gamma}_2} = 2\boldsymbol{\Sigma}\boldsymbol{\gamma}_2 - 2a_2\boldsymbol{\gamma}_2 - b\boldsymbol{\gamma}_1 = \mathbf{0} \tag{5.9}$$

Multiplikation von links mit $\boldsymbol{\gamma}_1^\top$ ergibt

$$2\boldsymbol{\gamma}_1^\top\boldsymbol{\Sigma}\boldsymbol{\gamma}_2 - 2a_2\boldsymbol{\gamma}_1^\top\boldsymbol{\gamma}_2 - b\boldsymbol{\gamma}_1^\top\boldsymbol{\gamma}_1 = 0 - 0 - b \cdot 1 = 0 \ ,$$

und damit muss $b = 0$ sein. Somit reduziert sich (5.9) auf

$$\boldsymbol{\Sigma}\boldsymbol{\gamma}_2 = a_2\boldsymbol{\gamma}_2 \ ,$$

und somit ist $\boldsymbol{\gamma}_2$ Eigenvektor von $\boldsymbol{\Sigma}$ zum nächst größten Eigenwert a_2, und z_2 wird als *zweite Hauptkomponente* bezeichnet.

Weitere Hauptkomponenten:

Die k-te Hauptkomponente ($2 < k \leq p$) wird analog zu oben definiert, nämlich über Maximierung von $\mathrm{Var}(z_k)$ und der Bedingung $\boldsymbol{\gamma}_k^\top\boldsymbol{\gamma}_k = 1$, sowie Unkorreliertkheit zu allen vorigen Hauptkomponenten, also $Cov(z_k, z_j) = 0$ für $k > j$, gleichbedeutend mit Orthogonalität $\boldsymbol{\gamma}_k^\top\boldsymbol{\gamma}_j = 0$.

Wenig überraschend ist die Lösung, dass $\boldsymbol{\gamma}_k$ Eigenvektor von $\boldsymbol{\Sigma}$ zum k-t größten Eigenwert a_k ist.

Sammelt man nun alle gefundenen Eigenvektoren als Spalten in der Matrix $\boldsymbol{\Gamma}$, also $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p)$, und die entsprechenden Eigenwerte $a_1, \ldots, a_p$ der Größe nach sortiert ($a_1 \geq a_2 \geq \ldots \geq a_p \geq 0$) in der Diagonale der Matrix $\boldsymbol{A}$, also $\boldsymbol{A} = \mathrm{Diag}(a_1, \ldots, a_p)$, dann kann die Hauptkomponenten-Lösung in Matrixschreibweise ausgedrückt werden als

$$\boldsymbol{\Sigma}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{A} \tag{5.10}$$

oder auch $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^\top$, siehe Spektralzerlegungssatz.

Mit diesen Definitionen ist die Lineartransformation

$$\boldsymbol{z} = \boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu}) \tag{5.11}$$

bekannt unter dem Namen *Hauptkomponententransformation*, das i-te Element z_i des Vektors $\boldsymbol{z}$ nennt man *i-te Hauptkomponente* (*HK*).

Der Erwartungswert der *HK* ist Null, da

$$E(\boldsymbol{z}) = \boldsymbol{\Gamma}^\top\Big[E(\boldsymbol{x} - \boldsymbol{\mu})\Big] = \mathbf{0}, \tag{5.12}$$

und die Kovarianz ist

$$\mathrm{Cov}(\boldsymbol{z}) = \boldsymbol{\Gamma}^\top\mathrm{Cov}(\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^\top\boldsymbol{\Sigma}\boldsymbol{\Gamma} = \boldsymbol{A} \ . \tag{5.13}$$

Die Varianz der i-ten *HK* entspricht also a_i, dem i-ten Eigenwert von $\boldsymbol{\Sigma}$. Verschiedene *HK* sind unkorreliert, weil $\boldsymbol{A}$ eine Diagonalmatrix ist. Die Gesamtvarianz der

HK ist die Summe aller Eigenwerte. Dies ist natürlich auch gleichzeitig die Gesamtvarianz oder Gesamtvariation von $\boldsymbol{x}$.

Die Matrix $\boldsymbol{\Gamma}$, die die Beziehung zwischen $\boldsymbol{x}$ und $\boldsymbol{z}$ herstellt, wird auch *Ladungsmatrix* (engl. loadings matrix) genannt. Deren Elemente γ_{ij} widerspiegeln den Einfluss von x_i auf z_j.

Ein interessantes Maß für den Zusammenhang zwischen $\boldsymbol{x}$ und $\boldsymbol{z}$ ist auch die Korrelation. Speziell das Quadrat dieser Korrelation stellt als Bestimmtheitsmaß die erklärte Variation von z_j durch x_i dar. Die Kovarianz zwischen den Variablen und den *HK* ist

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{x}, \boldsymbol{z}) &= E\Big[(\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{z}^\top\Big] \\ &= E\Big[\boldsymbol{\Gamma}\boldsymbol{z}\boldsymbol{z}^\top\Big] \\ &= \boldsymbol{\Gamma}\boldsymbol{A} \end{aligned} \tag{5.14}$$

oder

$$\mathrm{Cov}(x_i, z_j) = \gamma_{ij} a_j \qquad \text{für} \qquad i, j = 1, \ldots, p\ . \tag{5.15}$$

Somit ist die Korrelation zwischen Variablen und *HK*

$$Corr(x_i, z_j) = \lambda_{ij} = \frac{\mathrm{Cov}(x_i, z_j)}{\sqrt{\mathrm{Var}(x_i)}\sqrt{\mathrm{Var}(z_j)}} = \frac{\gamma_{ij} a_j}{\sigma_{ii}^{\frac{1}{2}} a_j^{\frac{1}{2}}} = \gamma_{ij}\sqrt{\frac{a_j}{\sigma_{ii}}} \tag{5.16}$$

oder

$$Corr(\boldsymbol{x}, \boldsymbol{z}) = \boldsymbol{\Lambda} = \Big(\mathrm{Diag}(\boldsymbol{\Sigma})\Big)^{-\frac{1}{2}} \boldsymbol{\Gamma}\boldsymbol{A}^{\frac{1}{2}}\ . \tag{5.17}$$

Bemerkung: Nach obigen Aussagen erkennt man sofort, dass die *HK*-Transformation nicht skaleninvariant ist und daher die Resultate von den Einheiten abhängen. Oft standardisiert man daher vorher die Variablen auf Mittel 0 und Varianz 1,

$$y_i = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}} \quad \text{für} \quad i = 1, \ldots, p, \tag{5.18}$$

um Skaleninvarianz zu erreichen.

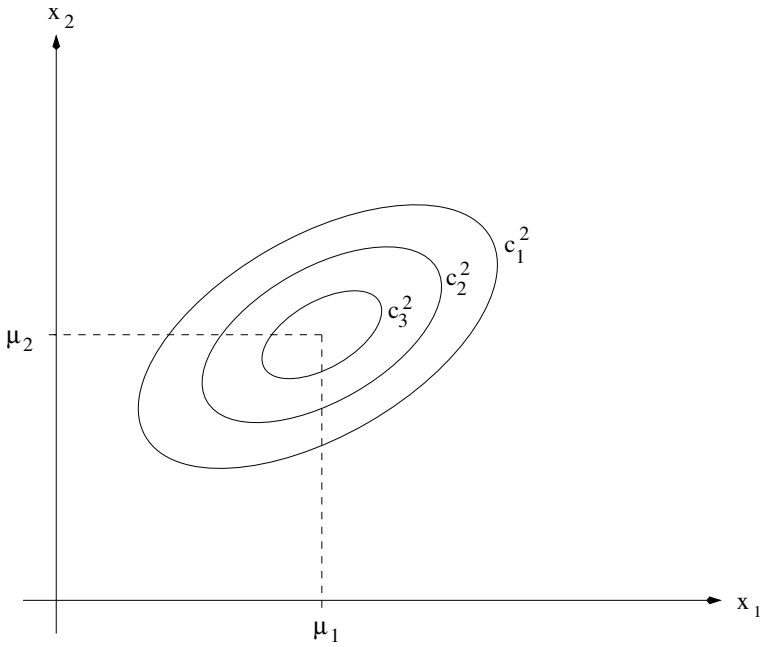
5.3 Geometrische Interpretation von Hauptkomponenten

In Kapitel 1.1 wurde die Mahalanobis-Distanz definiert. Setzt man, wie in (1.27) dieses Distanzmaß gleich einer Konstanten, also

$$(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) := c^2 \quad , \tag{5.19}$$

so wird dadurch ein Ellipsoid im p-dimensionalen Raum mit Zentrum $\boldsymbol{\mu}$ beschrieben. Die Dichte der p-variaten Normalverteilung, die ja im Exponent die Mahalanobis-Distanz enthält, ist also auf solchen Ellipsoiden konstant.

Abbildung 5.1: Ellipsen mit konstanter Dichte bei der zweidimensionalen Normalverteilung



Durch die *HK*-Transformation

$$\boldsymbol{z} = \boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu}) \tag{5.20}$$

erhält man:

$$\begin{aligned} c^2 &= (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Gamma} \boldsymbol{A}^{-1} \boldsymbol{\Gamma}^\top (\boldsymbol{x} - \boldsymbol{\mu}) \\ &= \left[\boldsymbol{\Gamma}^\top(\boldsymbol{x} - \boldsymbol{\mu})\right]^\top \boldsymbol{A}^{-1} \boldsymbol{\Gamma}^\top (\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{z}^\top \boldsymbol{A}^{-1} \boldsymbol{z} \\ &= \sum_{i=1}^{p} \frac{z_i^2}{a_i} = \frac{z_1^2}{a_1} + \frac{z_2^2}{a_2} + \ldots + \frac{z_p^2}{a_p} \quad . \end{aligned}$$

Die Komponenten $(z_1, \ldots, z_p)$ von $\boldsymbol{z}$ (d.h. die *HK*) repräsentieren also die Hauptachsen des Ellipsoids.

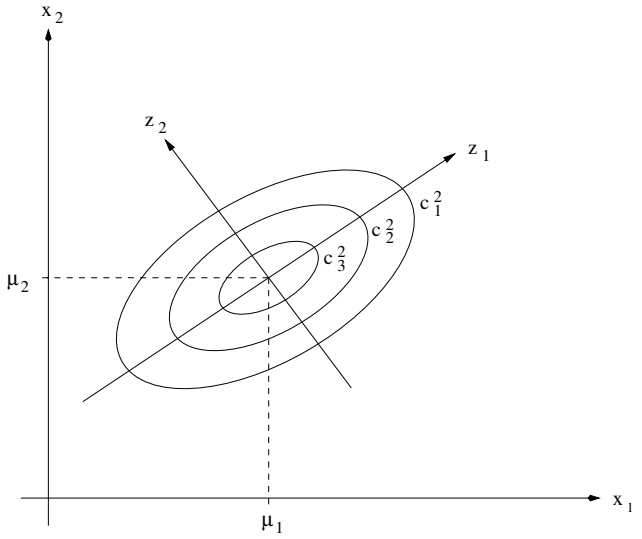
Die erste *HK* z_1 liegt in der größten Ausdehnung des Ellipsoids, die zweite *HK* z_2 liegt in der größten Ausdehnung, die orthogonal zu z_1 ist, die 3. *HK* z_3 liegt in der größten Ausdehnung, die orthogonal zu z_1 und z_2 ist usw. (siehe Abbildung 5.2).

5.4 Hauptkomponenten von Stichproben

Sei nun $\boldsymbol{X}$ eine $(n \times p)$ Datenmatrix. Mittelvektor und Stichproben-Kovarianzmatrix können geschätzt werden durch

$$\widehat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_{i\cdot} \ , \tag{5.21}$$

Abbildung 5.2: Hauptkomponenten im zweidimensionalen Fall



$\boldsymbol{x}_{i.}$ ist die i-te Zeile von $\boldsymbol{X}$, und

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})(\boldsymbol{x}_{i.} - \bar{\boldsymbol{x}})^{\top} \ . \tag{5.22}$$

In Analogie zu (5.11) werden die *HK* der Stichprobe berechnet durch

$$\boldsymbol{Z} = (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^{\top})\hat{\boldsymbol{\Gamma}} \tag{5.23}$$

oder

$$\boldsymbol{z}_j = (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^{\top})\hat{\boldsymbol{\gamma}}_j \qquad \text{für} \qquad j = 1, \ldots, p \ . \tag{5.24}$$

$\mathbf{1}$ ist ein Spaltenvektor bestehend aus n Einsen, und $\hat{\boldsymbol{\Gamma}} = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_p)$ ist die Matrix der Eigenvektoren von $\boldsymbol{S}$. Ähnlich zu den theoretischen *HK* erhalten wir die Zerlegung

$$\hat{\boldsymbol{\Gamma}}^{\top}\boldsymbol{S}\hat{\boldsymbol{\Gamma}} = \hat{\boldsymbol{A}} \ , \tag{5.25}$$

wobei $\hat{\boldsymbol{A}} = \mathrm{Diag}(\hat{a}_1, \ldots, \hat{a}_p)$ eine Diagonalmatrix der Eigenwerte (angeordnet in absteigender Größe) zu den zugehörigen Eigenvektoren von $\boldsymbol{S}$ ist.

Die Matrix $\boldsymbol{Z}$ der *HK* hat natürlich die selbe Dimension wie die Matrix $\boldsymbol{X}$. Die einzelnen *HK* $\boldsymbol{z}_j$ kann man sich als Beobachtungen in einem orthogonal rotierten Raum vorstellen. Die Werte z_{ij} werden meist mit *Werte der HK* oder *scores* bezeichnet.

Analog zu (5.16) werden die Korrelationen zwischen Variablen und *HK* berechnet als

$$\hat{\lambda}_{ij} = \frac{\hat{\gamma}_{ij}\hat{a}_j^{\frac{1}{2}}}{s_{ii}^{\frac{1}{2}}} \qquad \text{für} \qquad i, j = 1, \ldots, p \ , \tag{5.26}$$

wobei s_{ii} das i-te Diagonalelement von $\boldsymbol{S}$ ist. In Matrixschreibweise wird die Korrelation durch

$$\hat{\boldsymbol{\Lambda}} = \Big(\mathrm{Diag}(\boldsymbol{S})\Big)^{-\frac{1}{2}} \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{A}}^{\frac{1}{2}} \tag{5.27}$$

berechnet.

Beispiel 5.4.1 *Gegeben seien wieder die Prüfungsdaten aus Tabelle 2.1. Für diese Daten sollen nun die* HK *bestimmt werden.*

Der Vektor der Spaltenmittel beträgt für diese Daten

$$\bar{\boldsymbol{x}} = (39.0 \quad 50.6 \quad 50.6 \quad 46.7 \quad 42.3)^\top \quad .$$

Die empirische Kovarianzmatrix ergibt sich als

$$\boldsymbol{S} = \begin{pmatrix} 305.8 & 127.2 & 101.6 & 106.3 & 117.4 \\ & 172.8 & 85.2 & 94.7 & 99.0 \\ & & 112.9 & 112.1 & 121.9 \\ & & & 220.4 & 155.5 \\ & & & & 297.8 \end{pmatrix} \quad .$$

Durch eine Spektralzerlegung von $\boldsymbol{S}$ erhält man die Matrix der Eigenvektoren,

$$\hat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.51 & -0.75 & -0.30 & 0.30 & 0.08 \\ 0.37 & -0.21 & 0.42 & -0.78 & 0.19 \\ 0.35 & 0.08 & 0.15 & 0.00 & -0.92 \\ 0.45 & 0.30 & 0.60 & 0.52 & 0.29 \\ 0.53 & 0.55 & -0.60 & -0.18 & 0.15 \end{pmatrix},$$

sowie die Diagonalmatrix der Eigenwerte,

$$\hat{\boldsymbol{A}} = \mathit{Diag}(687.0, 202.1, 103.7, 84.6, 32.2).$$

Die HK *werden dann berechnet durch*

$$\boldsymbol{z}_j = (\boldsymbol{X} - \mathbf{1}\bar{\boldsymbol{x}}^\top)\hat{\boldsymbol{\gamma}}_j = (\boldsymbol{x}_1\hat{\gamma}_{1j} + \ldots + \boldsymbol{x}_p\hat{\gamma}_{pj}) - \mathbf{1}\bar{\boldsymbol{x}}^\top\hat{\boldsymbol{\gamma}}_j,$$

bzw. eingesetzt

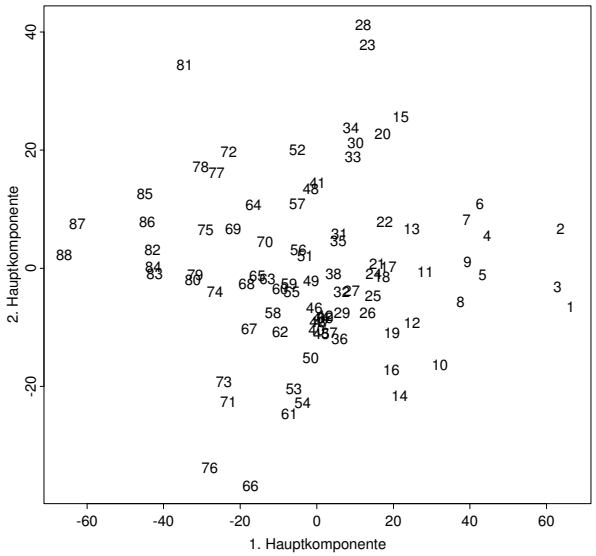
$$\begin{aligned}
\boldsymbol{z}_1 &= 0.51\boldsymbol{x}_1 + 0.37\boldsymbol{x}_2 + 0.35\boldsymbol{x}_3 + 0.45\boldsymbol{x}_4 + 0.53\boldsymbol{x}_5 + 99.5 \cdot \mathbf{1} \\
\boldsymbol{z}_2 &= -0.75\boldsymbol{x}_1 - 0.21\boldsymbol{x}_2 + 0.08\boldsymbol{x}_3 + 0.30\boldsymbol{x}_4 + 0.55\boldsymbol{x}_5 + 1.4 \cdot \mathbf{1} \\
\boldsymbol{z}_3 &= -0.30\boldsymbol{x}_1 + 0.42\boldsymbol{x}_2 + 0.15\boldsymbol{x}_3 + 0.60\boldsymbol{x}_4 - 0.60\boldsymbol{x}_5 + 19.2 \cdot \mathbf{1} \\
\boldsymbol{z}_4 &= 0.30\boldsymbol{x}_1 - 0.78\boldsymbol{x}_2 + 0.00\boldsymbol{x}_3 + 0.52\boldsymbol{x}_4 - 0.18\boldsymbol{x}_5 - 11.5 \cdot \mathbf{1} \\
\boldsymbol{z}_5 &= 0.08\boldsymbol{x}_1 + 0.19\boldsymbol{x}_2 - 0.92\boldsymbol{x}_3 + 0.29\boldsymbol{x}_4 + 0.15\boldsymbol{x}_5 - 14.4 \cdot \mathbf{1}
\end{aligned}$$

Die Varianzen der HK *entsprechen den Eigenwerten,*

$$\begin{aligned}
\widehat{a}_1 &= \widehat{Var}(\boldsymbol{z}_1) = 687.0 \\
\widehat{a}_2 &= \widehat{Var}(\boldsymbol{z}_2) = 202.1 \\
\widehat{a}_3 &= \widehat{Var}(\boldsymbol{z}_3) = 103.7 \\
\widehat{a}_4 &= \widehat{Var}(\boldsymbol{z}_4) = 84.6 \\
\widehat{a}_5 &= \widehat{Var}(\boldsymbol{z}_5) = 32.2 \quad .
\end{aligned}$$

In Abbildung 5.3 sind die ersten beiden HK *einander gegenübergestellt.*

Abbildung 5.3: 1. und 2. Hauptkomponente der Prüfungsdaten



5.5 Anzahl der relevanten Hauptkomponenten

Eine Methode, die Anzahl k der relevanten *HK* festzustellen, ist ein Test auf Gleichheit der letzten (d.h. kleinsten) $(p-k)$ Eigenwerte, also $a_p = a_{p-1} = \ldots = a_{k+1}$. Die letzten $(p-k)$ *HK* beschreiben somit gleiche Variation und enthalten daher gleich viel Information.

In der Praxis macht man oft eine Folge dieses Tests: Man startet mit $k = 0$ und erhöht k so lange, bis die Nullhypothese nicht mehr abgelehnt werden kann. Man kommt auf folgende Teststatistik:

$$\left(n - \frac{2p+11}{6}\right)(p-k)\ln\left(\frac{m_a}{m_g}\right) \sim \chi^2_{(p-k+2)(p-k-1)/2} \tag{5.28}$$

mit

$$m_a = \frac{\hat{a}_{k+1} + \ldots + \hat{a}_p}{p-k}\ , \qquad m_g = \sqrt[p-k]{\hat{a}_{k+1} \ldots \hat{a}_p}\ . \tag{5.29}$$

Wenn man mit standardisierten Variablen arbeitet, müssen die Eigenwerte von der Korrelationsmatrix $\boldsymbol{\rho}$ bestimmt werden. Eine Schätzung für $\boldsymbol{\rho}$ ist die Stichproben-Korrelationsmatrix $\boldsymbol{R}$. Ein Test für die Hypothese, daß die kleinsten $(p-k)$ Eigenwerte von $\boldsymbol{R}$ gleich sind, ist dem obigen Test ähnlich. Die Teststatistik für diesen Fall ist

$$(n-1)(p-k)\ln\left(\frac{m_a}{m_g}\right) \sim \chi^2_{(p-k+2)(p-k-1)/2} \quad , \tag{5.30}$$

wobei m_a und m_g arithmetisches und geometrisches Mittel der letzten $(p-k)$ Eigenwerte von $\boldsymbol{R}$ sind.

Beispiel 5.5.1 *Gegeben seien wiederum die Prüfungsdaten. Zuerst untersuchen wir, ob die letzten vier Eigenwerte gleich sind, also*

$$H_0: \qquad a_2 = a_3 = a_4 = a_5 \ .$$

Durch Anwendung der Formel für die Teststatistik erhält man eine χ^2_9*–Verteilung. Der Wert der Teststatistik ist* 26.11 *. Diese Nullhypothese ist signifikant zum* 1%*–Niveau und wird verworfen. Die neue Nullhypothese lautet nun*

$$H_0: \qquad a_3 = a_4 = a_5 \quad .$$

Man erhält eine χ^2_5*–Verteilung und der Wert der Teststatistik ist* 7.30 *. H_0 ist nicht einmal signifikant zum* 5%*–Niveau und wird daher angenommen.*

5.5.1 „Faustregeln“

In der Literatur gibt es viele Regeln für die richtige Wahl der Anzahl von *HK*, die nicht statistisch begründet sind. Trotz dieses Mangels sind die Kriterien dafür weit verbreitet und werden auch in Statistik-Programmpaketen verwendet.

Eine Kenngröße, die die Entscheidung über die relevante Anzahl von *HK* erleichtern soll, ist der oben beschriebene Anteil der ersten k *HK* an der Gesamtvariation. Demnach sollte man so viele *HK* berücksichtigen, dass dieser Anteil zumindest 90% ist,

$$\frac{\sum_{j=1}^{k} \hat{a}_j}{\sum_{i=1}^{p} \hat{a}_i} \geq 90\% \ . \tag{5.31}$$

Ein anderes Kriterium besagt, dass man jene *HK* ausschließen soll, deren Eigenwerte geringer als der Durchschnitt sind.

Beispiel 5.5.2 *Betrachten wir wieder unsere Prüfungsdaten, so berechnet man den Anteil jeder* HK *bzw. den Anteil der ersten j* HK *als*

$\frac{\hat{a}_j}{\sum_i \hat{a}_i}$ in %	61.9	18.2	9.3	7.6	2.9
gesamt %	61.9	80.2	89.5	97.1	100

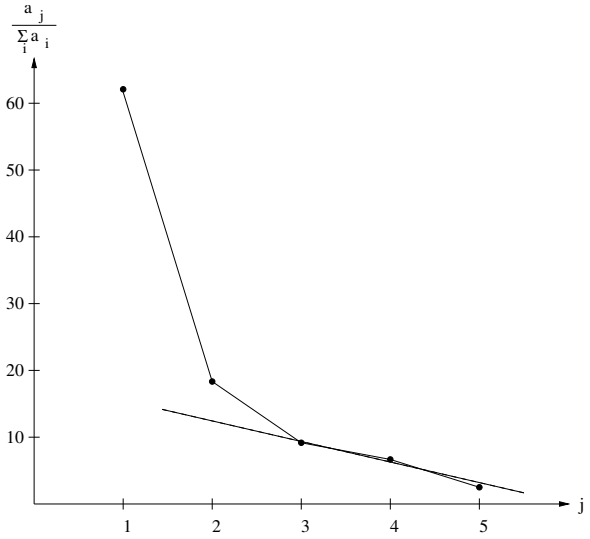
Nach dem 90%-Kriterium sollte man somit die ersten vier HK *nehmen und nur die 5. Komponente weglassen. Das Kriterium mit dem Durchschnitt würde nur die erste Komponente einschließen.*

Eine weitere Möglichkeit zum Bestimmen der maßgeblichen Variation bietet der *Scree-Graph* (“Geröll”-Grafik), der eine Gegenüberstellung vom Anteil jeder *HK* an der Gesamtvariation zur jeweiligen Nummer der *HK* darstellt. Jene Punkte, die auf einer Geraden liegen, werden weggelassen.

Beispiel 5.5.3 *Im Fall der Prüfungsdaten ergibt sich der Scree-Graph in Abbildung 5.4.*

Die Komponenten 3, 4 *und* 5 *werden weggelassen.*

Abbildung 5.4: Scree-Graph zu den Prüfungsdaten mit 5 HK



Eine andere Möglichkeit, den Anteil der Variation jeder Variable durch die *HK* zu erkennen, ist der durch (5.26) beschriebene Korrelationskoeffizient $\hat{\lambda}_{jk}$ der Variable $\boldsymbol{x}_j$ mit der Komponente $\boldsymbol{z}_k$, bzw. das Quadrat dieser Zahl (Bestimmtheitsmaß). Sind die quadrierten Korrelationen mit den letzten *HK* nur mehr sehr gering, so können diese Komponenten ohne viel Informationsverlust weggelassen werden.

Beispiel 5.5.4 *Bei den Prüfungsdaten ist der Anteil der letzten beiden* HK *an der Gesamtvariation* 10.5%. *Gefragt ist nun, wieviel Information verloren geht, wenn man diese Komponenten weglässt. Es ergeben sich folgende Werte für* $\hat{\lambda}^2_{jk}$ *(j sind die Variablen, k die Komponenten):*

$\hat{\lambda}^2_{jk}$	$\boldsymbol{z}_1$	$\boldsymbol{z}_2$	$\boldsymbol{z}_3$	$\boldsymbol{z}_4$	$\boldsymbol{z}_5$
$\boldsymbol{x}_1$	0.574	0.371	0.030	0.024	0.001
$\boldsymbol{x}_2$	0.539	0.050	0.104	0.300	0.007
$\boldsymbol{x}_3$	0.727	0.018	0.019	0.000	0.243
$\boldsymbol{x}_4$	0.634	0.083	0.168	0.103	0.012
$\boldsymbol{x}_5$	0.660	0.204	0.125	0.009	0.002

Die erste HK *erklärt* 57.4% *von* $\boldsymbol{x}_{.1}$, 53.9% *von* $\boldsymbol{x}_2$, 72.7% *von* $\boldsymbol{x}_3$, 63.4% *von* $\boldsymbol{x}_4$ *und* 66.0% *von* $\boldsymbol{x}_5$. *Die ersten beiden* HK *erklären* 94.5% *von* $\boldsymbol{x}_1$ *etc. Man beachte, dass die letzte Komponente* 24.3% *der Variation von* $\boldsymbol{x}_3$, *aber nur sehr wenig von den restlichen Variablen erklärt. Wenn man daher die letzte* HK *weglässt, verliert man viel mehr Information von der 3. Variablen als von den anderen. Wenn man die letzten beiden* HK *weglässt, verliert man die meiste Information von* $\boldsymbol{x}_2$ *und* $\boldsymbol{x}_3$, *die wenigste von Variable* 5.

In der Praxis wird oft ein Kompromiss zwischen verschiedenen Kriterien eingegangen.

Es gibt auch „modernere“ Verfahren zur Bestimmung der Anzahl von relevanten *HK* wie sogenannte *Resampling*-Methoden. Das sind nichtparametrische Schätzmethoden, die die Voraussetzung der Normalverteilung nicht benötigen (außer beim Testen). Die bekanntesten Resampling-Methoden sind *Jackknife* (Efron, 1982), *Bootstrap* (Efron, 1979; 1981) und *Cross Validation* (Stone, 1974; Eastment und Krzanowski, 1982).

5.6 Singulärwertzerlegung

Die Singulärwertzerlegung oder *singular value decomposition (SVD)* ist ein alternativer Algorithmus zur Berechnung der Hauptkomponenten, der keine Schätzung der Kovarianzmatrix benötigt. Dies hat vor allem Vorteile bei Daten mit $n < p$, wo die geschätzte Kovarianzmatrix immer singulär ist, und damit die Hauptkomponenten der Ordnung $> n$ automatisch Varianz 0 haben.

Wir gehen im folgenden davon aus, dass die Datenmatrix $\boldsymbol{X}$ eine *zentrierte* $(n \times p)$-Matrix mit reellen Zahlen ist. Es sind also die Spaltenmittel 0. Dann existiert eine orthogonale $(n \times n)$-Matrix $\boldsymbol{U}$ und eine orthogonale $(p \times p)$-Matrix $\boldsymbol{V}$, sodass

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \tag{5.32}$$

gilt. $\boldsymbol{D}$ ist eine $(n \times p)$-Matrix mit $d_{ii} \geq 0$ für $i = 1, \ldots, \min(n, p)$, die restlichen Elemente sind null. Die positiven Werte d_{ii} werden *Singulärwerte von* $\boldsymbol{X}$ genannt. Die Anzahl dieser positiven Werte entspricht dem Rang von $\boldsymbol{X}$.
Wenn der Rang von $\boldsymbol{X}$ gleich $k < min(n, p)$ ist, dann kann $\boldsymbol{X}$ auch dargestellt werden durch

$$\boldsymbol{X} = \sum_{i=1}^{k} d_{ii}\boldsymbol{u}_i\boldsymbol{v}_i^\top \ . \tag{5.33}$$

$\boldsymbol{u}_i$ bzw. $\boldsymbol{v}_i$ bezeichnet die i-te Spalte von $\boldsymbol{U}$ bzw. $\boldsymbol{V}$. Durch die Orthogonalität von $\boldsymbol{U}$ und $\boldsymbol{V}$ ergibt sich

$$\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{u}_i = d_{ii}^2\boldsymbol{u}_i \tag{5.34}$$

und

$$\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{v}_i = d_{ii}^2\boldsymbol{v}_i \ . \tag{5.35}$$

Daraus erkennt man, dass $\boldsymbol{u}_i$ der i-te Eigenvektor von $\boldsymbol{X}\boldsymbol{X}^\top$ zum Eigenwert d_{ii}^2, und $\boldsymbol{v}_i$ der i-te Eigenvektor von $\boldsymbol{X}^\top\boldsymbol{X}$ zum selben Eigenwert d_{ii}^2 ist. Die Eigenwerte für $i = 1, \ldots, k$ sind dabei größer als 0, die restlichen sind null.
Zusammenfassend heißt das, dass $\boldsymbol{U}$ n orthogonale Eigenvektoren von $\boldsymbol{X}\boldsymbol{X}^\top$ in den Spalten hat, und dass $\boldsymbol{V}$ p orthogonale Eigenvektoren von $\boldsymbol{X}^\top\boldsymbol{X}$ in den Spalten hat.

Man kann nun leicht den Zusammenhang zu einer Kovarianz-basierten Schätzung der Hauptkomponenten herstellen, so wie sie in Abschnitt 5.4 durchgeführt wurde. Wenn wir wiederum zentrierte Daten betrachten, dann waren die *HK* definiert als

$\boldsymbol{Z} = \boldsymbol{X}\hat{\boldsymbol{\Gamma}}$, und damit ist $\boldsymbol{X} = \boldsymbol{Z}\hat{\boldsymbol{\Gamma}}^\top$. Weiters ist die empirische Kovarianzmatrix in diesem Fall

$$\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{X}^\top\boldsymbol{X}.$$

Die Matrix $\hat{\boldsymbol{\Gamma}}$ ist die Matrix der normierten Eigenvektoren von $\boldsymbol{S}$. Bei *SVD* ist $\boldsymbol{V}$ die Matrix der normierten Eigenvektoren von $\boldsymbol{X}^\top\boldsymbol{X}$, woraus wir schließen können, dass $\hat{\boldsymbol{\Gamma}} \equiv \boldsymbol{V}$ ist. Aus Gleichung (5.35) schließen wird, dass $d_{ii}^2 = (n-1)\hat{a}_i$ ist. Wir haben somit die Darstellung

$$\boldsymbol{X} = \boldsymbol{Z}\boldsymbol{V}^\top = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top, \tag{5.36}$$

und damit ist $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}$.

Mit diesen Überlegungen kann die Berechnung der Hauptkomponenten auch über eine andere Zielfunktion formuliert werden. Vorerst definieren wir noch die Frobenius Norm: Seien $\boldsymbol{x}_{i.}$ die Zeilen der Matrix $\boldsymbol{X}$, für $i = 1, \ldots, n$. Dann ist die Frobenius Norm von $\boldsymbol{X}$ definiert als

$$\|\boldsymbol{X}\|_F = \sqrt{\sum_{i=1}^{n} \|\boldsymbol{x}_{i.}\|_2^2} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} x_{ij}^2}$$

Laut oben haben wir

$$\boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D} = \boldsymbol{Z}.$$

Sei $\boldsymbol{V}_m = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m)$, und m kleiner als der Rang von $\boldsymbol{X}$. Dann stellen $\boldsymbol{X}\boldsymbol{V}_m$ genau die ersten m *HK* dar, was gleichbedeutend ist mit der Projektion von $\boldsymbol{X}$ auf einen m-dimensionalen Unterraum. Man kann zeigen, dass

$$\boldsymbol{V}_m = arg\max_{\boldsymbol{B}} \|\boldsymbol{X}\boldsymbol{B}\|_F^2$$

für eine beliebige $p \times m$ Matrix $\boldsymbol{B}$ mit $rank(\boldsymbol{B}) \leq m$ und $\boldsymbol{B}^\top\boldsymbol{B} = \boldsymbol{I}$.

Eine äquivalente Formulierung ist folgende: Wir haben

$$\boldsymbol{X} = \boldsymbol{X}\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{X}\boldsymbol{V}_m\boldsymbol{V}_m^\top + \boldsymbol{E}.$$

Dann gilt

$$\boldsymbol{V}_m = arg\min_{\boldsymbol{B}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top\|_F^2$$

mit gleicher Definition von $\boldsymbol{B}$ wie oben. Man bemerke, dass $\hat{\boldsymbol{X}} := \boldsymbol{X}\boldsymbol{B}\boldsymbol{B}^\top$ die gleiche Dimension wie $\boldsymbol{X}$ hat. Man kann somit $\hat{\boldsymbol{X}}$ als Rang m Approximation von $\boldsymbol{X}$ auffassen, die in obigem Sinne optimal ist.

5.7 Biplots

Ein *Biplot* ist eine zweidimensionale grafische Darstellung von Objekten und Variablen in einer einzigen Abbildung. Biplots wurden von Gabriel (1971) entwickelt.

Das "Bi" in der Bezeichnung Biplot bezieht sich nicht auf die Dimension 2, sondern darauf, dass beide Größen (Objekte und Variablen) gleichzeitig repräsentiert werden. Selbstverständlich sind auch dreidimensionale Biplots denkbar, wenn auch etwas unpraktisch (Gabriel, 1986).

Wir werden hier Biplots kennenlernen, die den Zweck haben, Loadings und Scores von der Hauptkomponentenanalyse grafisch in einem Plot darzustellen. Es gibt auch andere Methoden von Biplots wie z.B. Multidimensionale Skalierung, Korrespondenzanalyse, Kanonische Biplots, Nichtlineare Biplots etc. Ein gutes Buch über diese Gebiete ist Gower und Hand (1996).

Zweidimensionale Plots sind gut handhabbar und übersichtlich. Eine Projektion der Daten auf zwei Dimensionen sollte allerdings auch voraussetzen, dass die Daten annähernd Rang 2 haben, dass also mit zwei Dimensionen der Großteil der Variabilität dargestellt werden kann. Für eine Datenmatrix mit höherem Rang sollten die ersten beiden Hauptkomponenten die Daten möglichst gut repräsentieren.

Sei nun $\boldsymbol{X}$ eine zentrierte $(n \times p)$-Matrix mit Rang k. Ein Biplot zeigt eine Darstellung von $\boldsymbol{X}$ durch zwei Gruppen von Vektoren der Dimension n und p, die eine Approximation von $\boldsymbol{X}$ mit Rang 2 sind. Diese Rang-2-Approximation bezeichnen wir mit $\boldsymbol{X}_{(2)}$. Eine Approximation von $\boldsymbol{X}$ durch $\boldsymbol{X}_{(2)}$ im Sinne der Kleinsten Quadrate ist gegeben durch die ersten beiden Hauptkomponenten von $\boldsymbol{X}$, die entweder über die Schätzung der Kovarianzmatrix oder über SVD gewonnen werden können. Wir werden hier formal auf die Schätzung mit Singulärwertzerlegung eingehen. Wir benötigen also nur die ersten beiden Singulärwerte und die ersten beiden Spalten von $\boldsymbol{U}$ und $\boldsymbol{V}$ (nachfolgend trotzdem mit $\boldsymbol{U}$ und $\boldsymbol{V}$ bezeichnet):

$$\boldsymbol{X} \approx \boldsymbol{X}_{(2)} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_1^\top \\ \boldsymbol{v}_2^\top \end{pmatrix} \ . \tag{5.37}$$

$\boldsymbol{X}_{(2)}$ könnte aber auch anders aufgespalten werden, nämlich durch

$$\boldsymbol{X}_{(2)} = \boldsymbol{G}\boldsymbol{H}^\top \tag{5.38}$$

mit

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{1-c} \tag{5.39}$$

und

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix}^{c} \tag{5.40}$$

für $0 \leq c \leq 1$. Je nach Wahl von c werden also die ersten beiden Singulärwerte auf die Matrizen $\boldsymbol{G}$ und $\boldsymbol{H}$ "aufgeteilt". Ein Biplot besteht nun genau aus den Zeilen der Matrizen $\boldsymbol{G}$ und $\boldsymbol{H}$, also aus $n + p$ zweidimensionalen Vektoren.

Für eine Wahl von $c = 0.5$ werden die Vektoren für Objekte und Variablen auf gleiche Skalierung gebracht.

Für $c = 1$ und Umskalierung,

$$\boldsymbol{G} = \begin{pmatrix} \boldsymbol{g}_{1.}^\top \\ \vdots \\ \boldsymbol{g}_{n.}^\top \end{pmatrix} = \sqrt{n-1} \begin{pmatrix} \boldsymbol{u}_1 & \boldsymbol{u}_2 \end{pmatrix} \ , \tag{5.41}$$

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{h}_{1.}^\top \\ \vdots \\ \boldsymbol{h}_{p.}^\top \end{pmatrix} = \frac{1}{\sqrt{n-1}} \begin{pmatrix} \boldsymbol{v}_1 & \boldsymbol{v}_2 \end{pmatrix} \begin{pmatrix} d_{11} & 0 \\ 0 & d_{22} \end{pmatrix} , \tag{5.42}$$

erhalten wir folgende Eigenschaften (Um die Notation zu vereinfachen, betrachten wir die Variablen als zentriert, also $\frac{1}{n}\sum_{i=1}^n x_{ij} = 0$ für $j = 1, \ldots, p.$):

- Das innere Produkt zwischen den Zeilen von $\boldsymbol{G}$ und den Zeilen von $\boldsymbol{H}$ approximieren die Datenmatrix $\boldsymbol{X}$,

$$\boldsymbol{g}_{i.}^\top \boldsymbol{h}_{j.} = \sqrt{n-1}\, \boldsymbol{u}_{i.}^\top \frac{1}{\sqrt{n-1}} \left(\boldsymbol{v}_{j.}^\top \boldsymbol{D}\right)^\top = \boldsymbol{u}_{i.}^\top \boldsymbol{D} \boldsymbol{v}_{j.} \approx x_{ij} . \tag{5.43}$$

- Die inneren Produkte zwischen den Zeilen von $\boldsymbol{H}$ approximiert die Kovarianz,

$$\begin{aligned} \boldsymbol{H}\boldsymbol{H}^\top &= \left(\frac{1}{\sqrt{n-1}}\, \boldsymbol{V}\boldsymbol{D}\right)\left(\frac{1}{\sqrt{n-1}}\, \boldsymbol{D}\boldsymbol{V}^\top\right) = \frac{1}{n-1}\, \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top \\ &= \frac{1}{n-1}\, \left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^\top\right)\left(\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top\right) = \frac{1}{n-1}\, \boldsymbol{X}_{(2)}^\top \boldsymbol{X}_{(2)} \\ &\approx \frac{1}{n-1}\, \boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{S} . \end{aligned} \tag{5.44}$$

  Daraus folgt, dass die quadrierte Euklidische Norm der Zeilen von $\boldsymbol{H}$, $\|\boldsymbol{h}_{j.}\|^2 = \boldsymbol{h}_{j.}^\top \boldsymbol{h}_{j.}$, die Varianz approximiert. Darüber hinaus approximiert der Kosinus zwischen $\boldsymbol{h}_{i.}$ und $\boldsymbol{h}_{j.}$ $(i, j = 1, \ldots, p)$ die Korrelation zwischen den Variablen,

$$\cos(\boldsymbol{h}_{i.}, \boldsymbol{h}_{j.}) = \frac{\boldsymbol{h}_{i.}^\top \boldsymbol{h}_{j.}}{\|\boldsymbol{h}_{i.}\| \|\boldsymbol{h}_{j.}\|} \approx r_{ij} . \tag{5.45}$$

  Als Konsequenz von (5.44) ergibt sich

$$\frac{1}{n-1} \sum_{l=1}^{n} (x_{li} - x_{lj})^2 = \frac{1}{n-1}\, \left(\boldsymbol{x}_{.i} - \boldsymbol{x}_{.j}\right)^\top \left(\boldsymbol{x}_{.i} - \boldsymbol{x}_{.j}\right) \approx \|\boldsymbol{h}_{i.} - \boldsymbol{h}_{j.}\|^2 . \tag{5.46}$$

  D.h., dass die quadrierte Euklidische Distanz zwischen den Zeilen von $\boldsymbol{H}$ die mittlere quadrierte Differenz zwischen den Variablen approximiert.

- Die Euklidische Distanz zwischen den Zeilen von $\boldsymbol{G}$ approximiert die Mahalanobisdistanz zwischen den Beobachtungen,

$$\begin{aligned} \|\boldsymbol{g}_{i.} - \boldsymbol{g}_{j.}\|^2 &= \left(\boldsymbol{g}_{i.} - \boldsymbol{g}_{j.}\right)^\top \left(\boldsymbol{g}_{i.} - \boldsymbol{g}_{j.}\right) = (n-1)\left(\boldsymbol{u}_{i.} - \boldsymbol{u}_{j.}\right)^\top \left(\boldsymbol{u}_{i.} - \boldsymbol{u}_{j.}\right) \\ &\approx \left(\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\right)^\top \boldsymbol{S}^{-1} \left(\boldsymbol{x}_{i.} - \boldsymbol{x}_{j.}\right) , \end{aligned} \tag{5.47}$$

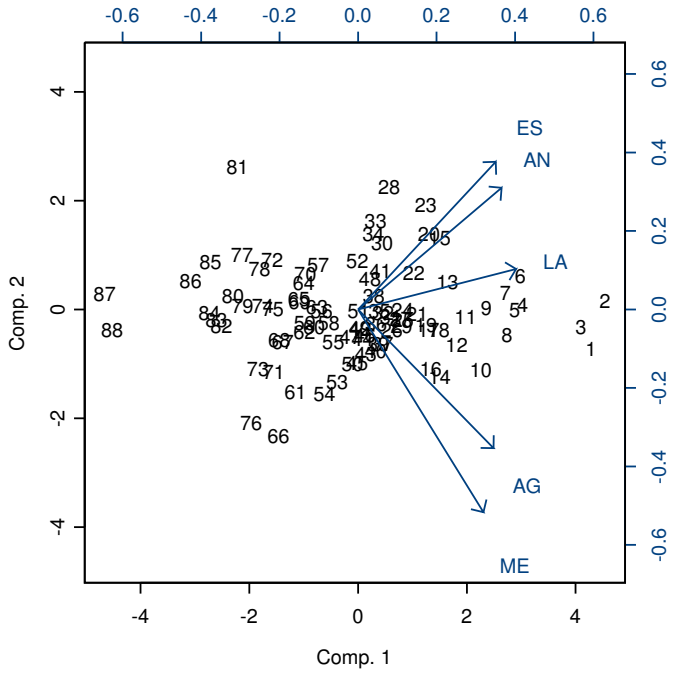
  da

$$\boldsymbol{x}_{i.}^\top \boldsymbol{S}^{-1} \boldsymbol{x}_{j.} \approx \left(\boldsymbol{u}_{i.}^\top \boldsymbol{D}\boldsymbol{V}^\top\right)(n-1)\left(\boldsymbol{V}\boldsymbol{D}^{-2}\boldsymbol{V}^\top\right)\left(\boldsymbol{V}\boldsymbol{D}\boldsymbol{u}_{j.}\right) = (n-1)\boldsymbol{u}_{i.}^\top \boldsymbol{u}_{j.} \tag{5.48}$$

  für $i, j = 1, \ldots, n$.

Beispiel 5.7.1 *Wir betrachten wiederum die Prüfungsdaten und erstellen damit einen Biplot (siehe Abbildung 5.5). Wir stellen fest, dass die Fächer ES und AN (und eventuell noch LA) in relativ starker Beziehung stehen, ebenso die Fächer AG und ME. Weiters ist ersichtlich, dass die x-Richtung zwischen guten und schlechten Resultaten generell unterscheidet, während die y-Richtung den Übergang von "mathematischen" zu "geometrischen" Fächern definiert. Die Einordnung jedes Studenten ist aus der Position im Plot (eigentlich aus der Projektion auf den Variablenvektor) leicht möglich. Z.B. ist Student 81 in den Fächern ME und AG sehr schlecht, bei den anderen Fächern liegt er im Mittelfeld.*

Abbildung 5.5: Biplot der skalierten Prüfungsdaten



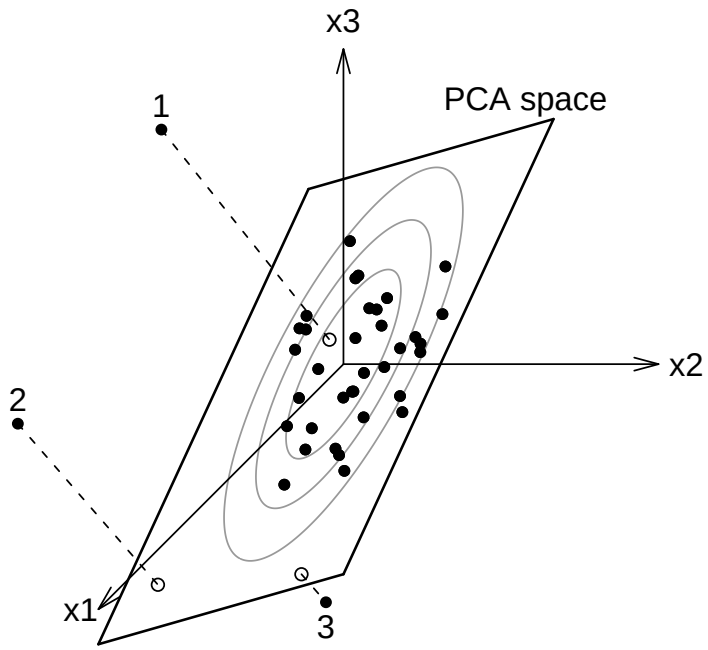
5.8 Diagnostik

Obwohl Hauptkomponentenanalyse keine typische Methodik zur multivariaten Ausreißererkennung darstellt, könnte man dennoch daran interessiert sein, ob einzelne Beobachtungen ungewöhnlich sind im Sinne von Abweichung von der Datenstruktur der Mehrheit der Daten. Dies wird durch zwei Distanz-Maße ermittelt, nämlich die *Score*-Distanz (SD) und die orthogonale Distanz (OD).

Abbildung 5.6 versucht dies, schematisch zu erläutern. Hier hat man Beobachtungen im 3-dimensionalen gegeben, und es wird der Raum der ersten beiden Hauptkomponenten visualisiert. Die (quadrierte) SD gibt die Mahalanobis-Distanz der Beobachtungen projiziert in diesen Raum an. Diese Distanz wird auch durch die

Ellipsen dargestellt. Die OD gibt die Distanz orthogonal zu diesem Raum an. Die meisten Beobachtungen sind im Raum der ersten beiden Hauptkomponenten gut dargestellt. Beobachtung 1 hat große OD, aber kleine SD. Beobachtung 2 hat große OD und große SD, und Beobachtung 3 hat große SD aber kleine OD. Analog zur Diagnostik in der Regression kann man hier Parallelen ziehen zu vertikalen Ausreißern (1), sowie guten (3) und schlechten (2) Hebelpunkten. Insbesondere schlechte Hebelpunkte können die klassische Schätzung der Hauptkomponenten stark beeinflussen. Daher ist diese Diagnostik erst sinnvoll für robust geschätzte Hauptkomponenten.

Abbildung 5.6: Diagnostik-Plot für Hauptkomponenten



Formal werden diese Distanzen folgendermaßen definiert. Sei $\hat{\mathbf{\Gamma}}_k = (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_k)$, also die Matrix der ersten k geschätzten HK *loadings*. Weiters sei $\boldsymbol{z}_{i.} = (z_{i1}, \ldots, z_{ik})^\top$ der i-te *score* Vektor der geschätzten HK, und $\hat{a}_1, \ldots, \hat{a}_k$ die entsprechenden geschätzten HK Varianzen. Die Anzahl k von HK wird üblicherweise so gewählt, dass in etwa 80% der Varianz erklärt sind.

Die SD für die i-te Beobachtung ($i = 1, \ldots, n$) ist definiert als

$$SD_i = \left(\sum_{j=1}^{k} \frac{z_{ij}^2}{\hat{a}_j} \right)^{1/2}.$$

Dies ist gleich mit der Mahalanobis-Distanz des *score* Vektors zum Zentrum der HK bezogen auf deren Kovarianzmatrix.

Die OD für die i-te Beobachtung $\boldsymbol{x}_{i.}$ $(i = 1, \ldots, n)$ ist definiert als

$$OD_i = \|\boldsymbol{x}_{i.} - \hat{\boldsymbol{\Gamma}}_k \boldsymbol{z}_{i.}\|_2,$$

mit der euklidischen Norm $\|\cdot\|_2$, also als euklidische Distanz der Beobachtung zur Projektion in den Raum der ersten k HK.

Ähnlich wie bei multivariater Ausreißererkennung kann man Grenzen (cutoff-Werte) für beide Distanzen bestimmen, die auf ungewöhnlich hohe SD und OD hinweisen würden. Nachdem SD eine Mahalanobis-Distanz ist, ist ein geeigneter cutoff-Wert $\sqrt{\chi^2_{k;0.975}}$. Der cutoff-Wert für die OD ergibt sich über eine Approximation als

$$\left(\text{median}_i(OD_i^{2/3}) + \text{MAD}_i(OD_i^{2/3}) z_{0.975}\right)^{3/2},$$

wobei $z_{0.975}$ das 0.975-Quantil der $N(0,1)$ ist. MAD steht für Median Absolute Deviation, und ist für univariate Werte $y_1, \ldots, y_n$ definiert als

$$\text{MAD} = 1.483 \cdot \text{median}_i(|\ y_i - \text{median}_j(y_j)\ |).$$

Beispiel 5.8.1 *Für unsere Prüfungsdaten können diese Diagnostik-Plots sehr einfach konstruiert werden. Wir verwenden nachfolgend eine robuste Methode zur Schätzung der HK, implementiert als Funktion* `PcaHubert()` *im Paket* `rrcov` *– für Details siehe help-Seiten, sowie klassische HK mit* `princomp()`*. Der R Code ist folgend:*

```
library(bootstrap)
data(scor) # Daten
library(rrcov)
res1 <- PcaHubert(scor,scale=TRUE)  # robust
plot(res1)

X <- scale(scor)
X.pca <- princomp(X)                # nicht robust
library(chemometrics)
res <- pcaDiagplot(X,X.pca,a=3,plot=TRUE) # 3 HK
```

Die resultierenden Diagnostik-Plots sind in Abbildung 5.7 dargestellt. Die robuste Analyse (links) zeigt ein paar mehr ungewöhnliche Beobachtungen im Vergleich zur klassischen Analyse (rechts).

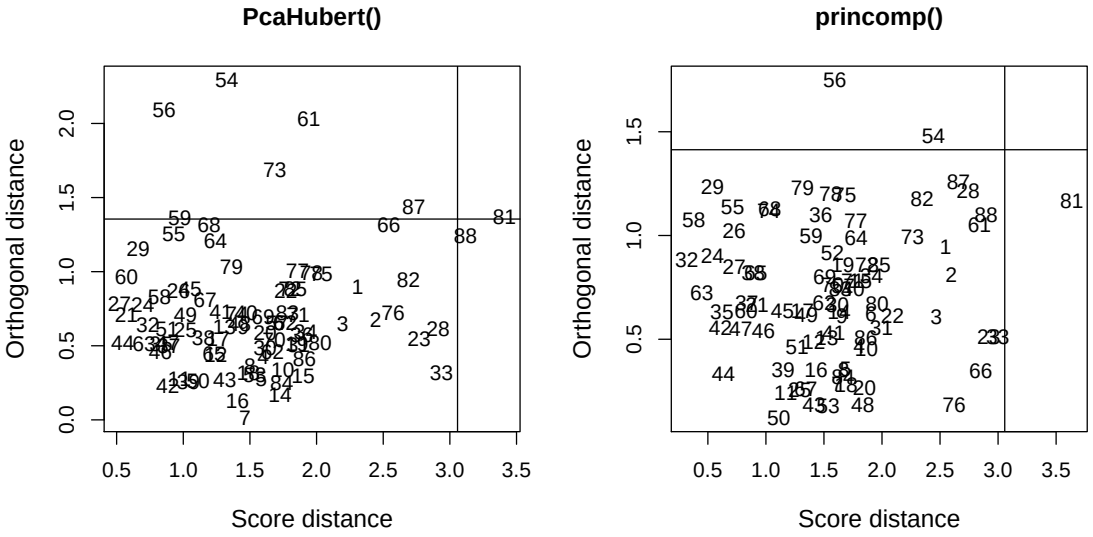
Literatur

A. Basilevsky. *Statistical Factor Analysis and Related Methods: Theory and Applications*. Wiley & Sons, New York, 1994.

T.F. Cox and A.A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994.

H.T. Eastment and W.J. Krzanowski. Cross-validatory choice of the number of components from a principal component analysis. *Technometrics*, 24:73–77, 1982.

Abbildung 5.7: Diagnostik-Plots für die Prüfungsdaten: links robust, rechts nicht robust



B. Efron. Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.*, 7:1–26, 1979.

B. Efron. Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap, and Other Methods. *Biometrika*, 68:589–599, 1981.

B. Efron. *The Jackknife, the Bootstrap and Other Resampling Plans.* J.W. Arrowsmith Ltd., Bristol, 1982.

K.R. Gabriel. The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

K.R. Gabriel and C.L. Odoroff. Use of three-dimensional biplots for diagnosis of models. In W. Gaul and M. Schader, editors, *Classification as a Tool of Research*, pages 153–159, 1986.

J.C. Gower and D.J. Hand. *Biplots.* Chapman & Hall, London, 1996.

J.C. Gower and S.A. Harding. Nonlinear biplots. *Biometrika*, 75(3):445–455, 1988.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24:417–441, 498–520, 1933.

J.E. Jackson. *A User's Guide To Principal Components.* Wiley & Sons, New York, 1991.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 4th edition, 1998.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Acad. Press, London, 1979.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Phil. Mag. (6)*, 2:559–572, 1901.

M. Stone. Cross-validatory choice and assessment of statistical prediction. *J. R. Stat. Soc. B*, 36:111–133, 1974.

Kapitel 6

Faktorenanalyse

6.1 Einleitung

Betrachtet man an n Objekten (Personen) p Merkmale (Körpergröße, Gewicht, Alter, ...), so werden die Merkmale in der Regel voneinander abhängen, d.h. auch miteinander korrelieren. Es ist anzunehmen, dass sie einander wechselseitig beeinflussen, oder dass eine dritte Größe, die man nicht direkt messen kann, den Einfluss bestimmt (z.B. Statur).

Das faktorenanalytische Modell geht immer davon aus, dass das Messbare nur eine Erscheinungsform von Größen ist, die im Hintergrund stehen, und die man direkt nicht messen kann. In vielen Fällen ist diese Annahme realistisch. Es ist daher von Interesse, ob sich aus den beobachteten Merkmalen eine Größe isolieren lässt, ein sogenannter Faktor, der die Zusammenhänge erklärt.

Das Hauptziel der Faktorenanalyse ist die Ableitung hypothetischer Größen oder Faktoren aus einer Menge beobachteter Merkmale. Die Faktoren sollen möglichst einfach sein und die Beobachtungen hinreichend genau (interpretierbar) beschreiben und erklären.

Der klassische Bereich der Faktorenanalyse ist die Theorie der Intelligenz. Man kann zunächst annehmen, dass sich Intelligenz aus zahlreichen Elementarfähigkeiten wie Gedächtnis, Aufmerksamkeit, Urteilsfähigkeit usw. zusammensetzt. Solche Größen versucht man durch psychologische Testverfahren zu messen, um dann mit Hilfe der Faktorenanalyse den Faktor „Intelligenz“ zu isolieren. Anhand dieses Faktors kann die beobachtete Personengruppe leichter unterteilt werden als auf Grund vieler einzelner Testergebnisse.

6.2 Das Faktorenmodell

6.2.1 Definition

Sei $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ ein stochastischer Vektor, $x_1, x_2, \ldots, x_p$ die Merkmale bzw. Variablen.

Sei weiters $\boldsymbol{y} = (y_1, y_2, \ldots, y_p)^\top$ der Vektor, der durch

$$y_i = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}}, \qquad i = 1, \ldots, p \quad ,$$

auf Mittel Null und Varianz Eins standardisierten Merkmale.

Es wird angenommen, dass die Elemente von $\boldsymbol{y}$, abgesehen von einem Fehlerterm $\boldsymbol{e}$, durch eine kleinere Anzahl $k < p$ von unbekannten Zufallsvariablen $\boldsymbol{f} = (f_1, \ldots, f_k)^\top$ dargestellt werden können. Dann lässt sich das k-Faktorenmodell folgendermaßen angeben:

$$\boldsymbol{y} = \boldsymbol{\Lambda f} + \boldsymbol{e} \quad . \tag{6.1}$$

Dabei ist $\boldsymbol{\Lambda} = (\lambda_{ij})$ eine $(p \times k)$-Matrix mit konstanten Werten. Sie wird *Ladungsmatrix* genannt und beschreibt den Zusammenhang zwischen Faktoren und Merkmalen.

Je nach Größe und Verteilung der *Ladungen* λ_{ij} $(i = 1, \ldots, p)$ spricht man bei f_j von einem

- *allgemeinen Faktor* (engl. *general factor*), wenn fast *alle* der oben angeführten Ladungen deutlich von Null verschieden sind;
- *gemeinsamen Faktor* (engl. *common factor*), wenn *mindestens zwei* Ladungen deutlich von Null verschieden sind.

Der Fehlerterm $\boldsymbol{e} = (e_1, \ldots, e_p)^\top$ wird *Einzelrestkomponente* (engl. *unique factor*) genannt.

Bei diesem Modell wird folgendes vorausgesetzt:

$$\begin{aligned} E(\boldsymbol{f}) = \boldsymbol{0},\ Cov(e_i, e_j) &= 0 \qquad (i \neq j), \\ E(\boldsymbol{e}) = \boldsymbol{0},\quad Cov(\boldsymbol{f}, \boldsymbol{e}) &= \boldsymbol{O}\ , \\ Var(f_i) &= 1\ . \end{aligned}$$

Die Kovarianzmatrix von $\boldsymbol{e}$ lässt sich somit in der Form

$$Cov(\boldsymbol{e}) = \boldsymbol{\Psi} = Diag(\psi_{11}, \ldots, \psi_{pp}) \tag{6.2}$$

schreiben.

Für die Korrelationsmatrix $\boldsymbol{\rho} = [(\rho_{ij})]_{i,j=1,\ldots,p}$ von $\boldsymbol{x}$ gilt nun:

$$\begin{aligned} \boldsymbol{\rho} &= Corr(\boldsymbol{x}) = Cov(\boldsymbol{y}) = Cov(\boldsymbol{\Lambda f} + \boldsymbol{e}) \\ &= \boldsymbol{\Lambda} \underbrace{Cov(\boldsymbol{f})}_{\boldsymbol{\Phi}} \boldsymbol{\Lambda}^\top + \boldsymbol{\Lambda} \underbrace{Cov(\boldsymbol{f}, \boldsymbol{e})}_{\boldsymbol{O}} + \underbrace{Cov(\boldsymbol{e}, \boldsymbol{f})}_{\boldsymbol{O}} \boldsymbol{\Lambda}^\top + \underbrace{Cov(\boldsymbol{e})}_{\boldsymbol{\Psi}} \\ &= \boldsymbol{\Lambda \Phi \Lambda}^\top + \boldsymbol{\Psi} \quad . \end{aligned}$$

$\boldsymbol{\Phi}$ ist die $(k \times k)$-Korrelationsmatrix der Faktoren.

Fordert man zusätzlich, dass die Faktoren unkorreliert sind, also

$$Cov(\boldsymbol{f}) = \boldsymbol{\Phi} = \boldsymbol{I} \quad , \tag{6.3}$$

so erhält man

$$\boldsymbol{\rho} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} \tag{6.4}$$

bzw.

$$\boldsymbol{\rho}_{red} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = \begin{pmatrix} \kappa_1^2 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \kappa_2^2 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \kappa_p^2 \end{pmatrix} = \boldsymbol{\rho} - \boldsymbol{\Psi} \quad . \tag{6.5}$$

$\boldsymbol{\rho}_{red}$ bezeichnet die *reduzierte Korrelationsmatrix*. Die Diagonalelemente $\kappa_i^2 = 1 - \psi_{ii} = \sum_{j=1}^{k} \lambda_{ij}^2 \quad (i = 1, \ldots p)$ werden *Kommunalitäten* genannt. Sie entsprechen den Zeilensummen der quadrierten Faktorenladungen und beschreiben den Anteil der Varianz von y_i, der durch die gemeinsamen Faktoren erklärt wird.

6.2.2 Nichteindeutigkeit der Faktorenladungen

Sei $\boldsymbol{G}$ eine orthogonale $(k \times k)$-Matrix. Dann gilt wegen $\boldsymbol{G}^{-1} = \boldsymbol{G}^\top$

$$\boldsymbol{y} = (\boldsymbol{\Lambda}\boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{f}) + \boldsymbol{e} \; . \tag{6.6}$$

Nachdem die neuen Faktoren $\boldsymbol{G}^\top \boldsymbol{f}$ die Modellannahmen erfüllen,

$$\begin{aligned} E(\boldsymbol{G}^\top \boldsymbol{f}) = \boldsymbol{0}, \quad & Cov(\boldsymbol{G}^\top \boldsymbol{f}) = \boldsymbol{I} \\ \text{und} \qquad & Cov(\boldsymbol{G}^\top \boldsymbol{f}, \boldsymbol{e}) = \boldsymbol{O} \; , \end{aligned}$$

ist das k-Faktorenmodell auch gültig mit den neuen Faktoren, d.h.

$$\boldsymbol{\rho} = (\boldsymbol{\Lambda}\boldsymbol{G})(\boldsymbol{G}^\top \boldsymbol{\Lambda}^\top) + \boldsymbol{\Psi} \; . \tag{6.7}$$

Das bedeutet aber, dass die Faktorenladungsmatrix nicht eindeutig bestimmt ist. Um eine eindeutige Lösung für die Faktorenladungsmatrix zu erhalten, fordern wir zusätzlich, dass entweder $\boldsymbol{\Lambda}^\top \boldsymbol{\Psi}^{-1} \boldsymbol{\Lambda}$ oder $\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda}$ eine Diagonalmatrix sein soll.

6.2.3 Parameterschätzung

In der Praxis ist eine Datenmatrix aus beobachteten Werten gegeben. Daraus kann die Korrelationsmatrix $\widehat{\boldsymbol{\rho}}$ geschätzt werden. Anhand von $\widehat{\boldsymbol{\rho}}$ sollen nun Schätzungen $\widehat{\boldsymbol{\Lambda}}$ und $\widehat{\boldsymbol{\Psi}}$ für $\boldsymbol{\Lambda}$ und $\boldsymbol{\Psi}$ gefunden werden, die sowohl

$$\widehat{\boldsymbol{\Lambda}}^\top \widehat{\boldsymbol{\Psi}}^{-1} \widehat{\boldsymbol{\Lambda}} = Diag \qquad \text{oder} \qquad \widehat{\boldsymbol{\Lambda}}^\top \widehat{\boldsymbol{\Lambda}} = Diag \tag{6.8}$$

als auch

$$\widehat{\boldsymbol{\rho}} \cong \widehat{\boldsymbol{\Lambda}} \widehat{\boldsymbol{\Lambda}}^\top + \widehat{\boldsymbol{\Psi}} \tag{6.9}$$

erfüllen.

Es stellt sich die Frage, ob das Faktorenmodell eine einfachere Interpretation der Daten liefert als die Korrelationsmatrix. Dazu kann die Anzahl der Parameter im uneingeschränkten Modell mit der Anzahl der freien Parameter im k-Faktorenmodell

verglichen werden. $\mathbf{\Lambda}$ und $\mathbf{\Psi}$ bestimmen gemeinsam $pk + p$ freie Parameter. Die Forderung, dass $\mathbf{\Lambda}^\top \mathbf{\Psi}^{-1} \mathbf{\Lambda}$ oder $\mathbf{\Lambda}^\top \mathbf{\Lambda}$ eine Diagonalmatrix ist, führt zu $\frac{1}{2}k(k-1)$ Einschränkungen. Es bleiben daher $pk + p - \frac{1}{2}k(k-1)$ Parameter. Die Anzahl der freien Parameter im uneingeschränkten Modell beträgt $\frac{1}{2}p(p+1)$. Die Differenz s ist somit

$$\begin{aligned} s &= \frac{1}{2}p(p+1) - \left(pk + p - \frac{1}{2}k(k-1)\right) \\ &= \frac{1}{2}(p-k)^2 - \frac{1}{2}(p+k) \quad . \end{aligned}$$

Je nach dem Wert von s können drei Fälle unterschieden werden:

1. $s < 0$: Es gibt mehr Parameter als Gleichungen. Daher können unendlich viele Lösungen für $\widehat{\mathbf{\Lambda}}$ und $\widehat{\mathbf{\Psi}}$ gefunden werden. Das Faktorenmodell ist somit nicht wohldefiniert.

2. $s = 0$: Es gibt gleich viele Parameter wie Gleichungen. $\widehat{\mathbf{\Lambda}}$ und $\widehat{\mathbf{\Psi}}$ können zwar eindeutig bestimmt werden, jedoch stellt das Faktorenmodell keine einfachere Beschreibung der Daten dar.

3. $s > 0$: Es gibt mehr Gleichungen als Parameter. $\widehat{\mathbf{\Lambda}}$ und $\widehat{\mathbf{\Psi}}$ können nicht exakt bestimmt werden, jedoch liefert das Faktorenmodell eine einfachere Interpretation der Daten als die Korrelationsmatrix.

Im folgenden wird nur der Fall $s > 0$ betrachtet.

6.3 Vorgangsweise

Der wesentliche Schritt in der Faktorenanalyse besteht in der Schätzung der Ladungsmatrix $\widehat{\mathbf{\Lambda}}$. Dazu gibt es verschiedene Verfahren, wie z.B. die Hauptkomponenten- bzw. Hauptfaktorenanalyse , die Maximum-Likelihood-Methode, die Zentroidmethode usw.

In diesem Abschnitt werden nur die Hauptfaktorenanalyse und die Maximum-Likelihood-Methode behandelt. Die Hauptkomponentenanalyse entspricht der Ermittlung von Hauptkomponenten der *nicht* reduzierten Korrelationsmatrix. Mit der Zentroidmethode werden die Achsen durch sogenannte *Schwerpunkte* gelegt.

6.3.1 Hauptfaktorenanalyse

Ausgehend von dem Modell

$$\boldsymbol{\rho} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi} \tag{6.10}$$

werden zuerst die Kommunalitäten geschätzt. Anschließend schätzt man die Ladungsmatrix.

Kommunalitätenschätzung

Die Kommunalitäten

$$\kappa_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 = 1 - \psi_{ii} \qquad (i = 1, \ldots, p) \tag{6.11}$$

beschreiben den Anteil der Varianz, der durch das gemeinsame Faktorenmodell erklärt wird. Sie können nur Werte zwischen 0 und 1 annehmen. Es gibt mehrere Möglichkeiten, die Kommunalitäten zu schätzen:

1. Der höchste Korrelationskoeffizient : $max_{i \neq j} \mid \rho_{ij} \mid$
   Für die Schätzung der Kommunalitäten wird in jeder Spalte von $\boldsymbol{\rho}$ das betragsmäßig größte Nichtdiagonalelement herangezogen. Diese Methode ist nur bei größerer Merkmalsanzahl empfehlenswert, da das Ergebnis zufallsabhängig ist.

2. Das Quadrat des multiplen Korrelationskoeffizienten: $\rho^2_{i,12\ldots)i(\ldots p}$
   Aus der Regressionsrechnung ist es als SMC (engl. *squared multiple correlation*) bekannt und beschreibt den Varianzanteil des i-ten Merkmals erklärt durch die anderen. Das Quadrat des multiplen Korrelationskoeffizienten des i-ten Merkmals kann folgendermaßen berechnet werden:

   $$\rho^2_{i,12\ldots)i(\ldots p} = 1 - \frac{1}{\rho^{ii}} \quad . \tag{6.12}$$

   Dabei ist ρ^{ii} das i-te Diagonalelement von $\boldsymbol{\rho}^{-1}$.

3. Die Iteration
   Bei der Iteration muss zuerst eine Entscheidung über die Anzahl k der zu extrahierenden Faktoren getroffen werden. Dazu kann man nach einem in der Hauptkomponentenanalyse besprochenen Kriterium vorgehen.

   Vorerst werden Schätzungen der Kommunalitäten (SMC oder Werte zwischen 0 und 1) in die Diagonale von $\boldsymbol{\rho}$ eingesetzt. Aus dieser reduzierten Korrelationsmatrix werden k Faktoren extrahiert (siehe Ende dieses Abschnitts). Aus dem Faktorenmuster (Ladungsmatrix) werden die neuen Kommunalitäten

   $$\kappa^2_{i_{neu}} = \sum_{j=1}^{k} \lambda_{ij}^2 \qquad \text{für} \quad i = 1, \ldots, p \tag{6.13}$$

   berechnet und als neue Schätzungen in die Diagonale von $\boldsymbol{\rho}$ eingesetzt. Die Iteration wird abgebrochen, wenn sich die Kommunalitäten nur mehr geringfügig ändern.

   Die Konvergenz dieses Verfahrens ist theoretisch nicht nachgewiesen. Außerdem wurde nicht gezeigt, unter welchen Bedingungen dieses Verfahren konvergiert und ob der Grenzwert mit den tatsächlichen Kommunalitäten übereinstimmt. Dennoch ist diese Methode am empfehlenswertesten.

Wählt man die Kommunalitäten zu hoch, so wird ein Teil der Einzelvarianz in die gemeinsame Varianz hineingepresst und das Faktorenmuster dadurch verändert. Wählt man sie zu niedrig, dann geht gemeinsame Varianz für die Bestimmung der Faktoren verloren.

Bei einer größeren Anzahl von Merkmalen ist der Einfluss ungenauer Kommunalitätenschätzungen auf die Lösung relativ gering. Bei kleinen Variablenanzahlen kann jedoch die Kommunalitätenschätzung entscheidend sein.

Schätzung der Ladungsmatrix

Nachdem nun die Kommunalitäten bekannt sind, kann man für die Schätzungen der Ladungen von der reduzierten Korrelationsmatrix

$$\boldsymbol{\rho}_{red} = \boldsymbol{\rho} - \boldsymbol{\Psi} \tag{6.14}$$

ausgehen, bei der in der Diagonale anstelle der Einsen die Kommunalitäten stehen.

Nach dem Spektralzerlegungssatz gilt

$$\boldsymbol{\rho} - \boldsymbol{\Psi} = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^{\top} = \sum_{i=1}^{p} a_i \boldsymbol{\gamma}_{.i}\boldsymbol{\gamma}_{.i}^{\top} \quad , \tag{6.15}$$

wobei $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_{.1}, \ldots, \boldsymbol{\gamma}_{.p})$ sowie $\boldsymbol{A} = Diag(a_1, \ldots, a_p)$ gilt und $\boldsymbol{\gamma}_{.i}$ bzw. a_i die Eigenvektoren bzw. die Eigenwerte von $\boldsymbol{\rho} - \boldsymbol{\Psi}$ sind $(i = 1, \ldots, p)$.

Das Faktorenmodell ist somit

$$\boldsymbol{\rho} - \boldsymbol{\Psi} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\top} = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Gamma}^{\top} \quad . \tag{6.16}$$

Für die Schätzung der Faktorenladungsmatrix gilt deshalb

$$\widehat{\boldsymbol{\Lambda}} = \boldsymbol{\Gamma}\boldsymbol{A}^{1/2} \quad . \tag{6.17}$$

Werden k Faktoren extrahiert $(1 \leq k \leq p)$, so kann diese Gleichung geschrieben werden als

$$\widehat{\boldsymbol{\lambda}}_{.j} = a_j^{1/2}\boldsymbol{\gamma}_{.j} \qquad \text{für} \quad j = 1, \ldots, k. \tag{6.18}$$

Für die Schätzung der merkmalseigenen Varianzen gilt

$$\hat{\psi}_{ii} = 1 - \sum_{j=1}^{k} \hat{\lambda}_{ij}^2 \qquad \text{für} \quad i = 1, \ldots, p \quad . \tag{6.19}$$

Die Lösung ist zulässig, wenn alle $\hat{\psi}_{ii} \geq 0$ sind.

Beispiel 6.3.1 *Gegeben seien die Prüfungsdaten aus Tabelle 2.1. Die empirische Korrelationsmatrix ergibt sich als*

$$\mathbf{R} = \begin{pmatrix} 1 & 0.553 & 0.547 & 0.410 & 0.389 \\ & 1 & 0.610 & 0.485 & 0.437 \\ & & 1 & 0.711 & 0.665 \\ & & & 1 & 0.607 \\ & & & & 1 \end{pmatrix} \quad .$$

Wählt man die Anzahl der Faktoren k größer als 2, so ist das Faktorenmodell nicht wohldefiniert ($k = 3 \Rightarrow s = -2 < 0$). Das bedeutet, dass entweder ein oder zwei Faktoren bestimmt werden können.

Zum Vergleich wurden die Kommunalitäten sowohl mit dem Quadrat des multiplen Korrelationskoeffizienten (SMC) *als auch mit dem* höchsten Korrelationskoeffizienten $max_{i \neq j} \mid r_{ij} \mid$ *geschätzt.*

Für die Inverse der Korrelationsmatrix ergibt sich

$$\mathbf{R}^{-1} = \begin{pmatrix} 1.603 & -0.558 & -0.510 & 0.001 & -0.041 \\ & 1.802 & -0.0659 & -0.152 & -0.039 \\ & & 3.047 & -1.113 & -0.864 \\ & & & 2.178 & -0.515 \\ & & & & 1.921 \end{pmatrix} .$$

Aus diesem Ergebnis wird nun der SMC-Schätzer berechnet. Somit ergibt sich für die Schätzungen der Kommunalitäten:

Schätzer	$\hat{\kappa}_1^2$	$\hat{\kappa}_2^2$	$\hat{\kappa}_3^2$	$\hat{\kappa}_4^2$	$\hat{\kappa}_5^2$
SMC	*0.376*	*0.445*	*0.672*	*0.541*	*0.479*
$max_{i \neq j} \mid r_{ij} \mid$	*0.553*	*0.610*	*0.711*	*0.711*	*0.665*

In diesem Beispiel werden die Kommunalitäten mit dem höchsten Korrelationskoeffizienten geschätzt. Aus der reduzierten Korrelationsmatrix

$$\mathbf{R} - \mathbf{\Psi} = \begin{pmatrix} 0.553 & 0.553 & 0.547 & 0.410 & 0.389 \\ & 0.610 & 0.610 & 0.485 & 0.437 \\ & & 0.711 & 0.711 & 0.665 \\ & & & 0.711 & 0.607 \\ & & & & 0.665 \end{pmatrix}$$

werden die Eigenwerte berechnet:

$$a_1 = 2.84 \qquad a_2 = 0.38 \qquad a_3 = 0.08 \qquad a_4 = 0.02 \qquad a_5 = -0.05$$

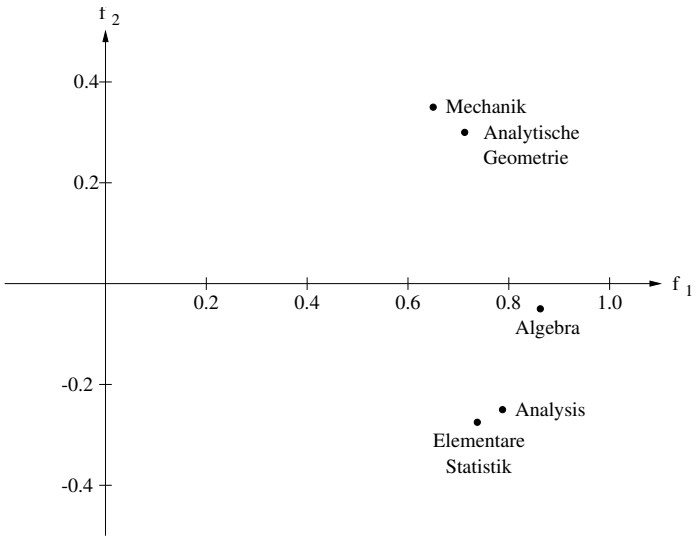
Aus der Größe der Eigenwerte kann man erkennen, dass die Annahme eines 1- bzw. 2-Faktorenmodells sinnvoll war. Für die Faktorenschätzungen und Kommunalitäten des 1- bzw. 2-Faktorenmodells ergibt sich:

	$\overbrace{\qquad}^{k=1}$		$\overbrace{\qquad\qquad\qquad}^{k=2}$		
Merkmale	κ_i^2	$\boldsymbol{\lambda}_1$	κ_i^2	$\boldsymbol{\lambda}_1$	$\boldsymbol{\lambda}_2$
ME	*0.417*	*0.646*	*0.543*	*0.646*	*0.354*
AG	*0.506*	*0.711*	*0.597*	*0.711*	*0.303*
LA	*0.746*	*0.864*	*0.749*	*0.864*	*-0.051*
AN	*0.618*	*0.786*	*0.680*	*0.786*	*-0.249*
ES	*0.551*	*0.742*	*0.627*	*0.742*	*-0.276*

Beim 1-Faktorenmodell kann man erkennen, dass die Ladungen in etwa gleich groß sind. In der „Ladungsmatrix“ ist daher keine Struktur ersichtlich. Anders ist das

beim 2-Faktorenmodell. Während sich die erste Spalte der Ladungsmatrix nicht ändert, zeigt die zweite Spalte positive Ladung bei den Merkmalen 1 und 2, keine Ladung bei Merkmal 3 und negative Ladung bei den Merkmalen 4 und 5. Der erste Faktor repräsentiert somit die durchschnittliche Leistung in den Prüfungsfächern, der zweite weniger wichtige Faktor ($a_2 = 0.38 << 2.84 = a_1$) beschreibt den Unterschied zwischen den Fächern (siehe auch Abbildung 6.1).

Abbildung 6.1: Faktorenladungen



Es muss darauf hingewiesen werden, dass alle Kommunalitäten wesentlich kleiner als 1 sind, d.h. es bleibt ein bedeutender Anteil der Varianz jeder Variablen unerklärt.

6.3.2 Maximum-Likelihood-Methode

Diese Methode wurde von Lawley (1940, 1941) eingeführt. Es wird angenommen, dass die betrachteten Variablen normalverteilt sind und dass die gemeinsamen Faktoren orthogonal sind. Wie auch in der Hauptfaktorenanalyse wird das Modell

$$\boldsymbol{\rho} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} \tag{6.20}$$

vorausgesetzt. Die Idee dieses Verfahrens ist, durch Maximierung der logarithmierten Plausibilitätsfunktion konsistente und effiziente Schätzungen der unbekannten Parameter $\boldsymbol{\Lambda}$ und $\boldsymbol{\Psi}$ zu erhalten.

Setzt man die SP-Korrelationsmatrix $\boldsymbol{R}$ als Realisierung für den Schätzer $\widehat{\boldsymbol{\rho}}$ ein, so erhält man die logarithmierte Plausibilitätsfunktion

$$l = -\frac{1}{2}n\ ln \mid \boldsymbol{\rho} \mid -\frac{1}{2}n\ tr\boldsymbol{\rho}^{-1}\boldsymbol{R} \quad . \tag{6.21}$$

l soll maximiert werden, d.h. die partiellen Ableitungen

$$\frac{\partial l}{\partial \lambda_{ij}} = -n\Big(\sum_t \lambda_{tj}\rho^{ti} - \sum_{t,u,v} \lambda_{tj}\rho^{tu}r_{uv}\rho^{vi}\Big) \tag{6.22}$$

und

$$\frac{\partial l}{\partial \psi_{ii}} = -\frac{1}{2}n\Big(\rho^{ii} - \sum_{t,u}\rho^{it}r_{tu}\rho^{ui}\Big) \tag{6.23}$$

werden Null gesetzt ($i = 1, \ldots, p$).

Die geschätzte Ladungsmatrix $\widehat{\boldsymbol{\Lambda}}$, die merkmalseigenen Varianzen $\widehat{\boldsymbol{\Psi}}$ und damit die geschätzte Korrelationsmatrix $\widehat{\boldsymbol{\rho}}$ sind durch die Gleichungen

$$\widehat{\boldsymbol{\Lambda}}^{\top}\widehat{\boldsymbol{\rho}}^{-1} - \widehat{\boldsymbol{\Lambda}}^{\top}\widehat{\boldsymbol{\rho}}^{-1}\boldsymbol{R}\,\widehat{\boldsymbol{\rho}}^{-1} = \boldsymbol{O} \tag{6.24}$$

und

$$diag(\widehat{\boldsymbol{\rho}}^{-1} - \widehat{\boldsymbol{\rho}}^{-1}\boldsymbol{R}\,\widehat{\boldsymbol{\rho}}^{-1}) = \boldsymbol{0} \tag{6.25}$$

gegeben, wobei mit $diag(\boldsymbol{X})$ die Diagonale der Matrix $\boldsymbol{X}$ bezeichnet wird. Die geschätzte Korrelationsmatrix ist dann

$$\widehat{\boldsymbol{\rho}} = \widehat{\boldsymbol{\Lambda}}\widehat{\boldsymbol{\Lambda}}^{\top} + \widehat{\boldsymbol{\Psi}} \quad . \tag{6.26}$$

Durch elementare Umformungen der Gleichungen (6.24) und (6.25) erhält man schließlich

$$diag(\widehat{\boldsymbol{\rho}} - \boldsymbol{R}) = \boldsymbol{0} \quad , \tag{6.27}$$

d.h. es gilt $\hat{\rho}_{ii} = r_{ii}$ für alle i oder

$$\widehat{\psi}_{ii} = r_{ii} - \sum_{j=1}^{k}\widehat{\lambda}_{ij}^2 = 1 - \sum_{j=1}^{k}\widehat{\lambda}_{ij}^2 \quad . \tag{6.28}$$

Durch Umformung von Gleichung (6.24) gilt außerdem

$$\widehat{\boldsymbol{\Lambda}}^{\top} = \widehat{\boldsymbol{J}}^{-1}\widehat{\boldsymbol{\Lambda}}^{\top}\widehat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{R} - \widehat{\boldsymbol{\Psi}}) \tag{6.29}$$

mit der Diagonalmatrix

$$\widehat{\boldsymbol{J}} = \widehat{\boldsymbol{\Lambda}}^{\top}\widehat{\boldsymbol{\Psi}}^{-1}\widehat{\boldsymbol{\Lambda}} \quad . \tag{6.30}$$

Daraus folgt, dass die Matrix

$$\widehat{\boldsymbol{H}} = \widehat{\boldsymbol{\Lambda}}^{\top}\widehat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{R} - \widehat{\boldsymbol{\Psi}})\widehat{\boldsymbol{\Psi}}^{-1}\widehat{\boldsymbol{\Lambda}} \tag{6.31}$$

gleich der Matrix $\widehat{\boldsymbol{J}}^2$ ist und daher diagonal ist.

Aus Gleichung (6.29) ist ersichtlich, dass die Elemente der Matrix $\widehat{\boldsymbol{J}}$ bzw. die Quadratwurzel der Elemente von $\widehat{\boldsymbol{H}}$ die Eigenwerte sowie die Zeilen der Matrix $\widehat{\boldsymbol{\Lambda}}^{\top}$ die Eigenvektoren von $\widehat{\boldsymbol{\Psi}}^{-1}(\boldsymbol{R} - \widehat{\boldsymbol{\Psi}})$ sind.

Eine **Lösung der Gleichungen** kann iterativ gewonnen werden: Seien $\widehat{\boldsymbol{\Lambda}}_{[1]}$ und $\widehat{\boldsymbol{\Psi}}_{[1]}$ Approximationen von $\widehat{\boldsymbol{\Lambda}}$ und $\widehat{\boldsymbol{\Psi}}$, die z.B. durch iterative Kommunalitätenschätzung und Extraktion der Faktoren erhalten werden (siehe voriger Abschnitt). Wenn $\widehat{\boldsymbol{\lambda}}_{j[1]}^{\top}$ die j-te Zeile von $\widehat{\boldsymbol{\Lambda}}_{[1]}^{\top}$ bezeichnet, so berechnet man $\widehat{\boldsymbol{\lambda}}_{j[2]}^{\top}$ durch

$$\widehat{\boldsymbol{\lambda}}_{j[2]}^{\top} = \frac{1}{\sqrt{\hat{h}_j}}\boldsymbol{u}_j^{\top} \tag{6.32}$$

mit

$$\boldsymbol{u}_j^\top = \widehat{\boldsymbol{\lambda}}_{j[1]}^\top \widehat{\boldsymbol{\Psi}}_{[1]}^{-1} \Big(\boldsymbol{R} - \widehat{\boldsymbol{\Psi}}_{[1]} - \sum_{t=1}^{j-1} \widehat{\boldsymbol{\lambda}}_{t[2]} \widehat{\boldsymbol{\lambda}}_{t[2]}^\top \Big) \tag{6.33}$$

und

$$\hat{h}_j = \boldsymbol{u}_j^\top \widehat{\boldsymbol{\Psi}}_{[1]}^{-1} \widehat{\boldsymbol{\lambda}}_{j[1]} \tag{6.34}$$

für $j = 1, \ldots, k$.

Eine bessere Approximation von $\hat{\psi}_{ii}$ kann durch Gleichung (6.28) erhalten werden. Diese neuen Schätzungen bilden die Matrix $\widehat{\boldsymbol{\Psi}}_{[2]}$. Mit den neuen Matrizen $\widehat{\boldsymbol{\Lambda}}_{[2]}$ und $\widehat{\boldsymbol{\Psi}}_{[2]}$ kann ein neuer Iterationszyklus gestartet werden. Der Prozess wird abgebrochen, wenn sich die Schätzung der Ladungsmatrix nur mehr geringfügig ändert.

Dieses Verfahren konvergiert oft sehr langsam, wobei die exakten Bedingungen für die Konvergenz nicht festgelegt sind.

Ein Nachteil der Maximum-Likelihood-Methode (auch der Hauptfaktorenanalyse) ist, dass die Anzahl k der zu extrahierenden Faktoren vorgegeben werden muss. Dieser Nachteil kann kompensiert werden, indem man einen *Test über die Anzahl der Faktoren* durchführt. Man kann mit einem Test zum Niveau α überprüfen, ob die gewählte Anzahl k der extrahierten gemeinsamen Faktoren zu gering war. Dazu wird die Hypothese

$$H_0 : \ \widehat{\boldsymbol{\rho}} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi} \quad , \tag{6.35}$$

wobei $\boldsymbol{\Lambda}$ eine $(p \times k)$-Matrix ist,
gegen die Hypothese

$$H_1 : \ \widehat{\boldsymbol{\rho}} = \boldsymbol{\Lambda}_1 \boldsymbol{\Lambda}_1^\top \tag{6.36}$$

getestet, wobei $\boldsymbol{\Lambda}_1$ eine $(p \times p)$-Matrix ist. Der Plausibilitätsquotiententest zum Testen der H_0 ist $\lambda = l_0/l_1$, wobei l_0 und l_1 die Plausibilitätsfunktionen unter der H_0 bzw. H_1 sind.

Die Statistik $-2 \ln \lambda$ ist asymptotisch χ^2-verteilt. Die Teststatistik

$$T = \Big[n - 1 - \frac{1}{6}(2p+5) - \frac{2}{3}k \Big] \ ln \frac{\mid \widehat{\boldsymbol{\rho}} \mid}{\mid \boldsymbol{R} \mid} \tag{6.37}$$

ist approximativ χ^2_ν-verteilt mit

$$\nu = \frac{1}{2}[(p-k)^2 - p - k] \tag{6.38}$$

Freiheitsgraden (siehe Lawley und Maxwell, 1971). Da die Anzahl der freien Parameter, $\frac{1}{2}p(p+1)$, größer sein sollte als die Anzahl der zu schätzenden Parameter, $pk + p - \frac{1}{2}k(k-1)$, erhält man die Bedingung

$$k \leq \frac{1}{2}\Big(1 + 2p - \sqrt{1+8p} \Big) \ . \tag{6.39}$$

Die Hypothese H_0, die besagt, dass k die richtige Anzahl von Faktoren ist, wird also zum Niveau α verworfen, falls

$$T > \chi^2_{\nu;1-\alpha} \tag{6.40}$$

gilt.

Die durch die Maximum-Likelihood-Methode (bzw. Hauptfaktorenanalyse) bestimmte Ladungsmatrix $\widehat{\mathbf{\Lambda}}$ ist i.a. nicht optimal bezüglich der Interpretierbarkeit der k extrahierten orthogonalen Faktoren. Es empfiehlt sich daher, auf $\widehat{\mathbf{\Lambda}}$ ein orthogonales bzw. schiefwinkeliges Rotationsverfahren anzuwenden (siehe nächster Abschnitt).

6.4 Faktorenrotation

Die Forderungen, dass entweder $\widehat{\mathbf{\Lambda}}^\top \widehat{\mathbf{\Psi}}^{-1} \widehat{\mathbf{\Lambda}}$ oder $\widehat{\mathbf{\Lambda}}^\top \widehat{\mathbf{\Lambda}}$ eine Diagonalmatrix sein soll, bewirken zwar die Eindeutigkeit der Lösung, jedoch sind die k Faktoren in den meisten Fällen nicht interpretierbar. Das Faktorenmuster ändert sich von Stichprobe zu Stichprobe und wird durch Einbeziehung neuer Merkmale entscheidend beeinflusst. Das Koordinatensystem der Faktoren soll jedoch die Merkmale möglichst einfach beschreiben. Das soll durch eine Faktorenrotation erreicht werden. Die Struktur, die man dadurch erlangen möchte, wird *Einfachstruktur* genannt.

Für die Einfachstruktur wurden folgende *fünf Postulate* festgelegt (siehe Thurstone, 1944):

1. Jede Zeile einer Ladungsmatrix soll mindestens eine Null enthalten, d.h. jedes Merkmal wird durch höchstens $k-1$ Faktoren beschrieben.

2. Jede Spalte einer Ladungsmatrix enthält wenigstens k Nulladungen, d.h. jeder Faktor trägt zur Beschreibung von höchstens $p-k$ der p Merkmale bei.

3. In jedem Spaltenpaar der Ladungsmatrix gibt es mehrere Merkmale, die auf dem einen Faktor eine hohe, auf dem anderen Faktor keine Ladung haben.

4. Wurden mehr als 4 Faktoren extrahiert, so soll jedes beliebige Spaltenpaar für eine große Zahl von Merkmalen in beiden Spalten Null enthalten.

5. Für jedes Spaltenpaar sollten nur wenige Merkmale in beiden Spalten hohe Ladungen haben.

Zum besseren Verständnis wird nun der Begriff der Einfachstruktur geometrisch veranschaulicht.

In den nachfolgenden Abbildungen stellen die Punkte die verschiedenen Merkmale dar.

In Abbildung 6.2 ist ein zweidimensionales Faktorenmuster dargestellt, wie es zustandekommt, wenn die Beziehungen zwischen den Variablen keine besondere Ordnung aufweisen. Diesen Fall nennt man *Zufallskonfiguration*. Hier gibt es keinen Hinweis darauf, wie man die Faktorenachsen legen soll.

In Abbildung 6.3 liegen die Merkmalspunkte deutlich in zwei Gruppen. Durch das rotierte Koordinatensystem $\widetilde{f}_1$, $\widetilde{f}_2$ erhält man eine einfachere Beschreibung der Merkmale als durch das ursprüngliche Koordinatensystem f_1, f_2. Projiziert man die Merkmalspunkte auf die Achsen, so kann man erkennen, dass die Ladungen auf einer Achse groß, auf der anderen Achse hingegen sehr klein sind. Da die rotierten Achsen

Abbildung 6.2: Zufallskonfiguration

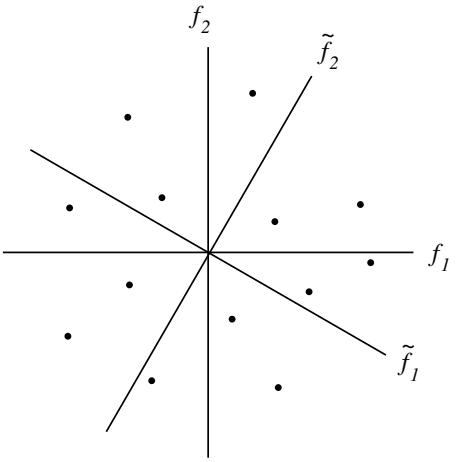
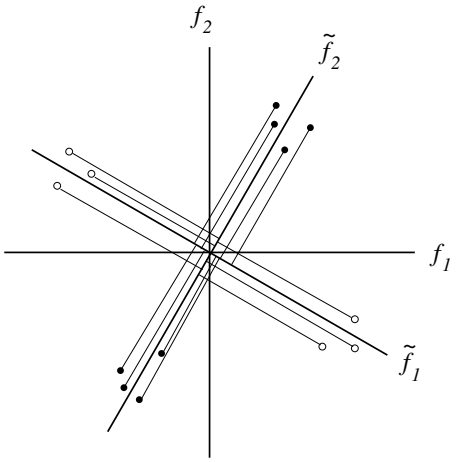


Abbildung 6.3: Orthogonale Einfachstruktur



ebenso wie das ursprüngliche Koordinatensystem aufeinander normal stehen, spricht man von einer *orthogonalen Einfachstruktur*.

Bei realitätsbezogenen Studien gibt es kaum eine so klare Struktur wie in Abbildung 6.3. Die beobachteten Merkmalspunkte sind im Faktorenraum meist in einer Form angeordnet, die ein Mittelding zwischen den beiden Extremen der Zufallskonfiguration und der orthogonalen Einfachstruktur sind. Die Punktwolken liegen meist auf Achsen, die nicht zueinander orthogonal sind. Dieser Fall ist in Abbildung 6.4 veranschaulicht. Man spricht von einer *schiefwinkeligen Einfachstruktur*.

Die Projektionen der Punkte auf die Achsen des rotierten Koordinatensystems unterscheiden sich längs einer Achse kaum, auch wenn sie von Punkten des anderen Punktschwarmes stammen. Man kann daher die Punktgruppen schlecht nach der Höhe ihrer Ladungen unterscheiden, wie dies bei der orthogonalen Einfachstruktur der Fall ist.

Abbildung 6.4: Schiefwinkelige Einfachstruktur

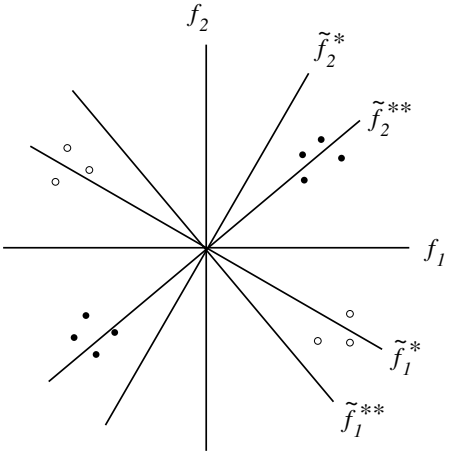
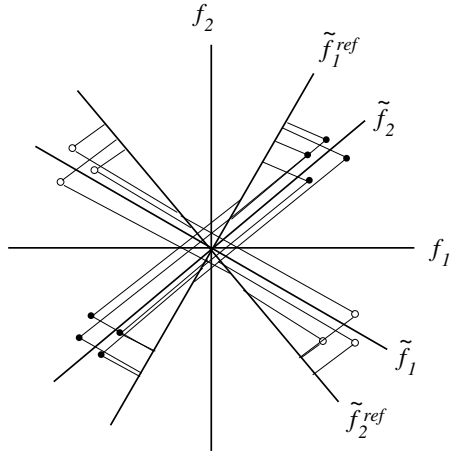


Abbildung 6.5: Referenzvektoren



Dieses Problem wurde mit der Einführung von *Referenzvektoren* $\widetilde{f}_1^{ref}$ und $\widetilde{f}_2^{ref}$ gelöst (siehe Thurstone, 1961). Durch die Schwerpunkte der Punktwolken werden Achsen gelegt. Die darauf normal stehenden Vektoren (Referenzvektoren) bilden das neue Koordinatensystem. Somit können die Punktgruppen wieder nach der Höhe ihrer Ladungen unterschieden werden (siehe Abbildung 6.5).

Die Faktorenrotation entspricht einer Transformation der Ladungsmatrix. Man unterscheidet zwischen *schiefwinkeliger* und *orthogonaler Rotation*.

Bei der orthogonalen Rotation bleibt die Orthogonalität, d.h. die Unkorreliertheit der Faktoren, erhalten. Meist werden jedoch die Forderungen nach Einfachstruktur durch eine schiefwinkelige Rotation, die auch korrelierte Faktoren zulässt, besser erfüllt.

6.4.1 Orthogonale Rotationsverfahren

Ausgegangen wird von der Idee, dass ein Punkt im zweidimensionalen Koordinatensystem dann am einfachsten dargestellt wird, wenn er auf einer der beiden Achsen liegt. Je näher ein Punkt einer Achse kommt, desto kleiner wird das Produkt aus den beiden Koordinaten des Punktes. Da diese Produkte aus den Faktorenladungen positiv oder negativ sein können und die Summe aller Produkte als Kriterium der Einfachheit in Frage kommt, werden die Produkte erst quadriert. Damit ergibt sich folgendes Kriterium der Einfachheit, das ein Minimum annehmen muss, wenn möglichst viele Punkte nahe an den Achsen liegen:

$$\sum_{s<j=1}^{k} \sum_{i=1}^{p} \Big(\lambda_{is}\lambda_{ij}\Big)^2 = Min \quad . \tag{6.41}$$

Die Summe über alle Variablen $i = 1, \ldots, p$ und über alle Paare von Faktoren über diesen Ausdruck muss also minimiert werden.

Wenn $\boldsymbol{T}$ als orthogonale $(k \times k)$-Transformationsmatrix und $\widetilde{\boldsymbol{\Lambda}}$ als orthogonal rotierte $(p \times k)$-Ladungsmatrix bezeichnet wird, so wird die ursprüngliche Ladungsmatrix $\boldsymbol{\Lambda}$ durch die Transformation

$$\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{T} \tag{6.42}$$

rotiert. Wenn die Transformation orthogonal ist, bleibt die Kommunalität invariant:

$$\kappa_i^2 = \sum_{j=1}^{k} \lambda_{ij}^2 = \sum_{j=1}^{k} \tilde{\lambda}_{ij}^2 \qquad \text{für} \quad i = 1, \ldots, p \quad . \tag{6.43}$$

Auch das Quadrat der Kommunalitäten bleibt durch eine orthogonale Rotation konstant:

$$(\kappa_i^2)^2 = \Big(\sum_{j=1}^{k} \tilde{\lambda}_{ij}^2\Big)^2 = \sum_{j=1}^{k} \tilde{\lambda}_{ij}^4 + 2 \sum_{s<j=1}^{k} \tilde{\lambda}_{is}^2 \tilde{\lambda}_{ij}^2 = konst \quad . \tag{6.44}$$

Auch wenn über alle Variablen p summiert wird, bleibt die erhaltene Größe für orthogonale $\boldsymbol{T}$ konstant:

$$\sum_{i=1}^{p} \sum_{j=1}^{k} \tilde{\lambda}_{ij}^4 + 2 \sum_{i=1}^{p} \sum_{s<j=1}^{k} \tilde{\lambda}_{is}^2 \tilde{\lambda}_{ij}^2 = konst \quad . \tag{6.45}$$

Da die Summe der beiden Ausdrücke konstant ist, wird der eine Summand maximiert, wenn der andere minimiert wird und umgekehrt.

Das bekannte **Quartimax-Kriterium** geht daher von der Maximierung der Funktion

$$QMAX = \sum_{i=1}^{p} \sum_{j=1}^{k} \tilde{\lambda}_{ij}^4 \tag{6.46}$$

aus. Nachteil dieser Methode ist, dass ein allgemeiner Faktor zu stark herausgehoben wird; allerdings wird manchmal bewusst ein sogenanntes Einfaktor-Modell gefordert.

Ein anderes und wahrscheinlich das bekannteste orthogonale Rotationskriterium ist das **Varimax-Kriterium**, das von Kaiser (1958) entwickelt wurde. Kaiser betrachtet die Varianz der quadrierten Faktorenladungen für den Faktor j, nämlich

$$s_j^2 = \frac{1}{p} \sum_{i=1}^{p} \left(\widetilde{\lambda}_{ij}^2\right)^2 - \frac{1}{p^2} \Big[\sum_{i=1}^{p} \widetilde{\lambda}_{ij}^2\Big]^2 \quad . \tag{6.47}$$

Diese Varianz wird über alle Faktoren summiert. Die so erhaltene Größe muss maximal werden, wenn im gesamten Faktorenmuster die Varianz der quadrierten Faktorenladungen möglichst groß sein soll. Nachteil dieser Größe ist, dass Variablen mit höherer Kommunalität stärker eingehen. Deshalb wird jede Faktorenladung durch die jeweilige Kommunalität dividiert. Das Varimax-Kriterium lautet dann (nach Multiplikation mit p^2)

$$VMAX = p \sum_{j=1}^{k} \sum_{i=1}^{p} \Big(\frac{\widetilde{\lambda}_{ij}}{\kappa_i}\Big)^4 - \sum_{j=1}^{k} \Big[\sum_{i=1}^{p} \Big(\frac{\widetilde{\lambda}_{ij}}{\kappa_i}\Big)^2\Big]^2 \quad . \tag{6.48}$$

6.4.2 Schiefwinkelige Rotationsverfahren

Wie bereits in der Einleitung zur Faktorenrotation erwähnt wurde, kann i.a. eine Einfachstruktur durch orthogonale Rotation nicht optimal erreicht werden. Die schiefwinkelige Rotation ist zwar rechenintensiver, aber in Hinblick auf die Interpretierbarkeit der Faktoren ist dieses Verfahren in den meisten Fällen besser geeignet.

Auch hier gibt es eine Reihe von verschiedenen Rotationskriterien, von denen nur zwei vorgestellt werden.

Der gravierende Unterschied zu den orthogonalen Verfahren liegt am Ausgangsmodell

$$\boldsymbol{\rho} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{\top} + \boldsymbol{\Psi} \quad . \tag{6.49}$$

Da die schiefwinkeligen Faktoren miteinander korrelieren, muss auch die Korrelationsmatrix $\boldsymbol{\Phi}$ der Faktoren miteingebracht werden, die bei den orthogonalen Verfahren gleich der Einheitsmatrix ist.

Das zu minimierende **Quartimin-Kriterium** $QMIN$ ist analog zu (6.41) definiert durch

$$QMIN = \sum_{s<j=1}^{k} \sum_{i=1}^{p} \widetilde{\lambda}_{is}^2 \widetilde{\lambda}_{ij}^2 \quad , \tag{6.50}$$

das eine Einfachstruktur der Faktorenladungen bewirken soll.

Die rotierte Ladungsmatrix ist

$$\widetilde{\boldsymbol{\Lambda}} = \boldsymbol{\Lambda}\boldsymbol{T} \ . \tag{6.51}$$

Durch Einsetzen in das Modell erhält man

$$\begin{aligned} \boldsymbol{\rho} - \boldsymbol{\Psi} &= \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{\top} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{T}^{-1}\boldsymbol{\Phi}(\boldsymbol{T}^{-1})^{\top}\widetilde{\boldsymbol{\Lambda}}^{\top} = \widetilde{\boldsymbol{\Lambda}}\boldsymbol{T}^{-1}\mathrm{Cov}(\boldsymbol{f})(\boldsymbol{T}^{-1})^{\top}\widetilde{\boldsymbol{\Lambda}}^{\top} \\ &= \widetilde{\boldsymbol{\Lambda}}\mathrm{Cov}(\boldsymbol{T}^{-1}\boldsymbol{f})\widetilde{\boldsymbol{\Lambda}}^{\top} = \widetilde{\boldsymbol{\Lambda}}\mathrm{Cov}(\widetilde{\boldsymbol{f}})\widetilde{\boldsymbol{\Lambda}}^{\top} \quad . \end{aligned} \tag{6.52}$$

Die rotierten Faktoren $\widetilde{\boldsymbol{f}}$ berechnen sich somit durch

$$\widetilde{\boldsymbol{f}} = \boldsymbol{T}^{-1}\boldsymbol{f} \ . \tag{6.53}$$

Die Transformationsmatrix $\boldsymbol{T}$ muss somit so konstruiert werden, dass ihre Inverse existiert.

Ein allgemeineres und vielleicht besseres Kriterium als das Quartimin-Kriterium ist das **Oblimin-Kriterium**

$$OBMIN = \sum_{s<j=1}^{k} \left(\sum_{i=1}^{p} \widetilde{\lambda}_{is}^2 \widetilde{\lambda}_{ij}^2 - \frac{\gamma}{p} \sum_{i=1}^{p} \widetilde{\lambda}_{is}^2 \sum_{i=1}^{p} \widetilde{\lambda}_{ij}^2 \right) \quad . \tag{6.54}$$

Durch die Wahl von $\gamma = 0$ erhält man das Quartimin-Kriterium, $\gamma = 1$ liefert das sogenannte *Covarimin-Kriterium*.

Beispiel 6.4.1 *Es werden wiederum die Prüfungsdaten betrachtet. Die durch die Maximum-Likelihood-Methode berechneten Faktoren werden sowohl nach dem Varimax-Kriterium (VMAX) als auch nach dem Quartimin-Kriterium (QMIN) rotiert. Man erhält folgende Ergebnisse:*

	$\overbrace{\quad VMAX \quad}$		$\overbrace{\quad QMIN \quad}$	
Merkmale	$\widetilde{\boldsymbol{\lambda}}_{.1}$	$\widetilde{\boldsymbol{\lambda}}_{.2}$	$\widetilde{\boldsymbol{\lambda}}_{.1}$	$\widetilde{\boldsymbol{\lambda}}_{.2}$
ME	*0.271*	*0.675*	*-0.033*	*0.752*
AG	*0.354*	*0.680*	*0.078*	*0.707*
LA	*0.734*	*0.513*	*0.694*	*0.250*
AN	*0.742*	*0.319*	*0.821*	*-0.019*
ES	*0.704*	*0.285*	*0.789*	*-0.041*

In Abbildung 6.6 kann man erkennen, dass in diesem Beispiel die orthogonale Rotation zu keiner Verbesserung der Interpretierbarkeit der Faktoren beiträgt. In Abbildung 6.7 hingegen ist erkennbar, dass Elementare Statistik, Analysis und Lineare Algebra durch den Faktor $\widetilde{f}_1$ erklärt werden. Mechanik und Analytische Geometrie werden durch den Faktor $\widetilde{f}_2$ erklärt. Die schiefwinkelige Rotation lässt somit eine Interpretation des Faktors $\widetilde{f}_1$ als algorithmisch abstraktes Denkvermögen *und des Faktors $\widetilde{f}_2$ als* räumlich abstraktes Denkvermögen *zu.*

Abbildung 6.6: Varimax-Rotation (orthogonal)

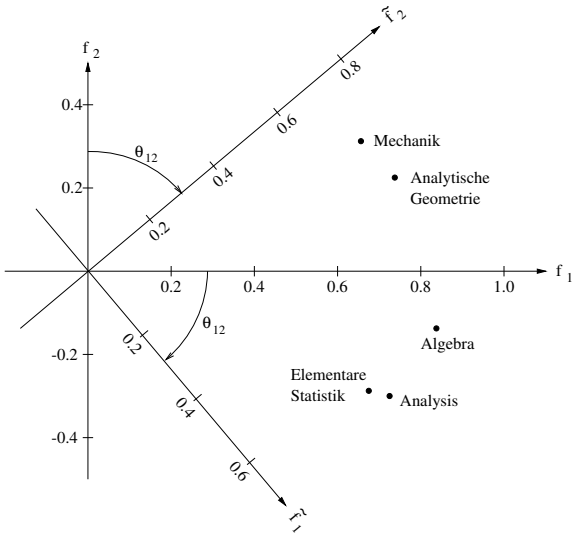
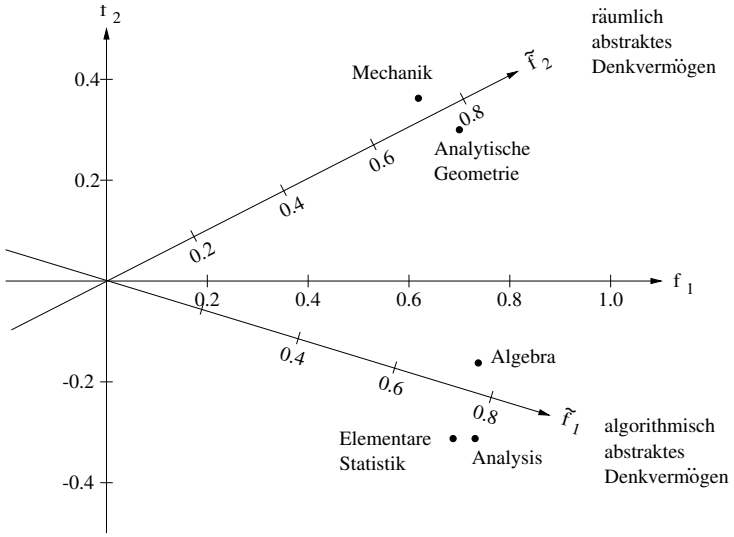


Abbildung 6.7: Quartimin-Rotation (schiefwinkelig)



6.5 Schätzung von Faktorenwerten

Das Hauptinteresse der Faktorenanalyse liegt in der Schätzung der Parameter im Faktorenmodell. Oft möchte man jedoch auch die Werte der gemeinsamen Faktoren (*common factor scores*) ermitteln. Diese Größen können dann für Diagnostik bzw. zur weiteren Analyse (z.B. Clusteranalyse) herangezogen werden.

Die Faktorenwerte sind nicht Schätzungen von unbekannten Parametern im üblichen Sinn. Sie sind Schätzungen von **unbeobachteten** Zufallsvektoren $f_{j\cdot}$ $(j = 1, \ldots, n)$. Es gibt verschiedene Möglichkeiten, die Faktorenwerte zu schätzen, zwei der bekanntesten werden nun vorgestellt.

6.5.1 Gewichtete Kleinste-Quadratsummen-Schätzung

Unser Faktorenmodell lautet

$$\boldsymbol{y} = \boldsymbol{\Lambda f} + \boldsymbol{e} \tag{6.55}$$

mit dem Fehlerterm $\boldsymbol{e} = (e_1, \ldots, e_p)^\top$. Dieses Modell kann auch als Regressionsmodell mit zu schätzenden Regressionsparametern $\boldsymbol{f}$ aufgefasst werden. Allerdings ist die Forderung nach Homoskedastizität nicht erfüllt, da $Var(e_i) = \psi_i$ für $i = 1, \ldots, p$ nicht notwendigerweise gleich ist (Heteroskedastizität). Man kann das Problem aber leicht in den Griff bekommen, indem man die Gleichung mit $\boldsymbol{\Psi}^{-1/2}$ multipliziert (gewichtet),

$$\boldsymbol{\Psi}^{-1/2}\boldsymbol{y} = \boldsymbol{\Psi}^{-1/2}\boldsymbol{\Lambda f} + \boldsymbol{\Psi}^{-1/2}\boldsymbol{e} \ . \tag{6.56}$$

Für den neuen Fehlerterm $\boldsymbol{\Psi}^{-1/2}\boldsymbol{e}$ ist die Gleichheit der Varianzen wieder gegeben, und der Kleinste-Quadratsummenschätzer (*Least Squares*) liefert nun eine unverzerrte Schätzung. Wichtig ist dabei, dass die vorher geschätzten Ladungen und merkmalseigenen Varianzen jetzt als **wahre Werte** behandelt werden. Es ergibt sich

$$\hat{\boldsymbol{f}} = \left(\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{y} \tag{6.57}$$

für die Schätzung der Faktorenwerte (Zufallsvektor). Für eine konkrete Stichprobe sind somit die geschätzten Faktorenwerte der j-ten Beobachtung

$$\hat{\boldsymbol{f}}_{j.} = \left(\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{y}_{j.} \ . \tag{6.58}$$

Die $(n \times k)$-Matrix der geschätzten Faktorenwerte ist daher

$$\hat{\boldsymbol{F}} = \boldsymbol{Y}\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1} \ . \tag{6.59}$$

6.5.2 Regressionsmethode

Wie auch bei der vorigen Methode werden sowohl die Ladungsmatrix $\boldsymbol{\Lambda}$ als auch die Matrix mit den merkmalseigenen Varianzen $\boldsymbol{\Psi}$ als bekannt vorausgesetzt. Ausgehend vom Faktorenmodell (6.55) können nun die Faktorenwerte als Regressionsproblem bezüglich der Variablen $\boldsymbol{y}$ betrachtet werden,

$$\boldsymbol{f} = \boldsymbol{B}\boldsymbol{y} + \boldsymbol{\delta} \ , \tag{6.60}$$

wobei $\boldsymbol{B}$ eine $(p \times k)$-Matrix der Regressionskoeffizienten bezeichnet. Die Regressionskoeffizienten sollen im Sinne der kleinsten Quadrate geschätzt werden. Minimierung von $\boldsymbol{\delta}^\top\boldsymbol{\delta}$ ergibt den Regressionsschätzer

$$\widehat{\boldsymbol{B}} = \boldsymbol{f}\boldsymbol{y}^\top(\boldsymbol{y}\boldsymbol{y}^\top)^{-1} \ . \tag{6.61}$$

Da $\boldsymbol{f}$ unbekannt ist, erhält man eine Schätzung durch

$$\begin{aligned} \widehat{\boldsymbol{f}} &= \widehat{\boldsymbol{B}}\boldsymbol{y} \\ &= \boldsymbol{f}\boldsymbol{y}^\top(\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\boldsymbol{y} \\ &= \boldsymbol{\Phi}\boldsymbol{\Lambda}^\top\boldsymbol{\rho}^{-1}\boldsymbol{y} \ . \end{aligned} \tag{6.62}$$

$\boldsymbol{\rho}$ ist die Korrelationsmatrix der Variablen. $\boldsymbol{\Lambda}$ bezeichnet hier sinnvollerweise die rotierte Ladungsmatrix, und daher erhält man mit $\widehat{\boldsymbol{f}}$ auch schon die Faktorenwerte bezüglich des rotierten Koordinatensystems. Für eine konkrete Stichprobe erhält man daher als Schätzung der Faktorenwerte für die j-te Beobachtung ($j = 1, \ldots, n$)

$$\widehat{\boldsymbol{f}}_{j.} = \boldsymbol{\Phi}\boldsymbol{\Lambda}^\top \boldsymbol{R}^{-1}\boldsymbol{y}_{j.} \ , \tag{6.63}$$

die gesamte Matrix ist

$$\widehat{\boldsymbol{F}} = \boldsymbol{Y}\boldsymbol{R}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi} \ . \tag{6.64}$$

Im Falle von orthogonaler Rotation gilt $\boldsymbol{\Phi} = \boldsymbol{I}$, womit sich die Formel reduziert auf

$$\widehat{\boldsymbol{F}} = \boldsymbol{Y}\boldsymbol{R}^{-1}\boldsymbol{\Lambda} \ . \qquad \text{(orthogonale Faktoren)} \ . \tag{6.65}$$

Es kann auch die Beziehung der Faktorenwerte bezüglich der beiden Schätzmethoden dargestellt werden. Zu diesem Zweck führen wir die Bezeichnung $\hat{\boldsymbol{f}}^{LS}$ für Kleinste-Quadratsummen-Schätzung und $\hat{\boldsymbol{f}}^{R}$ für Regressionsschätzung ein. Man kann leicht zeigen, dass die Identität

$$\boldsymbol{\Lambda}^\top \left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}\right)^{-1} = \left(\boldsymbol{I} + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1} \tag{6.66}$$

immer erfüllt ist. $\boldsymbol{\Lambda}$ soll hier gleich die (schiefwinkelig) rotierte Ladungsmatrix bezeichnen. Es folgt daraus

$$\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{y} = \left(\boldsymbol{I} + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\right)\left[\boldsymbol{\Lambda}^\top\left(\boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}\right)^{-1}\boldsymbol{y}\right] \ . \tag{6.67}$$

Der Ausdruck in der eckigen Klammer ist aber genau $\boldsymbol{\Phi}^{-1}\hat{\boldsymbol{f}}^{R}$, und daher ist

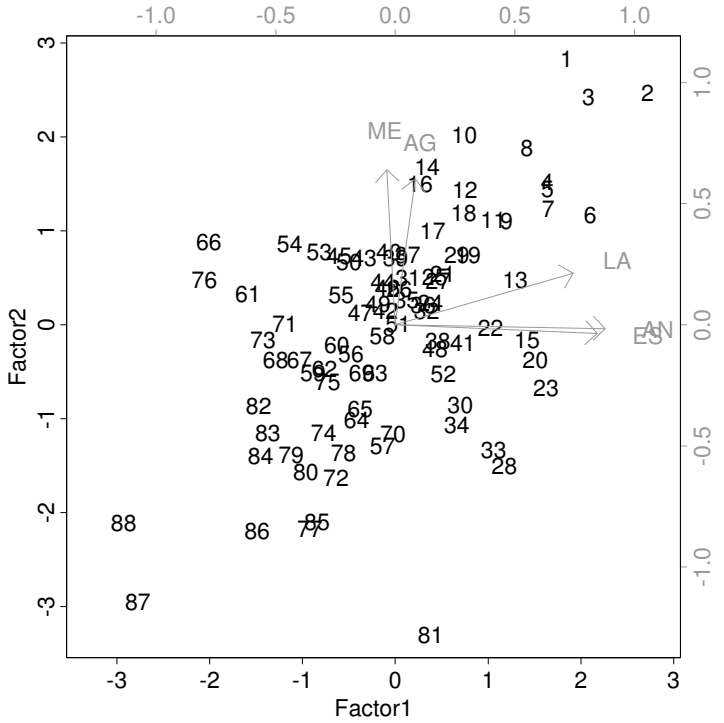
$$\begin{aligned}\hat{\boldsymbol{f}}^{LS} &= \left(\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1}\left(\boldsymbol{I} + \boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi}\right)\boldsymbol{\Phi}^{-1}\hat{\boldsymbol{f}}^{R} \\ &= \left[\left(\boldsymbol{\Phi}\boldsymbol{\Lambda}^\top\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\right)^{-1} + \boldsymbol{I}\right]\hat{\boldsymbol{f}}^{R}\end{aligned} \tag{6.68}$$

Wenn die Elemente der Matrix in runden Klammern nahe 0 sind, werden die beiden Schätzmethoden annähernd das gleiche Ergebnis liefern.

Beispiel 6.5.1 *Zur Veranschaulichung der Schätzung von Faktorenwerten wird wieder das Beispiel mit den Prüfungsdaten herangezogen.*

Nach erfolgter schiefwinkeliger Quartimin-Rotation (siehe letztes Beispiel) können die Faktorenwerte z.B. nach der Gewichteten Kleinste-Quadratsummen-Schätzung berechnet werden. Das Ergebnis ist in Abbildung 6.8 in Form eines Biplots graphisch veranschaulicht. Hohe Werte auf Faktor 1 bedeuten algorithmisch abstraktes Denkvermögen, hohe Werte auf Faktor 2 bedeuten räumlich abstraktes Denkvermögen.

Abbildung 6.8: Schätzung der Faktorenwerte



Literatur

K. Adel, R. Dutter, H. Filzmoser, and P. Filzmoser. *Tiefenstrukturen der Sprache: Untersuchung regionaler Unterschiede mit statistischen Methoden*. WUV-Universitätsverlag, Wien, 1994.

D.B. Clarkson and R.I. Jennrich. Quartic Rotation Criteria and Algorithms. *Psychometrika*, 53(2):251–259, 1988.

W.R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications*. John Wiley & Sons, New York, 1984.

L. Guttman. Multiple rectilinear prediction and the resolution into components I. *Psychometrika*, 5:75–99, 1940. Cynthia O, Williamsburg, Virginia.

H.H. Harman. *Modern Factor Analysis*. The University of Chicago Press, Chicago and London, 2nd edition, 1967.

R.I. Jennrich and P.F. Sampson. Rotation for Simple Loadings. *Psychometrika*, 31(3):313–323, 1966.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, London, 4th edition, 1998.

H.F. Kaiser. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200, 1958.

D.N. Lawley and A.E. Maxwell. *Factor Analysis as a Statistical Method*. Butterworths, London, 1963.

D. Revenstorf. *Lehrbuch der Faktorenanalyse*. Verlag W. Kohlhammer, Stuttgart, 1976.

G.A.F. Seber. *Multivariate observations*. John Wiley & Sons, New York, 1984.

L.L. Thurstone. Second-order factors. *Psychometrika*, 9:71–100, 1944. Cynthia O, Williamsburg, Virginia.

L.L. Thurstone. *Multiple factor analysis*. University Press Chicago, Chicago, sixth edition, 1961.

K. Überla. *Faktorenanalyse*. Springer, Berlin, 1971.

Kapitel 7

Korrelationsanalyse

Bei statistischen Analysen von Daten ist man oft daran interessiert, Zusammenhänge und Abhängigkeiten von Merkmalen festzustellen. Ein Instrumentarium dafür ist die Regressionsanalyse. Möchte man auch das Vorhandensein und die Stärke von Abhängigkeiten messen, so kann die *Korrelationsanalyse* dafür herangezogen werden. Die Korrelation misst den linearen Zusammenhang zwischen Merkmalen.

7.1 Multiple Korrelation

Die *multiple Korrelation* ist ein Maß für die Abhängigkeit eines Merkmals x von einem p-dimensionalen Merkmal $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$. Wir nehmen an, dass sowohl x als auch $y_1, \ldots, y_p$ Zufallsvariablen mit gemeinsamer Verteilung sind. Mittelvektor $\boldsymbol{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$ dieser Verteilung (nicht notwendigerweise Normalverteilung) sind somit

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \boldsymbol{\mu_y} \end{pmatrix} \quad \text{bzw.} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{xx} & \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \\ \boldsymbol{\sigma}_{\boldsymbol{y}x} & \boldsymbol{\Sigma_{yy}} \end{pmatrix} .$$

Wird nun x durch $\boldsymbol{y}$ (linear) vorhergesagt, ergibt sich analog zur Regression der Fehler

$$x - a_0 - a_1 y_1 - \ldots - a_p y_p .$$

Dieser Fehler ist zufallsabhängig, und daher möchte man die Koeffizienten a_0 und $\boldsymbol{a} = (a_1, \ldots, a_p)^\top$ so wählen, dass der mittlere quadratische Fehler (*mean squared error*)

$$\text{MSE} = E(x - a_0 - \boldsymbol{a}^\top \boldsymbol{y})^2$$

minimal wird. Dieser MSE hängt aber von der gemeinsamen Verteilung von x und $\boldsymbol{y}$ ab, und somit von den Parametern $\boldsymbol{\mu}$ und $\boldsymbol{\Sigma}$.

Satz 7.1.1 *Die lineare Vorhersagefunktion $a_0 + \boldsymbol{a}^\top \boldsymbol{y}$ mit den Koeffizienten*

$$\boldsymbol{a} = \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x} \quad \textit{und} \quad a_0 = \mu_x - \boldsymbol{a}^\top \boldsymbol{\mu_y}$$

hat minimalen MSE von allen linearen Vorhersagefunktionen von x. Es gilt

$$MSE = E(x - a_0 - \boldsymbol{a}^\top \boldsymbol{y})^2 = E(x - \mu_x - \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu_y}))^2 = \sigma_{xx} - \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x} .$$

Außerdem ist

$$a_0 + \boldsymbol{a}^\top \boldsymbol{y} = \mu_x + \boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_{\boldsymbol{y}})$$

jene lineare Vorhersagefunktion, die maximale Korrelation mit x aufweist, nämlich

$$Corr(x, a_0 + \boldsymbol{a}^\top \boldsymbol{y}) = \sqrt{\frac{\boldsymbol{a}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}} \boldsymbol{a}}{\sigma_{xx}}} = \sqrt{\frac{\boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x}}{\sigma_{xx}}} \ .$$

Beweis: Siehe z.B. Johnson und Wichern (1998).

Die Korrelation zwischen x und der besten linearen Vorhersagefunktion wird bezeichnet mit *multipler Korrelationskoeffizient* (der Grundgesamtheit)

$$\rho_{x,\boldsymbol{y}} = +\sqrt{\frac{\boldsymbol{\sigma}_{\boldsymbol{y}x}^\top \boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1} \boldsymbol{\sigma}_{\boldsymbol{y}x}}{\sigma xx}} \ .$$

Sein Quadrat $\rho_{x,\boldsymbol{y}}^2$ wird mit *multiples Bestimmtheitsmaß* (der Grundgesamtheit) bezeichnet, und es gibt an, wie gut das Merkmal x durch die Merkmale $\boldsymbol{y} = (y_1, \ldots, y_p)^\top$ erklärt wird.

Sind anstelle der Kovarianzen Korrelationen gegeben, so kann der multiple Korrelationskoeffizient auch definiert werden als

$$\rho_{x,\boldsymbol{y}}^2 = \boldsymbol{\rho}_{\boldsymbol{y}x}^\top \boldsymbol{\rho}_{\boldsymbol{yy}}^{-1} \boldsymbol{\rho}_{\boldsymbol{y}x} \quad .$$

Für konkrete Stichproben kann man die theoretischen Größen durch die entsprechenden Realisierungen schreiben, und man erhält für den multiplen Korrelationskoeffizienten

$$r_{x,\boldsymbol{y}}^2 = \boldsymbol{R}_{\boldsymbol{y}x}^\top \boldsymbol{R}_{\boldsymbol{yy}}^{-1} \boldsymbol{R}_{\boldsymbol{y}x} \quad .$$

Ein Test für die Hypothese, dass die multiple Korrelation Null ist, ist gleichwertig dazu, dass alle einfachen Korrelationen Null sind. Man testet also die Hypothese

$$H_0 : \rho_{x,\boldsymbol{y}} = 0 \ (= \rho_{xy_1} = \ldots = \rho_{xy_p})$$

gegen die Alternativhypothese

$$H_1 : \exists \ i \in \{1, \ldots, p\} \quad \text{mit} \quad \rho_{xy_i} \neq 0$$

zum Niveau α. Wenn n die Anzahl der Objekte in der Stichprobe ist, besitzt die Teststatistik

$$F = \frac{(n-1-p) \ r_{x,\boldsymbol{y}}^2}{p \ (1 - r_{x,\boldsymbol{y}}^2)}$$

eine $F_{p,n-1-p}$ -Verteilung und die Nullhypothese wird zum Niveau α verworfen, falls

$$F > F_{p,n-1-p;1-\alpha}$$

gilt.

7.2 Kanonische Korrelation

Bei der kanonischen Korrelation wird die lineare Abhängigkeit zwischen zwei Gruppen von Variablen berechnet. Wie sich herausstellen wird, kann diese Abhängigkeit nicht mehr durch *einen* Korrelationskoeffizienten ausgedrückt werden, sondern es ergibt sich ein Unterraum, der die lineare Abhängigkeit zwischen den Gruppen beschreibt.

Satz 7.2.1 *Sei $\boldsymbol{x}$ eine p-dimensionale und $\boldsymbol{y}$ eine q-dimensionale Zufallsgröße ($p \leq q$) mit den Erwartungswerten*

$$E(\boldsymbol{x}) = \boldsymbol{\mu}_1 \qquad \textit{und} \qquad E(\boldsymbol{y}) = \boldsymbol{\mu}_2 \ .$$

Die Kovarianzmatrizen $\boldsymbol{\Sigma}_{ij}$ mit $i, j = 1, 2$ sind definiert durch

$$\begin{aligned} \boldsymbol{\Sigma}_{11} &= E[(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top] \quad , \\ \boldsymbol{\Sigma}_{22} &= E[(\boldsymbol{y} - \boldsymbol{\mu}_2)(\boldsymbol{y} - \boldsymbol{\mu}_2)^\top] \quad \textit{und} \\ \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top &= E[(\boldsymbol{x} - \boldsymbol{\mu}_1)(\boldsymbol{y} - \boldsymbol{\mu}_2)^\top] \end{aligned}$$

und haben vollen Rang. Wir betrachten die Linearkombinationen $\varphi = \boldsymbol{a}^\top \boldsymbol{x}$ und $\eta = \boldsymbol{b}^\top \boldsymbol{y}$, wobei $\boldsymbol{a}$ ein p-dimensionaler und $\boldsymbol{b}$ ein q-dimensionaler Vektor ist.

Dann ist die einfache Korrelation zwischen φ und η gegeben durch

$$\max_{\boldsymbol{a},\boldsymbol{b}} \mathit{Corr}(\varphi, \eta) = \rho_1 \ .$$

Das Maximum wird erreicht durch die Linearkombinationen

$$\varphi_1 = \underbrace{\boldsymbol{e}_1^\top \boldsymbol{\Sigma}_{11}^{-1/2}}_{\boldsymbol{a}_1^\top} \boldsymbol{x} \quad \textit{und} \quad \eta_1 = \underbrace{\boldsymbol{f}_1^\top \boldsymbol{\Sigma}_{22}^{-1/2}}_{\boldsymbol{b}_1^\top} \boldsymbol{y} \ ,$$

die mit **erstes Paar von kanonischen Variablen** *bezeichnet werden. ρ_1 heißt* **erster kanonische Korrelationskoeffizient**.

Das ***k*-te Paar von kanonischen Variablen** *($k = 2, 3, \ldots, p$) ist gegeben durch*

$$\varphi_k = \underbrace{\boldsymbol{e}_k^\top \boldsymbol{\Sigma}_{11}^{-1/2}}_{\boldsymbol{a}_k^\top} \boldsymbol{x} \quad \textit{und} \quad \eta_k = \underbrace{\boldsymbol{f}_k^\top \boldsymbol{\Sigma}_{22}^{-1/2}}_{\boldsymbol{b}_k^\top} \boldsymbol{y}$$

und maximiert

$$\mathit{Corr}(\varphi_k, \eta_k) = \rho_k$$

über alle Linearkombinationen, die unkorreliert mit den vorherigen $1, 2, \ldots, k-1$ kanonischen Variablen sind. ρ_k heißt ***k*-ter kanonische Korrelationskoeffizient**.

$\rho_1^2 \geq \rho_2^2 \geq \cdots \geq \rho_p^2$ sind Eigenwerte der Matrix $\boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1/2}$ und $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$ die zugehörigen Eigenvektoren (Dimension p).

$\rho_1^2, \rho_2^2, \ldots, \rho_p^2$ sind auch die p größten Eigenwerte der Matrix $\boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1/2}$ mit zugehörigen Eigenvektoren $\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_p$ (Dimension q). Jedes $\boldsymbol{f}_i$ ist proportional zu $\boldsymbol{\Sigma}_{22}^{-1/2} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1/2} \boldsymbol{e}_i$.

Die kanonischen Variablen haben die Eigenschaften

$$\begin{aligned} Var(\varphi_k) &= Var(\eta_k) = 1 \\ Cov(\varphi_k, \varphi_l) &= Corr(\varphi_k, \varphi_l) = 0 \qquad k \neq l \\ Cov(\eta_k, \eta_l) &= Corr(\eta_k, \eta_l) = 0 \qquad k \neq l \\ Cov(\varphi_k, \eta_l) &= Corr(\varphi_k, \eta_l) = 0 \qquad k \neq l \end{aligned}$$

für $k, l = 1, 2, \ldots, p$, *bzw. mit der Notation* $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_p)^\top$ *und* $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top$ *und* $\boldsymbol{\rho} = Diag(\rho_1, \ldots, \rho_p)$ *gilt somit:*

$$Cov\begin{pmatrix} \boldsymbol{\varphi} \\ \boldsymbol{\eta} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{\rho} \\ \boldsymbol{\rho} & \boldsymbol{I} \end{pmatrix} .$$

Beweis: Siehe z.B. Johnson und Wichern (1998).

Bemerkung: Die Koeffizienten $\boldsymbol{a}_i$ und $\boldsymbol{b}_i$, die das Maximierungsproblem von Satz 7.2.1 lösen, können auch auf folgende Weise bestimmt werden:

ρ_i^2 von Satz 7.2.1 ist auch EW von $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ zum EV $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_i = \boldsymbol{a}_i$,

ρ_i^2 von Satz 7.2.1 ist auch EW von $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ zum EV $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{f}_i = \boldsymbol{b}_i$

$(i = 1, \ldots, p)$.

Die kanonischen Korrelationskoeffizienten sind invariant gegenüber linearen Transformationen, was u.a. im folgenden gezeigt wird.

Satz 7.2.2 *Ist* $\boldsymbol{x}^* = \boldsymbol{U}^\top\boldsymbol{x} + \boldsymbol{u}$ *und* $\boldsymbol{y}^* = \boldsymbol{V}^\top\boldsymbol{y} + \boldsymbol{v}$, *wobei* $\boldsymbol{U}$ *und* $\boldsymbol{V}$ *reguläre* $(p \times p)$- *bzw.* $(q \times q)$-*Matrizen und* $\boldsymbol{u}$ *und* $\boldsymbol{v}$ p- *bzw.* q-*dimensionale feste Vektoren sind. Dann gilt:*

(a) Die kanonischen Korrelationskoeffizienten zwischen $\boldsymbol{x}^*$ *und* $\boldsymbol{y}^*$ *sind dieselben wie jene zwischen* $\boldsymbol{x}$ *und* $\boldsymbol{y}$.

(b) Die Linearkombinationen, die die lineare Abhängigkeit zwischen $\boldsymbol{x}^*$ *und* $\boldsymbol{y}^*$ *maximieren, sind gegeben durch*

$$\boldsymbol{a}_i^* = \boldsymbol{U}^{-1}\boldsymbol{a}_i \qquad und \qquad \boldsymbol{b}_i^* = \boldsymbol{V}^{-1}\boldsymbol{b}_i \qquad für \quad i = 1, \ldots, p \ ,$$

wobei $\boldsymbol{a}_i$ *und* $\boldsymbol{b}_i$ *die entsprechenden Linearkombinationen für* $\boldsymbol{x}$ *und* $\boldsymbol{y}$ *sind.*

Beweis: Berechnet man statt den Matrizen $\boldsymbol{\Sigma}_{ij}$ für $\boldsymbol{x}$ und $\boldsymbol{y}$ die Matrizen $\boldsymbol{\Sigma}_{ij}^*$ für $\boldsymbol{x}^*$ und $\boldsymbol{y}^*$ $(i = 1, 2)$, so erhält man für $\boldsymbol{M}_1^* = \boldsymbol{\Sigma}_{11}^{*-1}\boldsymbol{\Sigma}_{12}^*\boldsymbol{\Sigma}_{22}^{*-1}\boldsymbol{\Sigma}_{21}^*$:

$$\boldsymbol{M}_1^* = (\boldsymbol{U}^\top\boldsymbol{\Sigma}_{11}\boldsymbol{U})^{-1}\boldsymbol{U}^\top\boldsymbol{\Sigma}_{12}\boldsymbol{V}(\boldsymbol{V}^\top\boldsymbol{\Sigma}_{22}\boldsymbol{V})^{-1}\boldsymbol{V}^\top\boldsymbol{\Sigma}_{21}\boldsymbol{U} = \boldsymbol{U}^{-1}\boldsymbol{M}_1\boldsymbol{U}$$

mit $\boldsymbol{M}_1 = \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$. Anwendung von Satz 1.4.1 mit $\boldsymbol{C} = \boldsymbol{U}^{-1}$ bzw. mit $\boldsymbol{C} = \boldsymbol{V}^{-1}$ ergibt die Aussage. □

Anwendung: Wählt man $\boldsymbol{U} = (diag\boldsymbol{\Sigma}_{11})^{-1/2}$ und $\boldsymbol{V} = (diag\boldsymbol{\Sigma}_{22})^{-1/2}$, so erhält man anstelle von $\boldsymbol{\Sigma}_{ij}$ die Korrelationsmatrizen $\boldsymbol{\rho}_{ij}$. Statt $\boldsymbol{M}_1$ erhält man $\boldsymbol{M}_1^* = \boldsymbol{\rho}_{11}^{-1}\boldsymbol{\rho}_{12}\boldsymbol{\rho}_{22}^{-1}\boldsymbol{\rho}_{21}$.
Sind also nicht die Kovarianzmatrizen sondern die Korrelationsmatrizen $\boldsymbol{\rho}_{ij}$ gegeben, so sind

(a) die kanonischen Korrelationskoeffizienten bestimmt durch die Wurzeln aus den Eigenwerten von $\boldsymbol{M}_1^*$ (oder $\boldsymbol{M}_2^*$),

(b) die Linearkombinationen $\boldsymbol{a}_i$ und $\boldsymbol{b}_i$ bestimmt durch die Transformation $\boldsymbol{a}_i = \boldsymbol{U}\boldsymbol{a}_i^*$ bzw. $\boldsymbol{b}_i = \boldsymbol{V}\boldsymbol{b}_i^*$, wobei $\boldsymbol{a}_i^*$ bzw. $\boldsymbol{b}_i^*$ die EV von $\boldsymbol{M}_1^*$ bzw. $\boldsymbol{M}_2^*$ sind. Es müssen allerdings die Varianzen von x_i und y_i bekannt sein.

Beispiel 7.2.1 *Alle bisher formulierten Sätze lassen sich natürlich auf Stichproben übertragen. Um den Unterschied zwischen Grundgesamtheit und Stichprobe deutlich zu machen, schreiben wir $\boldsymbol{S}_{ij}$ anstelle von $\boldsymbol{\Sigma}_{ij}$, $\boldsymbol{R}_{ij}$ statt $\boldsymbol{\rho}_{ij}$ und r_i statt ρ_i .*

Gegeben seien die Prüfungsdaten aus Tabelle 2.1. Es wird die kanonische Korrelation zwischen den Fächern Mechanik (ME), Analytische Geometrie (AG) und den Fächern Lineare Algebra (LA), Analysis (AN), Elementare Statistik (ES) berechnet. Die Korrelationsmatrizen der Daten sind

$$\boldsymbol{R}_{11} = \begin{pmatrix} 1 & 0.5534 \\ 0.5534 & 1 \end{pmatrix} \quad , \qquad \boldsymbol{R}_{22} = \begin{pmatrix} 1 & 0.7108 & 0.6647 \\ 0.7108 & 1 & 0.6072 \\ 0.6647 & 0.6072 & 1 \end{pmatrix} \quad ,$$

$$\boldsymbol{R}_{12} = \boldsymbol{R}_{21}^\top = \begin{pmatrix} 0.5468 & 0.4094 & 0.3891 \\ 0.6096 & 0.4851 & 0.4364 \end{pmatrix} \quad .$$

Aus diesen Korrelationsmatrizen können nach Satz 7.2.2 die Matrizen $\boldsymbol{M}_1^*$ und $\boldsymbol{M}_2^*$ berechnet werden. Deren Eigenwerte sind 0.4396 *und* 0.0017 *. Die kanonischen Korrelationskoeffizienten sind die Wurzeln aus den Eigenwerten, d.h. $r_1 = 0.6631$ und $r_2 = 0.0409$. Möchte man zusätzlich die kanonischen Variablen erhalten, so berechnet man vorerst $\boldsymbol{a}_i^*$ und $\boldsymbol{b}_i^*$ für $i = 1, 2$, wobei in die entsprechenden Formeln statt den Kovarianzmatrizen die Korrelationsmatrizen eingesetzt werden. Es ergibt sich*

$$\boldsymbol{a}_1^* = \begin{pmatrix} 0.4517 \\ 0.6765 \end{pmatrix} \quad , \qquad \boldsymbol{a}_2^* = \begin{pmatrix} 1.1124 \\ -0.9918 \end{pmatrix} \quad ,$$

$$\boldsymbol{b}_1^* = \begin{pmatrix} 0.8703 \\ 0.1191 \\ 0.0596 \end{pmatrix} \quad , \qquad \boldsymbol{b}_2^* = \begin{pmatrix} 0.9600 \\ -1.4608 \\ 0.2473 \end{pmatrix} \quad .$$

Um die Vektoren $\boldsymbol{a}_i$ und $\boldsymbol{b}_i$ zu bestimmen, müssen die Matrizen

$$\boldsymbol{U} = (diag\boldsymbol{S}_{11})^{-1/2} = \begin{pmatrix} 0.0572 & 0 \\ 0 & 0.0761 \end{pmatrix} \quad \textit{und}$$

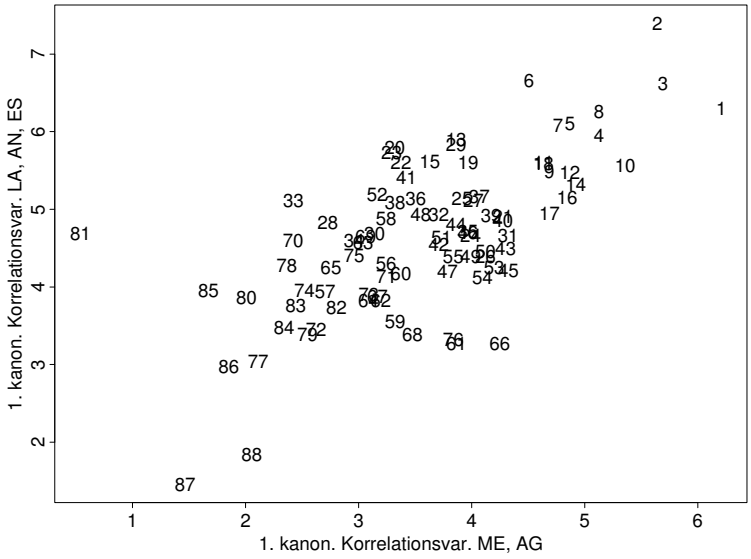
$$\boldsymbol{V} = (diag\boldsymbol{S}_{22})^{-1/2} = \begin{pmatrix} 0.0941 & 0 & 0 \\ 0 & 0.0674 & 0 \\ 0 & 0 & 0.0580 \end{pmatrix}$$

berechnet werden. Die kanonischen Variablen können nun direkt bestimmt werden, und es ergibt sich

$$\varphi_1 = 0.0258\ x_1 + 0.0515\ x_2 \quad , \qquad \eta_1 = 0.0819\ y_1 + 0.0080\ y_2 + 0.0035\ y_3 \quad ,$$
$$\varphi_2 = 0.0636\ x_1 - 0.0754\ x_2 \quad , \qquad \eta_2 = 0.0904\ y_1 - 0.0984\ y_2 + 0.0143\ y_3 \quad .$$

Die beiden ersten kanonischen Variablen sind in Abbildung 7.1 einander gegenübergestellt.

Abbildung 7.1: Erstes Paar von kanonischen Variablen ($r_1 = 0.663$)



In der kanonischen Korrelationsanalyse gibt es eine Reihe von **Tests**, von denen hier nur ausgewählte erwähnt werden, da die Verteilungen der Schätzer sehr kompliziert sind. Mit der Annahme, dass die Daten normalverteilt sind, gilt:

(a) Ein Plausibilitätsquotiententest für die Hypothese $H_0 : \boldsymbol{\Sigma}_{12} = \boldsymbol{O}$, d.h. für die Hypothese, dass $\boldsymbol{x}$ und $\boldsymbol{y}$ unkorreliert sind, ist durch die Teststatistik

$$\lambda^{2/n} = |\boldsymbol{I} - \boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}| = \prod_{i=1}^{p}(1 - r_i^2)$$

gegeben. Diese Teststatistik hat eine $\Lambda(q, n-1-p, p)$-Wilks-Verteilung. n ist der Stichprobenumfang und $r_1, \ldots, r_p$ sind die kanonischen Korrelationskoeffizienten.

(b) Da die Wilks-Verteilung als Produkt von Betaverteilungen relativ kompliziert ist, kann man sie (mit Bartlett's Approximation) für großen Stichprobenumfang n annähern durch

$$-\left[n - \frac{1}{2}\Big(p + q + 3\Big)\right] \ln \prod_{i=1}^{p}(1 - r_i^2) \sim \chi^2_{pq} \quad .$$

(c) Eine ähnliche Teststatistik kann auch für die Hypothese formuliert werden, dass nur s der kanonischen Korrelationskoeffizienten ungleich 0 sind, nämlich

$$-\left[n - \frac{1}{2}\Big(p + q + 3\Big)\right] \ln \prod_{i=s+1}^{p}(1 - r_i^2) \sim \chi^2_{(p-s)(q-s)} \quad .$$

Beispiel 7.2.2 *Für das Beispiel mit den Prüfungsdaten gilt* $n = 88$, $p = 2$ *und* $q = 3$. *Die Statistik zur Hypothese* $\rho_1 = \rho_2 = 0$ *ist nach (b) gegeben durch*

$$-\left(88 - \frac{8}{2}\right) \ln\left[(1 - 0.6631^2)(1 - 0.0409^2)\right] \approx 49 \ .$$

Dieser Wert liegt deutlich über $\chi^2_{6;0.95} \approx 12.6$, *die Hypothese wird daher verworfen.*
Testet man aber nur die Hypothese $\rho_2 = 0$, *so ist die Teststatistik nach (c)*

$$-\left(88 - \frac{8}{2}\right) \ln(1 - 0.0409^2) \approx 0.141 \ .$$

Dies ist beim Test gegen $\chi^2_{2;0.95} \approx 5.991$ *nicht signifikant, weshalb die Hypothese nicht verworfen werden kann.*

Literatur

W.R. Dillon and M. Goldstein. *Multivariate Analysis: Methods and Applications.* John Wiley & Sons, New York, 1984.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 4th edition, 1998.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis.* Acad. Press, London, 1979.

G.A.F. Seber. *Multivariate observations.* John Wiley & Sons, New York, 1984.

Kapitel 8

Diskriminanzanalyse

8.1 Einleitung

Der Ausdruck *Diskriminanzanalyse* wurde von Fisher (1938) geprägt. Es handelt sich dabei um eine multivariate Methode, die sich einerseits mit der Klassifizierung verschiedener Objektgruppen und andererseits mit der Zuweisung neuer Objekte zu vorher bestimmten Gruppen befasst. Im ersteren Fall wird versucht, die Unterschiede der Objekte, von denen man weiß, dass sie aus zwei oder mehreren Grundgesamtheiten stammen, grafisch oder algebraisch festzuhalten. Man sucht somit nach einer "Diskriminanzfunktion", die eine möglichst gute Trennung zulässt. Im zweiten Fall möchte man die Objekte in zwei oder mehrere Gruppen einteilen. Das Ziel ist dann, durch festgelegte Regeln *neue* Objekte zu klassifizieren. Obige Fälle stehen oft in unmittelbarem Zusammenhang, denn eine Funktion, die Objekte trennt, kann auch dazu benützt werden, um neue Objekte einzuordnen, bzw. umgekehrt.

8.2 Überlegungen zu Klassifikationsregeln

Die Objekte werden klassifiziert anhand von Messungen von p zugrundeliegenden Zufallsvariablen $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$. Geht man davon aus, dass zwei Gruppen vorhanden sind, so werden die zu messenden Objekte dann eingeteilt in die Klassen π_1 bzw. π_2. Die Gesamtheit der Werte der ersten Klasse sei die Population der $\boldsymbol{x}$-Werte von π_1, und jene der zweiten Klasse die Population der $\boldsymbol{x}$-Werte von π_2. Die beiden Populationen werden dann beschrieben durch die Wahrscheinlichkeitsverteilungen $f_1(\boldsymbol{x})$ und $f_2(\boldsymbol{x})$.

Beispielsweise könnten die beiden Populationen π_1 und π_2 gegeben sein durch Kompositionen von J.S. Bach und L.v. Beethoven. Die zu messenden Variablen $\boldsymbol{X}$ sind Häufigkeiten verschiedener Akkorde oder Tonfolgen. Es ist dann von Interesse, eine Diskriminanzfunktion zu finden, die beschreibt, inwiefern sich die beiden Komponisten anhand ihrer Werke unterscheiden.

Sei Ω der gesamte Stichprobenraum, also jener Raum, der alle Beobachtungen $\boldsymbol{x}$ enthält. Jedes Objekt $\boldsymbol{x}$ muss entweder aus der Population π_1 oder aus π_2 kommen. Weiters sei R_1 jener Raum von Beobachtungen $\boldsymbol{x}$, dem wir die Objekte aus π_1

zuordnen, und R_2 jener Raum, dem die restlichen Objekte aus π_2 zugeordnet werden. Ω ist die Vereinigung von R_1 und R_2.

Bei der Zuordnung kann nun passieren, dass Objekte, die in Wirklichkeit von der Population π_1 kommen, fälschlicherweise als π_2 klassifiziert werden. Sind die Wahrscheinlichkeitsfunktionen $f_1(\boldsymbol{x})$ und $f_2(\boldsymbol{x})$ bekannt, so kann diese Wahrscheinlichkeit der falschen Zuordnung als bedingte Wahrscheinlichkeit $P(2|1)$ berechnet werden durch

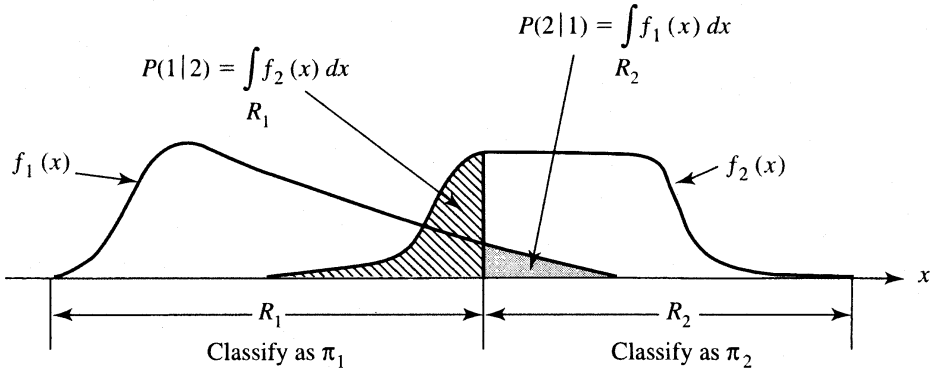
$$P(2|1) = P(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2=\Omega-R_1} f_1(\boldsymbol{x})d\boldsymbol{x} \ . \tag{8.1}$$

Umgekehrt kann der Fall eintreten, dass Objekte, die aus der Population π_2 stammen, fälschlicherweise π_1 zugeordnet werden. Die entsprechende Wahrscheinlichkeit ist

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x} \ . \tag{8.2}$$

Das Integral in (8.1) beschreibt das Volumen der Dichtefunktion $f_1(\boldsymbol{x})$ über dem Gebiet R_2, bzw. analog für (8.2).

Abbildung 8.1: Wahrscheinlichkeiten bei Missklassifikation



Sei nun p_1 die Wahrscheinlichkeit, dass die Objekte von π_1 kommen (*a-priori Wahrscheinlichkeit*), und p_2 jene für π_2, wobei gelten muss $p_1 + p_2 = 1$. Dann können durch Anwenden der Formel für bedingte Wahrscheinlichkeiten folgende Wahrscheinlichkeiten berechnet werden:

$$\begin{aligned}
P(\text{Beob. richtig als } \pi_1 \text{ klassifiziert}) &= P(\boldsymbol{X} \in R_1|\pi_1)P(\pi_1) = P(1|1)p_1 \\
P(\text{Beob. falsch als } \pi_1 \text{ klassifiziert}) &= P(\boldsymbol{X} \in R_1|\pi_2)P(\pi_2) = P(1|2)p_2 \\
P(\text{Beob. richtig als } \pi_2 \text{ klassifiziert}) &= P(\boldsymbol{X} \in R_2|\pi_2)P(\pi_2) = P(2|2)p_2 \\
P(\text{Beob. falsch als } \pi_2 \text{ klassifiziert}) &= P(\boldsymbol{X} \in R_2|\pi_1)P(\pi_1) = P(2|1)p_1
\end{aligned}$$

Missklassifikation ist oft direkt mit Kosten verbunden. Die Kosten sind natürlich 0, wenn korrekt klassifiziert wurde. Sie betragen $c(1|2)$, wenn eine Beobachtung von π_2 falsch als π_1 klassifiziert wird. Und die Kosten sind $c(2|1)$, wenn Beobachtungen von π_1 fälschlicherweise als π_2 klassifiziert werden. Gemeinsam mit den Wahrscheinlichkeiten für Missklassifikation können nun die *erwarteten Kosten bei Missklassifikation* (EKM) berechnet werden als

$$EKM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \ . \tag{8.3}$$

Ziel einer Klassifikationsregel ist es, EKM so klein wie möglich zu halten. Demnach könnte eine Klassifikationsregel, die EKM minimiert, folgendermaßen lauten:
Die Menge R_1 ist definiert für Beobachtungen $\boldsymbol{x}$, für die gilt:

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{8.4}$$

Die Menge R_2 ist definiert für Beobachtungen $\boldsymbol{x}$, für die gilt:

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \tag{8.5}$$

Daraus leiten sich folgende Spezialfälle ab:

(a) $p_1 = p_2$:

$$R_1: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)}{c(2|1)}$$

(b) $c(1|2) = c(2|1)$:

$$R_1: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq \frac{p_2}{p_1} \qquad R_2: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{p_2}{p_1} \tag{8.6}$$

(c) $p_1 = p_2$ und $c(1|2) = c(2|1)$:

$$R_1: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} \geq 1 \qquad R_2: \quad \frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < 1$$

Es können aber auch andere Kriterien zur Erstellung einer Klassifikationsregel herangezogen werden. Z.B. könnte man R_1 und R_2 so wählen, dass die *gesamte Wahrscheinlichkeit der Missklassifikation* (GWM) minimal wird, also

$$\begin{aligned} GWM &= P(\text{Missklassifizierung einer Beobachtung aus } \pi_1 \text{ oder } \pi_2) \\ &= p_1 \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \qquad (8.7) \\ &= p_1 P(2|1) + p_2 P(1|2)\ . \qquad (8.8) \end{aligned}$$

Unter der Voraussetzung, dass die Kosten bei Missklassifikation gleich sind, erkennt man sofort, dass Minimierung von (8.8) äquivalent zur Minimierung von (8.3) ist.

8.3 Der Zweigruppenfall

Wir beschränken uns hier auf multivariat normalverteilte Grundgesamtheiten. Somit sind $f_1(\boldsymbol{x})$ und $f_2(\boldsymbol{x})$ Dichtefunktionen multivariater Normalverteilungen mit Mittel $\boldsymbol{\mu}_1$ bzw. $\boldsymbol{\mu}_2$ und Kovarianzmatrizen $\boldsymbol{\Sigma}_1$ bzw. $\boldsymbol{\Sigma}_2$.

8.3.1 Spezialfall $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$

Die gemeinsame Dichtefunktion der Zufallsvariable $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ für die Grundgesamtheiten π_1 und π_2 ist gegeben durch

$$f_i(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right\} \quad \text{für } i = 1, 2\ . \tag{8.9}$$

Sind die Parameter $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ und $\boldsymbol{\Sigma}$ bekannt, dann ist entsprechend (8.4) der Bereich, der EKM minimiert:

$$\begin{aligned} R_1:\ \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} \\ \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \\ R_2:\ \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right\} \\ < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) \end{aligned} \tag{8.10}$$

Anhand dieser Bereiche R_1 und R_2 kann folgende Klassifikationsregel angegeben werden:

Satz 8.3.1 *Seien π_1 und π_2 normalverteilte Populationen mit gleicher Kovarianz. Dann lautet die Klassifikationsregel zur Minimierung von EKM:*
Eine Beobachtung $\boldsymbol{x}_0$ wird π_1 zugeteilt, wenn

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\ \frac{p_2}{p_1}\right)\ . \tag{8.11}$$

Ansonsten wird $\boldsymbol{x}_0$ π_2 zugeordnet.

Beweis: Nachdem alle Größen in (8.10) nichtnegativ sind, ändern sich die Regeln durch Logarithmieren nicht. Weiters gilt:

$$\begin{aligned} &-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) + \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) \\ &= \tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \tfrac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \tfrac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \tfrac{1}{2}\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 - \tfrac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \tfrac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \\ &= \boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \tfrac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + \tfrac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \end{aligned}$$

Obige Zusammenfassung kann erfolgen, weil die einzelnen Summanden Skalare sind und daher $a = a^\top$ gilt. Erweitert man nun mit $\frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2$, so ergibt sich weiters:

$$\begin{aligned} &(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \tfrac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \tfrac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \tfrac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 + \tfrac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_2 \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \tfrac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \tfrac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \tfrac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \end{aligned}$$

Daraus folgt unmittelbar die Aussage. □

Die in Satz 8.3.1 definierte Diskriminanzfunktion ist *linear* in $\boldsymbol{x}$. Meistens sind die Größen $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ und $\boldsymbol{\Sigma}$ nicht gegeben, und sie müssen aus der Stichprobe geschätzt werden. Seien von der Zufallsvariablen $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ für die Population π_1 n_1 Stichproben vorhanden und n_2 Messungen für π_2. Die entsprechenden Datenmatrizen der Dimensionen $(n_1 \times p)$ bzw. $(n_2 \times p)$ seien bezeichnet mit

$$\boldsymbol{X}_1 = \begin{pmatrix} \boldsymbol{x}_{11}^\top \\ \boldsymbol{x}_{12}^\top \\ \vdots \\ \boldsymbol{x}_{1n_1}^\top \end{pmatrix} \qquad \boldsymbol{X}_2 = \begin{pmatrix} \boldsymbol{x}_{21}^\top \\ \boldsymbol{x}_{22}^\top \\ \vdots \\ \boldsymbol{x}_{2n_2}^\top \end{pmatrix} \tag{8.12}$$

Die arithmetischen Mittelvektoren und empirischen Kovarianzmatrizen sind dann gegeben durch

$$\bar{\boldsymbol{x}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \boldsymbol{x}_{1j} \qquad \boldsymbol{S}_1 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1} (\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)(\boldsymbol{x}_{1j} - \bar{\boldsymbol{x}}_1)^\top$$

$$\bar{\boldsymbol{x}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \boldsymbol{x}_{2j} \qquad \boldsymbol{S}_2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2} (\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2)(\boldsymbol{x}_{2j} - \bar{\boldsymbol{x}}_2)^\top$$

Nachdem aber angenommen wurde, dass π_1 und π_2 Populationen mit gleicher Kovarianzmatrix $\boldsymbol{\Sigma}$ sind, werden $\boldsymbol{S}_1$ und $\boldsymbol{S}_2$ kombiniert zu einer "gepoolten" Kovarianzmatrix

$$\boldsymbol{S}_{pooled} = \left(\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right)\boldsymbol{S}_1 + \left(\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right)\boldsymbol{S}_2 \ .$$

$\boldsymbol{S}_{pooled}$ ist ein unverzerrter Schätzer von $\boldsymbol{\Sigma}$.

Werden nun in (8.11) die Schätzungen eingesetzt, so ergibt sich die Stichproben-Klassifikationsregel:

Satz 8.3.2 *Eine Beobachtung $\boldsymbol{x}_0$ wird π_1 zugeteilt, wenn*

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \geq \ln\left(\frac{c(1|2)}{c(2|1)}\ \frac{p_2}{p_1}\right) \ . \tag{8.13}$$

Ansonsten wird $\boldsymbol{x}_0$ π_2 zugeordnet.

Nachdem die Summanden in (8.13) Skalare sind, kann die Regel für den Fall, dass in (8.13) $\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right) = 1$ gilt $(\ln(1) = 0)$, auf folgende Weise vereinfacht werden:

$$\hat{y} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}\boldsymbol{x} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} \tag{8.14}$$

wird ausgewertet an der Stelle $\boldsymbol{x}_0$ und danach verglichen mit der Zahl

$$\hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) \ ,$$

wobei

$$\bar{y}_1 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \bar{\boldsymbol{x}}_1 = \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1$$

und

$$\bar{y}_2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \bar{\boldsymbol{x}}_2 = \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2 \ .$$

Somit reduziert die Regel bei p-dimensionalen Merkmalen die Entscheidung auf eine 1-dimensionale Größe y, die aus entsprechenden Linearkombinationen der Beobachtungen von π_1 und π_2 hervorgeht.

Beispiel 8.3.1 *Dieses Beispiel stammt aus einer Studie von Bouma et al. (1975), die sich mit der Erkennung der Bluterkrankheit beschäftigt. Es gibt verschiedene Typen dieser Krankheit, und die Studie zielt darauf ab, Träger von Hemophilie A zu erkennen. Gemessen wurden die beiden Variablen*

$$\begin{aligned} X_1 &= \log_{10}(\textit{AHF Aktivität}) \\ X_2 &= \log_{10}(\textit{AHF-ähnliches Antigen}) \ , \end{aligned}$$

wobei AHF den Anti-Hemophilie-Faktor bezeichnet. Es wurden die Werte von zwei Gruppen von Frauen gemessen, wobei die erste Gruppe mit $n_1 = 30$ Frauen das Hemophilie-Gen nicht trägt und die zweite Gruppe mit $n_2 = 22$ Frauen Träger von Hemophilie A sind. Die Messungen sind in Abbildung 8.2 eingetragen. Es sind außerdem die geschätzten Konturlinien mit Zentrum $\bar{\boldsymbol{x}}_1$ bzw. $\bar{\boldsymbol{x}}_2$ eingezeichnet, die 50% bzw. 95% der Beobachtungen enthalten. Die geschätzten Mittel sind

$$\bar{\boldsymbol{x}}_1 = \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix} \quad \textit{und} \quad \bar{\boldsymbol{x}}_2 = \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix}$$

und die gepoolte Kovarianzmatrix

$$\boldsymbol{S}_{pooled}^{-1} = \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix} \ .$$

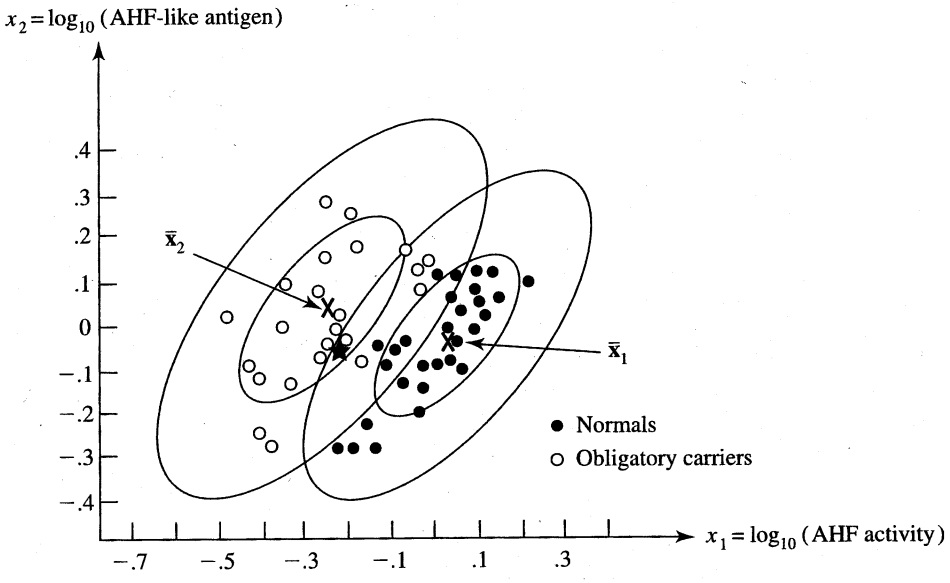
Unter der Annahme, dass die Kosten bei Missklassifikation sowie die a-priori Wahrscheinlichkeiten gleich sind, ergibt sich durch Einsetzen in (8.14)

$$\begin{aligned} \hat{y} &= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} \\ &= (0.2418 \quad -0.0652) \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= 37.61 x_1 - 28.92 x_2 \ . \end{aligned} \tag{8.15}$$

Weiters sind

$$\begin{aligned} \bar{y}_1 &= \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1 = (37.61 \quad -28.92) \begin{pmatrix} -0.0065 \\ -0.0390 \end{pmatrix} = 0.88 \\ \bar{y}_2 &= \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2 = (37.61 \quad -28.92) \begin{pmatrix} -0.2483 \\ 0.0262 \end{pmatrix} = -10.10 \end{aligned}$$

Abbildung 8.2: Plot der Hemophilie Daten



und der Mittelpunkt zwischen diesen Mitteln

$$\hat{m} = \frac{1}{2}(\bar{y}_1 + \bar{y}_2) = \frac{1}{2}(0.88 - 10.10) = -4.61 \ .$$

Es liegt nun bei einer Frau der Verdacht vor, dass sie Trägerin der Hemophilie A ist. Ihre Werte sind

$$x_1 = -0.210 \quad \textit{und} \quad x_2 = -0.044 \ .$$

Um nun festzustellen, ob sie der Gruppe π_1 (normal) oder der Gruppe π_2 (Träger) zuzuordnen ist, setzen wir in (8.15) die Werte ein und verwenden Klassifikationsregel (8.13):

$$\begin{aligned} &\textit{Weise } \boldsymbol{x}_0 \textit{ der Gruppe } \pi_1 \textit{ zu, wenn} \quad \hat{y}_0 = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_0 \geq \hat{m} = -4.61 \\ &\textit{Weise } \boldsymbol{x}_0 \textit{ der Gruppe } \pi_2 \textit{ zu, wenn} \quad \hat{y}_0 = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_0 < \hat{m} = -4.61 \end{aligned}$$

mit $\boldsymbol{x}_0 = (-0.210, -0.044)^\top$. Nachdem

$$\hat{y}_0 = \hat{\boldsymbol{a}}^\top \boldsymbol{x}_0 = (37.61 \quad -28.92)\begin{pmatrix} -0.210 \\ -0.044 \end{pmatrix} = -6.62 < -4.61$$

ist, wird die Frau der Gruppe π_2, also zur Gruppe der Träger der Hemophilie A zugewiesen. Diese Beobachtung ist in Abbildung 8.2 als Stern dargestellt, und man erkennt, dass sie innerhalb des geschätzten 50% Bereichs von π_2 liegt und etwa auf der 95% Konturlinie von π_1. Die Gruppen sind also nicht klar voneinander getrennt.

Es seien nun die a-priori Wahrscheinlichkeiten p_1 und p_2 für die Gruppenzugehörigkeiten bekannt. Z.B. sei die Frau, die vorhin klassifiziert wurde, eine Cousine mütterlicherseits von einer Trägerin der Hemophilie A. Dann ist nämlich die

genetische Wahrscheinlichkeit, dass auch diese Frau Trägerin ist gleich 0.25. *Somit sind* $p_1 = 0.75$ *(normal) und* $p_2 = 0.25$ *(Träger). Seien weiters die Kosten bei Missklassifikation gleich (unrealistisch), ergibt sich durch Einsetzen in (8.13)*

$$(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) = -2.01 \ .$$

Nachdem nun

$$-2.01 < \ln\left(\frac{p_2}{p_1}\right) = \ln\left(\frac{0.25}{0.75}\right) = -1.10$$

ist, wird die Frau wiederum der Gruppe π_2 *der Träger von Hemophilie A zugewiesen.*

Bemerkung: Die Koeffizienten $\hat{\boldsymbol{a}} = \boldsymbol{S}_{pooled}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ sind eindeutig bis auf eine multiplikative Konstante, denn jeder Vektor $c\hat{\boldsymbol{a}}$ mit $c \neq 0$ kann auch als Koeffizientenvektor zur Diskriminierung herangezogen werden. Üblicherweise wird man daher eine Standardisierung vornehmen, z.B.

$$\hat{\boldsymbol{a}}^* = \frac{\hat{\boldsymbol{a}}}{\sqrt{\hat{\boldsymbol{a}}^\top \hat{\boldsymbol{a}}}}$$

bringt $\hat{\boldsymbol{a}}^*$ auf Länge 1.

8.3.2 Spezialfall $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$

Die bisher formulierten Klassifikationsregeln waren basierend auf (8.4) und haben das Verhältnis der Dichtefunktionen $f_1(\boldsymbol{x})/f_2(\boldsymbol{x})$ betrachtet. Beim Fall gleicher Kovarianzmatrizen reduziert sich dieses Verhältnis zu einem relativ einfachen Term, was sich in der Regel (8.11) ausdrückt. Im Fall ungleicher Kovarianzmatrizen (und ungleicher Mittel) wird allerdings das Verhältnis der Dichtefunktionen ein komplizierterer Ausdruck. Wie bei (8.11) kann auch hier der Logarithmus des Verhältnisses betrachtet werden:

$$\begin{aligned}
\ln\left(\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}\right) &= \ln\left(\frac{|\boldsymbol{\Sigma}_2|^{1/2}}{|\boldsymbol{\Sigma}_1|^{1/2}}\right) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_1^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2) \\
&= \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\right) - \frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} \\
&\quad -\frac{1}{2}\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2
\end{aligned}$$

Damit ergibt sich folgende Klassifikationsregel:

$$\begin{aligned}
R_1: \ & -\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} - k \geq \ln\left(\frac{c(1|2)}{c(2|1)}\,\frac{p_2}{p_1}\right) \\
R_2: \ & -\frac{1}{2}\boldsymbol{x}^\top(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} - k < \ln\left(\frac{c(1|2)}{c(2|1)}\,\frac{p_2}{p_1}\right)
\end{aligned}$$

mit

$$k = \frac{1}{2} \ln \left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|} \right) + \frac{1}{2} (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} \boldsymbol{\mu}_1 - \frac{1}{2} \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1} \boldsymbol{\mu}_2) \ .$$

Die Konstante k hängt nur mehr von Mittel und Kovarianz der beiden Verteilungen ab. Für den Fall, dass $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ wird, reduziert sich obige Regel auf (8.11). Es ergibt sich jetzt direkt folgende Aussage:

Satz 8.3.3 *Seien π_1 und π_2 normalverteilte Populationen mit Mittelvektoren $\boldsymbol{\mu}_1$ bzw. $\boldsymbol{\mu}_2$ und Kovarianzmatrizen $\boldsymbol{\Sigma}_1$ bzw. $\boldsymbol{\Sigma}_2$. Dann lautet die Klassifikationsregel zur Minimierung von EKM:*
Eine Beobachtung $\boldsymbol{x}_0$ wird π_1 zugeteilt, wenn

$$-\frac{1}{2} \boldsymbol{x}_0^\top (\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1}) \boldsymbol{x}_0 + (\boldsymbol{\mu}_1^\top \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2^\top \boldsymbol{\Sigma}_2^{-1}) \boldsymbol{x}_0 - k \geq \ln \left(\frac{c(1|2)}{c(2|1)} \ \frac{p_2}{p_1} \right) \ . \qquad (8.16)$$

Ansonsten wird $\boldsymbol{x}_0$ π_2 zugeordnet.

Obige Diskriminanzfunktion ist *quadratisch* in $\boldsymbol{x}$. Nachdem die Populationsgrößen $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ und $\boldsymbol{\Sigma}_2$ i.a. nicht bekannt sind, werden sie durch die entsprechenden Stichprobengrößen $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$, $\boldsymbol{S}_1$ und $\boldsymbol{S}_2$ geschätzt.

8.3.3 Auswertung der Klassifikation

Obige Klassifikationregeln gehen von normalverteilten Daten aus. Ist diese Voraussetzung nicht erfüllt, so können die Regeln, die sich durch lineare oder quadratische Diskriminanzfunktionen ausdrücken, stark fehlschlagen. Man könnte daher vorher die Daten so transformieren, dass sie der (multivariaten) Normalverteilung besser entsprechen. Jedenfalls sollte aber überprüft werden, wie gut die Diskriminierung beider Populationen tatsächlich war. Ein Instrumentarium dafür ist die *tatsächliche Fehlerrate* (TFR),

$$TFR = p_1 \int_{\hat{R}_2} f_1(\boldsymbol{x}) d\boldsymbol{x} + p_2 \int_{\hat{R}_1} f_2(\boldsymbol{x}) d\boldsymbol{x} \ , \qquad (8.17)$$

die sich direkt aus der gesamten Wahrscheinlichkeit der Missklassifikation (GWM) (8.7) herleitet. $\hat{R}_1$ und $\hat{R}_2$ sind dabei die aus der Stichprobe ermittelten Bereiche, wenn z.B. Regel (8.13) verwendet wird (Voraussetzung gleiche Kovarianzmatrix).

Man könnte auch einfach den Anteil der falsch klassifizierten Objekte an der Gesamtanzahl der Objekte angeben. Aufgrund einer Stichprobe der Größe n_1 von π_1 und n_2 von π_2 kann dieser Anteil durch Anwenden einer Klassifikationsregel ermittelt werden.

Diese Vorgehensweisen setzen allerdings immer voraus, dass zumindest für einen "Trainingsdatensatz" bekannt ist, welche Objekte welcher Gruppe angehören. Daraus kann die Diskriminanzfunktion ermittelt werden und die Fehlerrate berechnet werden. Neue Objekte können dann entsprechend der vorher ermittelten Regel mit bekannter Fehlerrate klassifiziert werden.

Lachenbruch und Mickey (1968) schlugen folgende "jackknife"-Prozedur zur Schätzung des Fehleranteils vor:

1. Starte mit den Objekten der Gruppe π_1. Es wird eine Beobachtung dieser Gruppe weggelassen und eine Klassifikationsfunktion mit den restlichen $n_1 - 1$ und n_2 Beobachtungen bestimmt.

2. Die in Schritt 1 weggelassene Beobachtung wird aufgrund der ermittelten Klassifikationsfunktion klassifiziert.

3. Wiederhole Schritt 1 und 2 bis alle Objekte von π_1 klassifiziert sind. $\bar{n}_1$ sei die Anzahl der falsch klassifizierten Objekte.

4. Wiederhole Schritte 1 bis 3 für die Objekte von π_2. Es ergibt sich die Anzahl $\bar{n}_2$ der falsch klassifizierten Objekte von Gruppe π_2.

5. Nun können die bedingten Wahrscheinlichkeiten für Fehlklassifikation (8.1) und (8.2) geschätzt werden durch

$$\hat{P}(2|1) = \frac{\bar{n}_1}{n_1} \qquad \hat{P}(1|2) = \frac{\bar{n}_2}{n_2}$$

   und daraus die geschätzte Fehlerrate

$$\frac{\bar{n}_1 + \bar{n}_2}{n_1 + n_2}$$

   ermittelt werden.

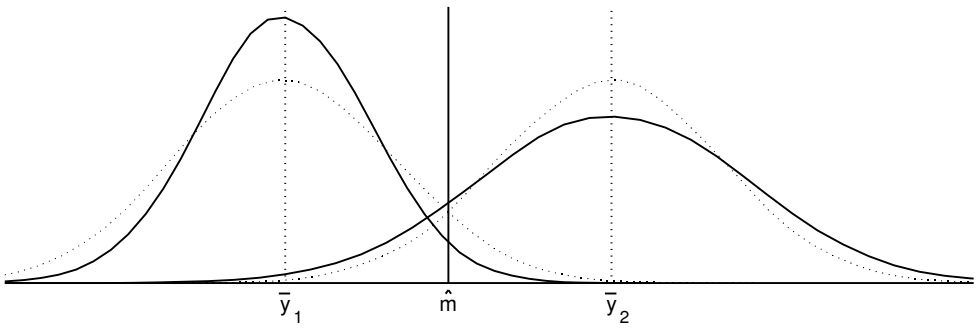
Die praktische Durchführung dieser Prozedur ist möglich bei relativ kleinen Anzahlen n_1 und n_2 bzw. bei Verwendung einer wenig rechenaufwendigen Klassifikationsregel.

Bemerkung: Bei Verwendung der linearen Diskriminanzfunktion aus Satz 8.3.1 ist darauf zu achten, dass die Varianzen der beiden Gruppen tatsächlich als gleich anzunehmen sind. Ist dies nämlich nicht der Fall, kann es passieren, dass die Wahrscheinlichkeiten bei Missklassifikation stark unterschiedlich werden. Dies ist in Abbildung 8.3 dargestellt. Es wurde die lineare Diskriminanzfunktion (8.14) verwendet. Die punktierten Kurven zeigen die Verteilungen der beiden Populationen um $\bar{y}_1$ und $\bar{y}_2$ mit der gepoolten Varianz. Anhand des Mittelpunktes $\hat{m}$ zwischen $\bar{y}_1$ und $\bar{y}_2$ wird nun für jede Beobachtung entschieden, zu welcher Population sie gezählt wird. Ist nun allerdings die Varianz der Populationen stark unterschiedlich, wie durch die tatsächlichen Verteilungen in der Abbildung angedeutet wird, kann es zu einem starken Ungleichgewicht der tatsächlichen Fehlerwahrscheinlichkeiten kommen. Es würden hier viel eher Werte aus π_2 falsch klassifiziert werden als Werte aus π_1.

8.3.4 Die Diskriminanzfunktion von Fisher

Fisher (1938) entwickelte eine lineare Diskriminanzfunktion analog zu (8.14), allerdings war die dahinterstehende Idee eine andere. Er versuchte nämlich, multivariate Beobachtungen auf univariate zu transformieren, sodass die beiden transformierten

Abbildung 8.3: Wahrscheinlichkeiten bei Missklassifikation bei Verwendung der linearen Diskriminanzfunktion



Gruppen möglichst stark voneinander separiert sind. Für eine feste Linearkombination der $\boldsymbol{x}$-Variablen ergeben sich univariate Werte $y_{11}, y_{12}, \ldots, y_{1n_1}$ für die Beobachtungen der ersten Gruppe und Werte $y_{21}, y_{22}, \ldots, y_{2n_2}$ für die Beobachtungen der zweiten Gruppe. Die Trennung der beiden Populationen geschieht dann so, dass die arithmetischen Mittel $\bar{y}_1$ und $\bar{y}_2$ der univariaten y-Werte möglichst stark voneinander abweichen. Diese Differenz wird ausgedrückt in Einheiten der Standardabweichung, und man erhält daher als Kriterium für die Separation

$$\frac{|\bar{y}_1 - \bar{y}_2|}{s_y}$$

mit der gepoolten Varianz der y-Werte

$$s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j} - \bar{y}_1)^2 + \sum_{j=1}^{n_2}(y_{2j} - \bar{y}_2)^2}{n_1 + n_2 - 2} \ .$$

Es wird nun versucht, eine Linearkombination von $\boldsymbol{x}$ zu finden, die eine maximale Separation der Stichprobenmittel $\bar{y}_1$ und $\bar{y}_2$ zulässt.

Satz 8.3.4 *Die Linearkombination*

$$\hat{y} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x} \tag{8.18}$$

maximiert das Verhältnis

$$\frac{(\bar{y}_1 - \bar{y}_2)^2}{s_y^2} = \frac{(\hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_1 - \hat{\boldsymbol{a}}^\top \bar{\boldsymbol{x}}_2)^2}{\hat{\boldsymbol{a}}^\top \boldsymbol{S}_{pooled} \hat{\boldsymbol{a}}} \tag{8.19}$$

über alle möglichen Vektoren $\hat{\boldsymbol{a}}$. $\boldsymbol{S}_{pooled}$ ist dabei analog definiert wie im Zweigruppenfall. Das maximale Verhältnis von (8.19) ist

$$D^2 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \ .$$

Beweis: Mit der Bezeichnung $\boldsymbol{d} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)$ folgt aus der erweiterten Cauchy-Schwarz Ungleichung (Johnson and Wichern, 1998):

$$\max_{\hat{\boldsymbol{a}}} \frac{(\hat{\boldsymbol{a}}^\top \boldsymbol{d})^2}{\hat{\boldsymbol{a}}^\top \boldsymbol{S}_{pooled} \hat{\boldsymbol{a}}} = \boldsymbol{d}^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{d} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) = D^2$$

□

Es ergibt sich somit folgende Klassifikationsregel (vgl. auch (8.14)):

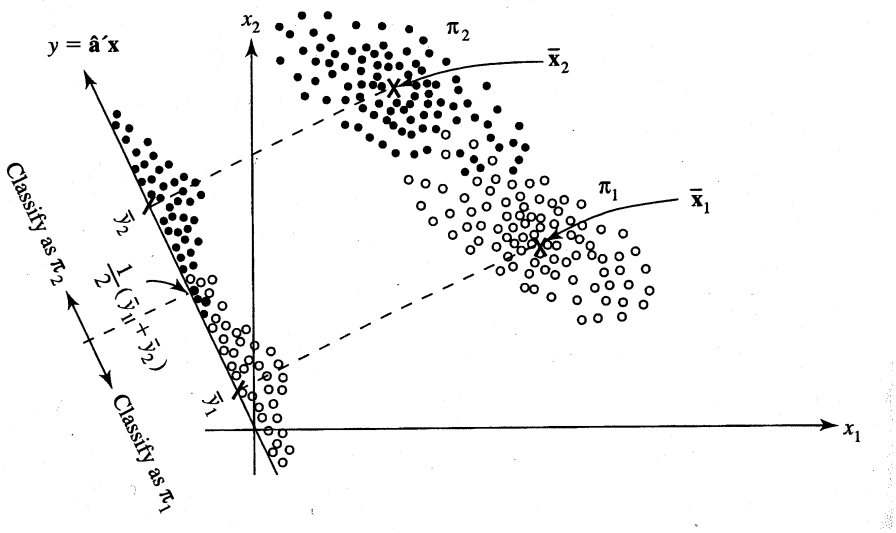
Satz 8.3.5 *Eine Beobachtung $\boldsymbol{x}_0$ wird π_1 zugeteilt, wenn*

$$\hat{y}_0 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x}_0 \geq \hat{m} = \frac{1}{2} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{8.20}$$

Ansonsten wird $\boldsymbol{x}_0$ π_2 zugeordnet.

Diese Klassifikationsregel wird für $p = 2$ in Abbildung 8.4 dargestellt. Die Beobachtungen werden auf eine Gerade projiziert. Die Richtung $\hat{\boldsymbol{a}}$ der Geraden wird so lange variiert, bis die beiden Gruppen maximal separiert sind.

Abbildung 8.4: Schematische Darstellung der Klassifikationsregel von Fisher.



Im Gegensatz zur Regel (8.13) wird bei der Klassifikationsregel von Fisher die Voraussetzung der Normalverteilung beider Populationen nicht benötigt. Allerdings geht man davon aus, dass die Populationen gleiche Kovarianzmatrix haben, nachdem eine gepoolte Schätzung der Kovarianz genommen wird. Man erkennt sofort, dass die lineare Diskriminanzfunktion von Fisher in (8.20) ein spezieller Fall von (8.13) ist. Sind nämlich bei Regel (8.13), die durch Minimierung der erwarteten Kosten bei Missklassifikation hervorgegangen ist, die a-priori Wahrscheinlichkeiten und die erwarteten Kosten bei Missklassifikation gleich, so ergibt sich genau (8.20).

Beispiel 8.3.2 *Wir betrachten nochmals die Hemophilie-Daten. Unter der Voraussetzung von gleichen a-priori Wahrscheinlichkeiten und gleichen Kosten bei Missklassifikation ergab sich in (8.15)*

$$\hat{y} = \hat{\boldsymbol{a}}^\top \boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} \boldsymbol{x} = 37.61 x_1 - 28.92 x_2 \ .$$

Diese Funktion entspricht der linearen Diskriminanzfunktion von Fisher in (8.18) und ermöglicht eine maximale Separation der beiden Gruppen. Die maximale Separation ist gegeben durch

$$\begin{aligned} D^2 &= (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)^\top \boldsymbol{S}_{pooled}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \\ &= (0.2418 \quad -0.0652) \begin{pmatrix} 131.158 & -90.423 \\ -90.423 & 108.147 \end{pmatrix} \begin{pmatrix} 0.2418 \\ -0.0652 \end{pmatrix} \\ &= 10.98 \ . \end{aligned}$$

Das heißt also, dass der Wert $D^2 = 10.98$ die maximale Separation der beiden Gruppen angibt, die durch Linearkombination der multivariaten Beobachtungen ermöglicht wird.

8.4 Klassifikation mehrerer Populationen

In diesem Abschnitt sollen nun die Konzepte für den Fall zweier Populationen erweitert werden auf $g > 2$ Gruppen. Diese Erweiterung ist mathematisch plausibel, allerdings ist bisher wenig bekannt über die Eigenschaften der entsprechenden Klassifikationsregeln der Stichprobe. Abweichungen von Normalverteilung der Gruppen oder verschiedene Kovarianzen könnten die guten theoretischen Eigenschaften dann stark verzerren.

8.4.1 Die Methode zur Minimierung der EKM

Sei $f_i(\boldsymbol{x})$ die Dichte der Beobachtungen von Population π_i mit $i = 1, \ldots, g$. Meistens nimmt man an, dass $f_i(\boldsymbol{x})$ Dichte einer multivariaten Normalverteilung ist, aber dies ist nicht Voraussetzung der nun folgenden Methode.

Die Notation ist angelehnt an den Zweigruppenfall. Demnach bezeichnet p_i die a-priori Wahrscheinlichkeit für die Population π_i $(i = 1, \ldots, g)$. Weiters sei R_k jener Raum von Beobachtungen $\boldsymbol{x}$, dem die Objekte aus π_k zugeordnet werden. $c(k|i)$ sind die Kosten für Missklassifikation, wenn Objekte aus π_i fälschlicherweise π_k zugeordnet werden $(k = 1, \ldots, g)$, wobei für $k = i$ natürlich $c(i|i) = 0$ gilt. Die Wahrscheinlichkeit dieser falschen Zuordnung ist

$$P(k|i) = P(\boldsymbol{X} \in R_k | \pi_i) = \int_{R_k} f_i(\boldsymbol{x}) d\boldsymbol{x} \ . \tag{8.21}$$

Die bedingten erwarteten Kosten bei Missklassifikation von $\boldsymbol{x}$ aus π_1 in $\pi_2, \ldots, \pi_g$ sind analog zu (8.3)

$$EKM(1) = \sum_{k=2}^{g} P(k|1) c(k|1) \ . \tag{8.22}$$

Diese erwarteten Kosten entstehen mit a-priori Wahrscheinlichkeit p_1. Analog werden $EKM(2), \ldots, EKM(g)$ definiert, und man erhält durch Multiplikation mit den a-priori Wahrscheinlichkeiten und Summation aller Beiträge die gesamten erwarteten Kosten bei Missklassifikation (EKM) als

$$EKM = \sum_{i=1}^{g} p_i EKM(i) = \sum_{i=1}^{g} p_i \left(\sum_{\substack{k=1 \\ k \neq i}}^{g} P(k|i)c(k|i) \right) . \tag{8.23}$$

Eine optimale Klassifikationsregel sollte nun solche Bereiche $R_1, \ldots, R_g$ (disjunkte und vollständige Zerlegung des Stichprobenraumes Ω) ergeben, sodass (8.23) minimiert wird.

Satz 8.4.1 *Die Bereiche zur Minimierung von EKM (8.23) sind dadurch gegeben, dass $\boldsymbol{x}$ jener Population π_k $(k = 1, \ldots, g)$ zugeordnet wird, für die der Ausdruck*

$$\sum_{\substack{i=1 \\ k \neq i}}^{g} p_i f_i(\boldsymbol{x}) c(k|i) \tag{8.24}$$

minimal wird.

Beweis: siehe Anderson (1984); vgl. auch (8.4)

Nimmt man an, dass die Kosten bei Missklassifikation gleich sind, so würde $\boldsymbol{x}$ jener Population π_k zugeordnet werden, für die

$$\sum_{\substack{i=1 \\ k \neq i}}^{g} p_i f_i(\boldsymbol{x}) \tag{8.25}$$

minimal wird, da in (8.24) die Kosten 1 gesetzt werden können. Minimierung von (8.25) entspricht aber einer Maximierung des weggelassenen Terms $p_k f_k(\boldsymbol{x})$. Damit vereinfacht sich die Klassifikationsregel zu:

Eine Beobachtung $\boldsymbol{x}$ wird π_k zugeordnet, wenn

$$p_k f_k(\boldsymbol{x}) > p_i f_i(\boldsymbol{x}) \quad \text{für alle } i \neq k . \tag{8.26}$$

Man beachte, dass obige Klassifikationsregeln nur angewandt werden können, wenn die a-priori Wahrscheinlichkeiten, die Missklassifikationskosten und die Dichtefunktionen bekannt sind.

8.4.2 Klassifikation bei Normalverteilung

Als wichtigen Spezialfall betrachten wir multivariat normalverteilte Grundgesamtheiten π_i mit Mittel $\boldsymbol{\mu}_i$ und Kovarianzmatrix $\boldsymbol{\Sigma}_i$,

$$f_i(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}_i|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \right\} \quad \text{für } i = 1, \ldots, g . \tag{8.27}$$

Verwendet man nun Regel (8.26) bzw. den Logarithmus dieser Regel, so ergibt sich: Eine Beobachtung $\boldsymbol{x}$ wird π_k zugeordnet, wenn

$$\ln p_k f_k(\boldsymbol{x}) = \ln p_k - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k) = \max_i \ln p_i f_i(\boldsymbol{x}) \ . \tag{8.28}$$

Die Konstante $(p/2)\ln(2\pi)$ kann in (8.28) weggelassen werden, da sie gleich ist für alle π_i. Wir definieren daher die *quadratischen Diskriminanzwerte* für die i-te Population als

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i) + \ln p_i \quad \text{für } i = 1,\ldots,g \ . \tag{8.29}$$

Somit erhält man folgende Klassifikationsregel:

Eine Beobachtung $\boldsymbol{x}$ wird π_k zugeordnet, wenn gilt:

$$d_k^Q(\boldsymbol{x}) \quad \text{ist am größten von } d_1^Q(\boldsymbol{x}),\ldots,d_g^Q(\boldsymbol{x}) \ . \tag{8.30}$$

Meistens sind die $\boldsymbol{\mu}_i$ und $\boldsymbol{\Sigma}_i$ unbekannt und müssen geschätzt werden. Ist ein "Trainingsset" von korrekt klassifizierten Beobachtungen verfügbar, so können mit den Stichprobengrößen n_i der i-ten Population ($i = 1,\ldots,g$) die Stichprobenmittel $\bar{\boldsymbol{x}}_i$ und Stichproben-Kovarianzmatrizen $\boldsymbol{S}_i$ geschätzt werden. Daraus ergeben sich die geschätzten quadratischen Diskriminanzwerte

$$\hat{d}_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{S}_i| - \frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i)^\top \boldsymbol{S}_i^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_i) + \ln p_i \ , \tag{8.31}$$

mit deren Hilfe Objekte analog zu (8.30) klassifiziert werden.

Eine Vereinfachung ergibt sich, wenn die Kovarianzmatrizen der Populationen gleich sind, also $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ für $i = 1,\ldots,g$. Dann sind die quadratischen Diskriminanzwerte aus (8.29)

$$d_i^Q(\boldsymbol{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \ . \tag{8.32}$$

Nachdem die ersten beiden Terme in (8.32) für alle Populationen gleich sind, können sie weggelassen werden, und man erhält die *linearen Diskriminanzwerte*

$$d_i(\boldsymbol{x}) = \boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_i^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \quad \text{für} \quad i = 1,\ldots,g \ . \tag{8.33}$$

Man erhält eine Schätzung für die linearen Diskriminanzwerte aus der Stichprobe, indem man zunächst eine gepoolte Kovarianzmatrix schätzt durch

$$\boldsymbol{S}_{pooled} = \frac{1}{n_1+\cdots+n_g-g}\Big((n_1-1)\boldsymbol{S}_1 + \cdots + (n_g-1)\boldsymbol{S}_g\Big) \ . \tag{8.34}$$

Damit erhält man Schätzungen

$$\hat{d}_i(\boldsymbol{x}) = \bar{\boldsymbol{x}}_i^\top\boldsymbol{S}_{pooled}^{-1}\boldsymbol{x} - \frac{1}{2}\bar{\boldsymbol{x}}_i^\top\boldsymbol{S}_{pooled}^{-1}\bar{\boldsymbol{x}}_i + \ln p_i \ . \tag{8.35}$$

Man wendet wieder folgende Klassifikationsregel an:

Eine Beobachtung $\boldsymbol{x}$ wird π_k zugeordnet, wenn gilt:

$$\hat{d}_k(\boldsymbol{x}) \quad \text{ist am größten von} \quad \hat{d}_1(\boldsymbol{x}), \ldots, \hat{d}_g(\boldsymbol{x}) \ . \tag{8.36}$$

Bemerkung: Der Ausdruck (8.33) ist eine lineare Funktion von $\boldsymbol{x}$. Man könnte aber eine analoge Regel erhalten, indem man in (8.29) den ersten Term ignoriert, der ja für den Fall gleicher Kovarianzmatrizen für alle Populationen gleich ist. Mit den entsprechenden geschätzten Werten erhält man eine quadratische Distanz

$$D_i^2(\boldsymbol{x}) = (\boldsymbol{x} - \bar{\boldsymbol{x}}_i)^\top \boldsymbol{S}_{pooled}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}_i) \ , \tag{8.37}$$

und die Klassifikationsregel lautet:

Weise $\boldsymbol{x}$ jener Population π_i zu, für die $-1/2 \ D_i^2(\boldsymbol{x}) + \ln p_i$ am größten ist. (8.38)

Diese Regel ist analog zur Regel (8.36), sie weist $\boldsymbol{x}$ jener Population zu, die "am nächsten" ist, wobei das Distanzmaß mit $\ln p_i$ belastet wird. Sind die a-priori Wahrscheinlichkeiten unbekannt, könnte man sie schätzen durch $p_i = 1/g$.

Beispiel 8.4.1 *Wir betrachten die bekannten Daten von Ruspini (1970) aus Tabelle 8.1. Die Daten bestehen aus 75 Punkten und wurden ursprünglich von Ruspini verwendet, um Fuzzy-Cluster Methoden zu illustrieren. Die Daten sind dargestellt in Abbildung 8.5, sie enthalten 4 Gruppen, die auch mit Zahlen dargestellt sind. Die Einteilung in die Gruppen erfolgte hier mit k-means Clusterung, und diese entspricht auch einer visuellen Einteilung. Die Anzahlen der Objekte in den einzelnen Gruppen sind*

$$n_1 = 17 \qquad n_2 = 20 \qquad n_3 = 23 \qquad n_4 = 15 \ .$$

Die Gruppenmittel sind mit + eingezeichnet und betragen

$$\bar{\boldsymbol{x}}_1 = \begin{pmatrix} 98.18 \\ 114.88 \end{pmatrix} \qquad \bar{\boldsymbol{x}}_2 = \begin{pmatrix} 20.15 \\ 64.95 \end{pmatrix} \qquad \bar{\boldsymbol{x}}_3 = \begin{pmatrix} 43.91 \\ 146.04 \end{pmatrix} \qquad \bar{\boldsymbol{x}}_4 = \begin{pmatrix} 68.93 \\ 19.40 \end{pmatrix} \ .$$

Die Kovarianzmatrizen der einzelnen Gruppen sind

$$\boldsymbol{S}_1 = \begin{pmatrix} 164.53 & 64.27 \\ 64.27 & 120.36 \end{pmatrix} \qquad \boldsymbol{S}_2 = \begin{pmatrix} 92.661 & -5.045 \\ -5.045 & 101.524 \end{pmatrix}$$
$$\boldsymbol{S}_3 = \begin{pmatrix} 91.81 & -23.27 \\ -23.27 & 52.59 \end{pmatrix} \qquad \boldsymbol{S}_4 = \begin{pmatrix} 51.2095 & 0.2429 \\ 0.2429 & 52.8286 \end{pmatrix}$$

Daraus ergibt sich die gepoolte Kovarianzmatrix bzw. deren Inverse als

$$\boldsymbol{S}_{pooled} = \begin{pmatrix} 100.419 & 5.972 \\ 5.972 & 81.004 \end{pmatrix} \qquad \boldsymbol{S}_{pooled}^{-1} = \begin{pmatrix} 0.0100021 & -0.0007374 \\ -0.0007374 & 0.0123995 \end{pmatrix}$$

Wir möchten nun eine neue Beobachtung $\boldsymbol{x}_0 = (55, 80)^\top$, die auch in Abbildung 8.5 mit $\times$ eingezeichnet ist, klassifizieren. Wir nehmen dazu an, dass die a-priori

Tabelle 8.1: Ruspini Daten.

x	y	x	y	x	y
4	53	41	150	98	124
5	63	38	145	99	119
10	59	38	143	99	128
9	77	32	143	101	115
13	49	34	141	108	111
13	69	44	156	110	111
12	88	44	149	108	116
15	75	44	143	111	126
18	61	46	142	115	117
19	65	47	149	117	115
22	74	49	152	70	4
27	72	50	142	77	12
28	76	53	144	83	21
24	58	52	152	61	15
27	55	55	155	69	15
28	60	54	124	78	16
30	52	60	136	66	18
31	60	63	139	58	13
32	61	86	132	64	20
36	72	85	115	69	21
28	147	85	96	66	23
32	149	78	94	61	25
35	153	74	96	76	27
33	154	97	122	72	31
38	151	98	116	64	30

Wahrscheinlichkeiten gleich sind, also $p_i = 0.25$. Wir schätzen die linearen Diskriminanzwerte anhand (8.35) und erhalten

$$\hat{d}_1 = 34.42 \qquad \hat{d}_2 = 43.08 \qquad \hat{d}_3 = 21.98 \qquad \hat{d}_4 = 25.81 \ .$$

Mit der Klassifikationsregel (8.36), die die neue Beobachtung jener Population zuordnet, die am nächsten liegt, wird $\boldsymbol{x}_0$ der Population π_2 zugeordnet, weil 43.08 der größte Diskriminanzwert ist.

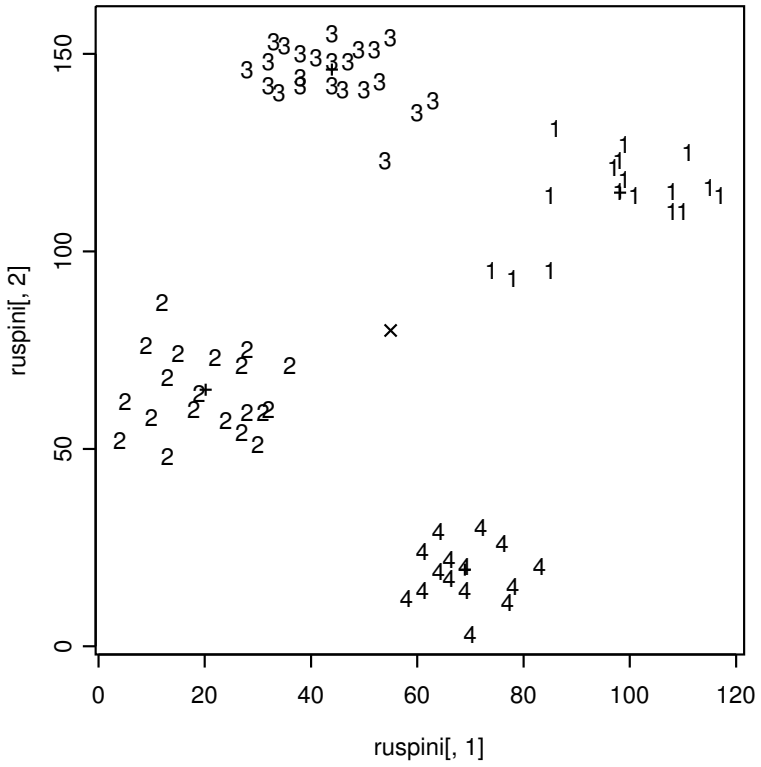
Wenn man nicht davon ausgehen möchte, dass die Kovarianzmatrizen als gleich anzunehmen sind, verwendet man die (geschätzten) quadratischen Diskriminanzwerte aus (8.31) und erhält

$$\hat{d}_1^Q = -13.59 \qquad \hat{d}_2^Q = -13.93 \qquad \hat{d}_3^Q = -49.06 \qquad \hat{d}_4^Q = -42.07 \ .$$

Der größte Wert ist -13.59, und daher wird $\boldsymbol{x}_0$ der Population π_1 zugeordnet. Die Entscheidung fällt aber hier denkbar knapp aus.

Ähnlich zum Zweigruppenfall kann auch hier die "jackknife"-Prozedur von Lachenbruch und Mickey (1968) zur Schätzung des Fehleranteils angewandt werden.

Abbildung 8.5: Ruspini Daten: Erläuterungen siehe Text.



Wenn $\bar{n}_i$ die Anzahl der missklassifizierten Objekte der i-ten Gruppe bezeichnet $(i = 1, \ldots, g)$, dann ist

$$\frac{\sum_{i=1}^{g} \bar{n}_i}{\sum_{i=1}^{g} n_i}$$

die geschätze Fehlerrate.

Bemerkung: Auch die Methode von Fisher kann auf den Fall mehrerer Gruppen übertragen werden. Die Diskriminanzfunktion ist ein Verhältnis von Variation zwischen den Gruppen zu Variation innerhalb der Gruppen (ähnlich zur Varianzanalyse). Details findet man z.B. in Johnson und Wichern (1998).

8.4.3 Diskriminanzanalyse nach Fisher für den Mehrgruppenfall

Die Diskriminanzanalyse nach Fisher für den Fall $g = 2$ kann erweitert werden auf den Mehrgruppenfall ($g > 2$) siehe Rao (1948). Zu diesem Zweck betrachten wir wieder univariate Projektionen der Form $y = \mathbf{a}^T\mathbf{x}$, mit $\mathbf{a} \in \Re^p$ und $\mathbf{a} \neq \mathbf{0}$, aber diesmal wird eine Projektionsrichtung $\mathbf{a}$ nicht ausreichend sein für die Beschreibung der Lösung.

Sei also $\mathbf{a} \neq \mathbf{0}$ eine gesuchte Projektionsrichtung. Der Erwartungswert für Population π_i (für $i = 1, \ldots, g$) der Zufallsvariable $y = \mathbf{a}^T\mathbf{x}$ ist dann

$$\mu_{i,y} = E(y|\mathbf{x} \in \pi_i) = \mathbf{a}^T E(\mathbf{x}|\mathbf{x} \in \pi_i) = \mathbf{a}^T\boldsymbol{\mu}_i$$

und die Varianz ist

$$\sigma^2_{i,y} = \mathrm{Var}(y|\mathbf{x} \in \pi_i) = \mathbf{a}^T\mathrm{Cov}(\mathbf{x}|\mathbf{x} \in \pi_i)\mathbf{a} = \mathbf{a}^T\boldsymbol{\Sigma}_i\mathbf{a} \ .$$

Das gesamte gewichtete Mittel der Populationen wird bezeichnet mit $\bar{\boldsymbol{\mu}} = \sum_{i=1}^g p_i\boldsymbol{\mu}_i$, und die entsprechende Projektion ins Univariate mit $\bar{\mu}_y = \mathbf{a}^T\bar{\boldsymbol{\mu}}$.

Wir treffen nun die Annahme, dass die Kovarianzen aller Gruppen gleich sind, also $\boldsymbol{\Sigma}_1 = \ldots = \boldsymbol{\Sigma}_g = \boldsymbol{\Sigma}$. Dann gilt auch für obige Varianz

$$\sigma^2_{1,y} = \ldots = \sigma^2_{g,y} = \sigma^2_y = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a} \ .$$

Im Fisher-Zweigruppenfall wurde ein Ausdruck $(\bar{y}_1 - \bar{y}_2)^2/s^2_y$ maximiert. In unserer Notation mit Zufallsgrößen entspricht das $(\mu_{1,y} - \mu_{2,y})^2/\sigma^2_y$. Wir betrachten jetzt zusätzlich a-priori Wahrscheinlichkeiten, und das gewichtete Mittel ist

$$\bar{\mu}_y = p_1\mu_{1,y} + p_2\mu_{2,y} = \mathbf{a}^T(p_1\boldsymbol{\mu}_1 + p_2\boldsymbol{\mu}_2) \ .$$

Es ist dann unschwer zu sehen, dass

$$p_1(\mu_{1,y} - \bar{\mu}_y)^2 + p_2(\mu_{2,y} - \bar{\mu}_y)^2 = p_1p_2(\mu_{1,y} - \mu_{2,y})^2 \ . \tag{8.39}$$

Letzterer Ausdruck (8.39) sollte also mit der Fisher Regel (bezogen auf die Varianz) maximiert werden, und dieser Ausdruck beschreibt die gewichtete Summe der quadrierten Abstände der Gruppenmittelwerte zum Gesamtmittel.

Die Verallgemeinerung von (8.39) auf den Mehrgruppenfall ist dann unmittelbar einsichtig; es soll nun

$$\frac{\sum_{i=1}^g p_i(\mu_{i,y} - \bar{\mu}_y)^2}{\sigma^2_y} \tag{8.40}$$

maximiert werden. Der Nenner ist laut oben $\sigma^2_y = \mathbf{a}^T\boldsymbol{\Sigma}\mathbf{a}$. Wenn die Gruppen-Kovarianzen nicht gleich sind, wird die resultierende Klassifikationsregel nicht mehr optimal sein. $\boldsymbol{\Sigma}$ würde dann am besten durch eine gepoolte Version ersetzt werden, also durch

$$\mathbf{W} = \sum_{i=1}^g p_i\boldsymbol{\Sigma}_i \ .$$

Die Matrix $\mathbf{W}$ beschreibt die *Variation innerhalb der Gruppen.*

Der Zähler von (8.40) kann dargestellt werden als

$$\sum_{i=1}^g p_i(\mu_{i,y} - \bar{\mu}_y)^2 = \sum_{i=1}^g p_i(\mathbf{a}^T(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}}))^2 = \mathbf{a}^T\mathbf{B}\mathbf{a} \ ,$$

mit der Matrix

$$\mathbf{B} = \sum_{i=1}^g p_j(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_i - \bar{\boldsymbol{\mu}})^T \ ,$$

die die *Variation zwischen den Gruppen* beschreibt.

Insgesamt kann also das Maximierungsproblem (8.40) ausgedrückt werden als

$$\frac{\mathbf{a}^T\mathbf{B}\mathbf{a}}{\mathbf{a}^T\mathbf{W}\mathbf{a}} \quad \text{für } \mathbf{a} \in \Re^p,\ \mathbf{a} \neq \mathbf{0}\ . \tag{8.41}$$

Satz 8.4.2 *Die Lösung des Maximierungsproblems (8.41) ist gegeben durch die Eigenvektoren* $\mathbf{a}_1, \ldots, \mathbf{a}_l$ *der Matrix* $\mathbf{W}^{-1}\mathbf{B}$*, die so skaliert werden, dass* $\mathbf{a}_j^T\mathbf{W}\mathbf{a}_j = 1$ *für* $i = j, \ldots, l$*. Die Anzahl* l *der strikt positiven Eigenwerte der Eigenwertzerlegung von* $\mathbf{W}^{-1}\mathbf{B}$ *ist dabei* $l \leq \min(g-1, p)$*.*

Beweis: Problem (8.41) ist invariant bezüglich einer Umskalierung von $\mathbf{a}$. D.h., für jedes $\tilde{\mathbf{a}} = \alpha\mathbf{a}$, mit $\alpha \neq 0$, erhält man das gleiche Maximum. Daher kann man den Nenner so skalieren, dass $\mathbf{a}^T\mathbf{W}\mathbf{a} = 1$ gilt. Dies vereinfacht das Optimierungsproblem (8.41) zu

$$\min_{\mathbf{a}}(-\mathbf{a}^T\mathbf{B}\mathbf{a}) \qquad \text{sodass} \quad \mathbf{a}^T\mathbf{W}\mathbf{a} = 1\ .$$

Die Minimierung erfolgt mittels Langrange'schen Ausdruck

$$\phi = -\frac{1}{2}\mathbf{a}^T\mathbf{B}\mathbf{a} + \frac{1}{2}\lambda(\mathbf{a}^T\mathbf{W}\mathbf{a} - 1)$$

mit dem Lagrange-Multiplikator λ. Die Terme $1/2$ sind bequem wenn wir nun die Ableitung bilden:

$$\frac{\partial\phi}{\partial\mathbf{a}} = -\mathbf{B}\mathbf{a} + \lambda\mathbf{W}\mathbf{a}$$

Nullsetzen der Ableitung ergibt

$$\mathbf{B}\mathbf{a} = \mathbf{W}\lambda\mathbf{a} \qquad \text{bzw.} \qquad \mathbf{W}^{-1}\mathbf{B}\mathbf{a} = \lambda\mathbf{a}\ .$$

Man erhält also wieder ein Eigenwertproblem, und die Lösungen für $\mathbf{a}$ sind die Eigenvektoren $\mathbf{a}_1, \ldots, \mathbf{a}_l$ von $\mathbf{W}^{-1}\mathbf{B}$ zu den Eigenwerten $\lambda_1, \ldots, \lambda_l$. Die Eigenwerte sind absteigend sortiert, und dann stellt $\mathbf{a}_1$ jene Richtung dar, entlang der die Gruppenmittel am besten separiert sind. Somit ist die Reihenfolge der erhaltenen Richtungen wichtig, und mit einer Projektion auf die erste Richtung erhält man – ähnlich wie bei Hauptkomponentenanalyse – die informativste Dimensionsreduktion der gesamten Information, mit dem Unterschied, dass man nun zusätzlich zur multivariaten Information der Beobachtungen auch Klasseninformation hat.
Die Anzahl $l \leq \min(g-1, p)$ erklärt sich unmittelbar aus den Rängen von $\mathbf{W}$, nämlich maximal p, und $\mathbf{B}$ (maximal $g-1$).
Man bemerke, dass $\mathbf{W}^{-1}\mathbf{B}$ nicht notwendigerweise symmetrisch ist, und somit können Eigenwerte- und vektoren Imaginärteile haben. Dies kann leicht umgangen werden, indem man die symmetrische Matrix $\mathbf{B}$ darstellt als $\mathbf{B} = \mathbf{B}^{1/2}\mathbf{B}^{1/2}$. Mit der Definition $\mathbf{b} = \mathbf{B}^{1/2}\mathbf{a}$, bzw. gleichwertig $\mathbf{a} = \mathbf{B}^{-1/2}\mathbf{b}$ ist obiges Eigenwertproblem $\mathbf{B}^{1/2}\mathbf{W}^{-1}\mathbf{B}^{1/2}\mathbf{b} = \lambda\mathbf{b}$, und man hat somit ein Eigenwertproblem der symmetrischen Matrix $\mathbf{B}^{1/2}\mathbf{W}^{-1}\mathbf{B}^{1/2}$. □

Man kann nun die *Fisher Diskriminanzfunktionen* definieren als $y_j = \mathbf{a}_j^T\mathbf{x}$, für $j = 1, \ldots, l$, die somit die Projektionen der Zufallsgröße $\mathbf{x}$ auf die Richtung $\mathbf{a}_j$ darstellen. Wenn nun konkrete Daten vorliegen, ist eine Visualisierung der beiden ersten

Diskriminanzfunktionen besonders interessant, weil dies jene Projektion der Daten darstellt, in der die Gruppenmittel am besten separiert erscheinen. Man bemerke, dass im Fall von $g = 3$ Gruppen $l \leq 2$ ist, egal ob p groß ist oder nicht.

Schließlich möchte man auch noch eine Klassifikationsregel erhalten. Dazu betrachtet man die *Fisher Diskriminanzwerte*

$$d_i^F(\mathbf{x}) = \sum_{j=1}^{l} (y_j - \mu_{i,y_j})^2 - 2 \log p_i \tag{8.42}$$

für $i = 1, \ldots, g$. Hier ist $\mu_{i,y_j} = \mathbf{a}_j^T \boldsymbol{\mu}_i$, und man hat somit ein Maß für Abweichung von $\mathbf{x}$ zum i-ten Gruppenmittel im Diskriminanzraum, adjustiert mit der a-priori Wahrscheinlichkeit (analog zu früher). Man bemerke, dass hier im Diskriminanzraum das Distanzmaß einfach die Euklidische Distanz ist. Eine neue Beobachtung $\mathbf{x}$ wird dann der Population π_k zugeordnet, wenn $d_k^F(\mathbf{x})$ der kleinste (!) Wert von allen Werten der Gruppen $d_1^F(\mathbf{x}), \ldots, d_g^F(\mathbf{x})$ ist.

Durch Anordnen der Eigenvektoren $\mathbf{a}_1, \ldots, \mathbf{a}_l$ in den Spalten der Matrix $\mathbf{A}$ kann man die *Fisher Diskriminanzwerte* auch schreiben als

$$d_i^F(\mathbf{x}) = \sum_{j=1}^{l} (\mathbf{a}_j^T (\mathbf{x} - \boldsymbol{\mu}_i))^2 - 2 \log p_i = (\mathbf{x} - \boldsymbol{\mu}_i)^T \mathbf{A}\mathbf{A}^T (\mathbf{x} - \boldsymbol{\mu}_i) - 2 \log p_i \ ,$$

was einer (quadrierten) Mahalanobis Distanz im ursprünglichen Raum entspricht.

Literatur

T.W. Anderson. *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, New York, 1984.

B.N. Bouma, et al. Evaluation of the detection rate of Hemophilia carriers. *Statistical Methods for Clinical Decision Making*, 7(2):339–350, 1975.

R.A. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376-386, 1938.

D.J. Hand. *Discrimination and Classification.* John Wiley & Sons, New York, 1981.

C.J. Huberty. *Applied Discriminant Analysis.* John Wiley & Sons, New York, 1994.

R. Johnson and D. Wichern. *Applied Multivariate Statistical Analysis.* Prentice-Hall, London, 4th edition, 1998.

P.A. Lachenbruch and M.R. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10(1):1–11, 1968.

C.R. Rao. The utilization to multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society, Series B*, 10:159–203, 1948.

E.H. Ruspini. Numerical methods for fuzzy clustering. *Inform. Sci.*, 2:319-350, 1970.

G.A.F. Seber. *Multivariate observations.* John Wiley & Sons, New York, 1984.

Kapitel 9

Projection Pursuit

9.1 Einleitung

9.1.1 Hintergrund

Der Ausdruck *Projection Pursuit* wurde von Friedman und Tukey (1974) eingeführt und bezeichnet eine Methode zur explorativen Analyse multivariater Daten. Mit dieser Methode werden „interessante“ Linearprojektionen der multivariaten Daten auf eine Gerade oder eine Ebene gesucht.

Für Projektionen vom zweidimensionalen in den eindimensionalen Raum ist es möglich, im wesentlichen alle solchen Projektionen zu untersuchen und dann jene herauszusuchen, die von Interesse sind. Im höherdimensionalen Raum ist das nicht mehr durchführbar, für solche Daten muss diese „manuelle“ Vorgangsweise automatisiert werden. Projection Pursuit (PP) ist eine Methode, die interessante Projektionen durch lokale Optimierung über alle Projektionsrichtungen auswählt. Das zugehörige Optimierungskriterium wird mit *Interessantheitsindex* bezeichnet. In der Praxis werden Projektionen in den ein- bzw. zweidimensionalen Raum durchgeführt, da bereits dreidimensionale Punktwolken graphisch schwer darstellbar sind.

Viele klassische multivariate Methoden sind (vom Optimierungskriterium betrachtet) Spezialfälle von PP. Ein Beispiel ist die Hauptkomponentenanalyse, bei der der Interessantheitsindex der Anteil an der Gesamtvarianz ist, der durch die projizierten Daten wiedergegeben wird. Ein anderes Beispiel ist die Quartimax-Methode in der Faktorenanalyse.

9.1.2 Definitionen und Notation

Sei $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ ein Zufallsvektor im $\mathbb{R}^p$. Eine Linearprojektion vom $\mathbb{R}^p$ in den $\mathbb{R}^k$ ist eine lineare Abbildung $\boldsymbol{A}$, wobei $\boldsymbol{A}$ eine $(p \times k)$-Matrix vom Rang k darstellt. Der projizierte Zufallsvektor $\boldsymbol{y}$ ist dann

$$\boldsymbol{y} = \boldsymbol{A}^\top \boldsymbol{x} \quad , \tag{9.1}$$

wobei $\boldsymbol{y} \in \mathbb{R}^k$. Wir sprechen von einer *orthonormalen* Projektion, wenn die Spaltenvektoren von $\boldsymbol{A}$ zueinander orthogonal sind und auf Länge Eins normiert sind, d.h. $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{I}$.

Sei F die Verteilung der p-dimensionalen Zufallsvariable $\boldsymbol{x}$, dann ist $\boldsymbol{y} = \boldsymbol{A}^\top \boldsymbol{x}$ eine k-dimensionale Zufallsvariable mit der Verteilung F_A. Für $k = 1$ ist $\boldsymbol{A}$ nur mehr ein Spaltenvektor, um den Unterschied hervorzuheben schreiben wir daher $\boldsymbol{a}$. Die Verteilung von $\boldsymbol{y}$ wird in diesem Fall mit F_a bezeichnet.

PP sucht nach einer Projektionsmatrix $\boldsymbol{A}$, die eine bestimmte „Interessantheitsfunktion“ $Q(F_A)$ maximiert. Da anzunehmen ist, dass mehrere verschiedene Projektionen interessante Ergebnisse liefern, ist man nicht nur am absoluten Extremum, sondern auch an lokalen Extrema interessiert. Q ist ein Funktional im Raum der Verteilungen in $\mathbb{R}^k$. Um die Notation zu vereinfachen, wird im folgenden anstelle von $Q(F_A)$ einfach $Q(\boldsymbol{A}^\top \boldsymbol{x})$ geschrieben.

Für eine feste und bekannte Datenmatrix $\boldsymbol{X}$ ist $Q(\boldsymbol{A}^\top \boldsymbol{X})$ eine Funktion $I(\boldsymbol{A})$ der Projektionsrichtung $\boldsymbol{A}$. Diese Funktion $I(\boldsymbol{A})$ wird mit *Projektionsindex* bezeichnet.

Es werden nur Funktionen Q berücksichtigt, die *affin invariant* sind, d.h. die durch beliebige reguläre Transformationen

$$Q(\boldsymbol{A_1}^\top \boldsymbol{x} + \boldsymbol{b_1}) = Q(\boldsymbol{x}) \tag{9.2}$$

die Ausgangsdaten $\boldsymbol{x}$ unverändert lassen.

Um die späteren Berechnungen zu vereinfachen, wird eine Transformation ausgeübt, welche die Dimension reduziert und die Daten skaliert. Die Kovarianzmatrix wird durch Spektralzerlegung auf die Form

$$Cov(\boldsymbol{x}) = \boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^\top \tag{9.3}$$

gebracht, wobei $\boldsymbol{\Gamma}$ die orthonormale Matrix der Eigenvektoren und $\boldsymbol{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_p)$ die $(p \times p)$-Diagonalmatrix der Eigenwerte von $\boldsymbol{\Sigma}$ ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$) ist. Ist q der Rang von $\boldsymbol{\Sigma}$, dann sind die Komponenten der standardisierten Zufallsvektoren $\boldsymbol{z}$ durch die Transformation

$$z_j = \lambda_j^{-1/2} \sum_{i=1}^{p} \gamma_{ij}(x_i - E(x_i)) \qquad \text{für } 1 \leq j \leq q \tag{9.4}$$

gegeben. γ_{ij} sind die Elemente der Matrix $\boldsymbol{\Gamma}$. Durch diese Transformation ist $E(\boldsymbol{z}) = \boldsymbol{0}$ und $Cov(\boldsymbol{z}) = \boldsymbol{I}$.

9.2 Der Projektionsindex von Friedman

9.2.1 Einführung

Der Projektionsindex misst die „Interessantheit“ einer Projektion. Die Frage, ob eine Projektion interessant ist, hängt von der Anwendung ab. Möchte man in den Daten eine Struktur finden, die nicht durch die Korrelationsstruktur erkennbar ist, so erscheint die Normalverteilung einer Projektion am wenigsten von Interesse zu sein. Diese hypothetische Annahme kann auch durch den Zentralen Grenzverteilungssatz begründet werden, der aussagt, dass die Vereinigung von Verteilungen die Normalverteilung approximiert. Möchte man also Projektionen finden, die eine starke Struktur der Daten aufzeigen, muss der Projektionsindex die Abweichung von der Normalverteilung messen.

9.2.2 Ein- und zweidimensionale Projektion

Friedman (1987) entwickelte einen Projektionsindex, der Abweichungen von der Normalverteilung im Zentrum der Verteilung mehr gewichtet als in den Schwänzen. Die grundlegende Idee ist eine Skalentransformation der projizierten Daten durch die kummulative Verteilungsfunktion. Die transformierte Verteilung wird dann mit der Gleichverteilung, die sich durch Anwendung der selben Transformation auf die Normalverteilung ergibt, verglichen.

Im eindimensionalen Fall von Projection Pursuit sucht man eine Linearkombination

$$y = \boldsymbol{a}^\top \boldsymbol{z} \quad , \tag{9.5}$$

sodass die Wahrscheinlichkeitsverteilung $p_a(y)$ von y relativ starke Struktur aufweist. Die oben beschriebene Transformation ist

$$v = 2\ \Phi(y) - 1 \quad , \tag{9.6}$$

wobei $\Phi(y)$ die kummulative Standardnormalverteilungsfunktion

$$\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-\frac{t^2}{2}} dt \tag{9.7}$$

ist. Durch diese Transformation nimmt v Werte im Interval $(-1, 1)$ an. Ist y standardnormalverteilt, dann ist v gleichverteilt auf diesem Intervall. Ein Maß für die Abweichung von der Gleichverteilung von v entspricht also einem Maß für die Abweichung von der Normalverteilung von y.

In Abbildung 9.1 ist anhand von Beispielen die Funktionsweise dieses Algorithmus illustriert. Die oberen fünf Bilder gehen von der Standardnormalverteilung aus, die nächsten basieren auf der χ^2-Verteilung mit 2 Freiheitsgraden, und die unteren fünf Bilder zeigen, wie der Algorithmus eine Bimodalverteilung (Komposition zweier Normalverteilungen) transformiert.

Die Dichtefunktion von v ist

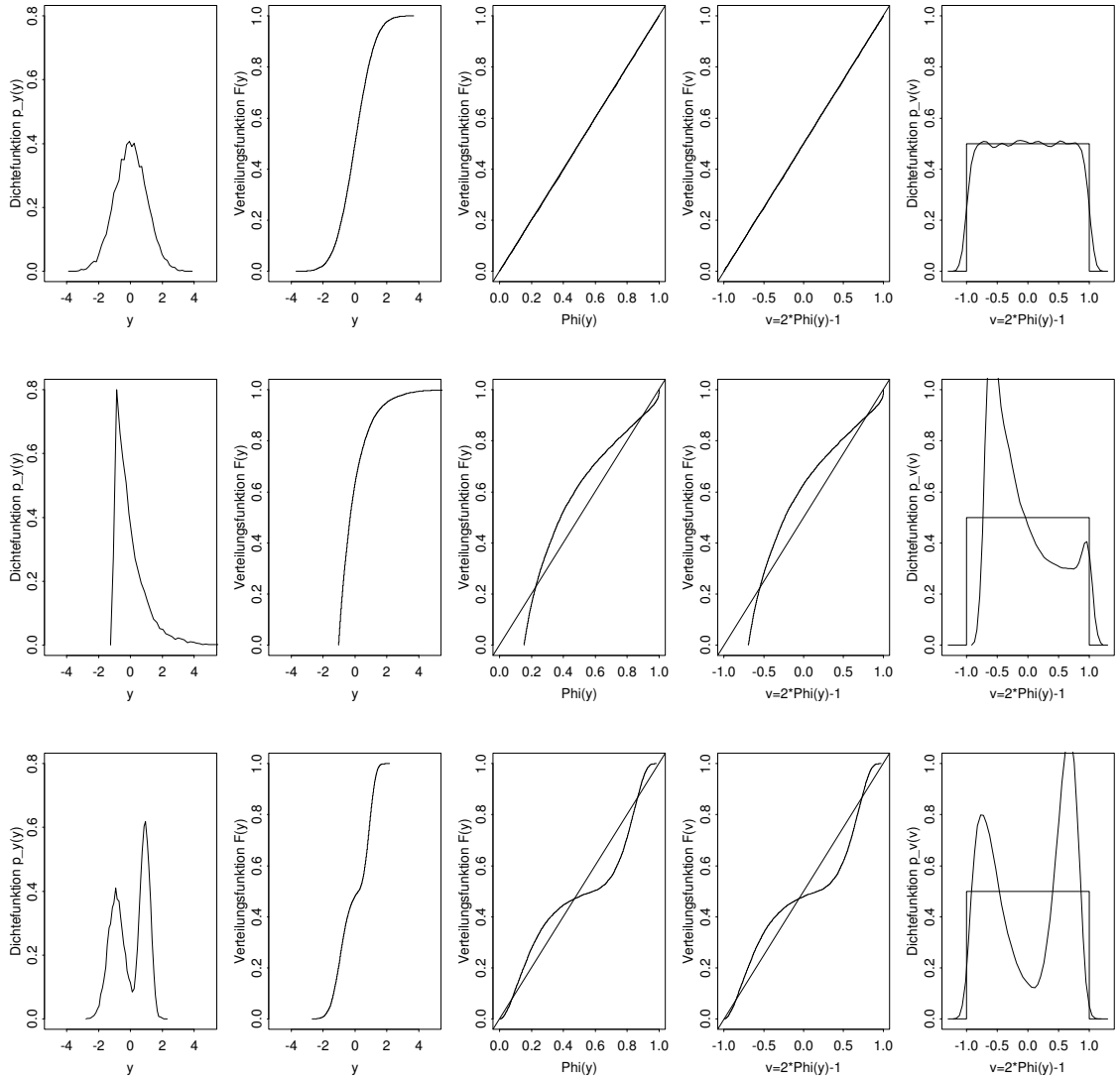
$$p_v(v) = \frac{p_a[\Phi^{-1}(\frac{v+1}{2})]}{2\ \phi[\Phi^{-1}(\frac{v+1}{2})]} \quad . \tag{9.8}$$

$\phi(y)$ ist die Dichte der Standardnormalverteilung. Als Maß für die Abweichung von der Gleichverteilung nimmt Friedman das Integral über das Quadrat der Differenz von $p_v(v)$ und der Dichte der Gleichverteilung $p_u(v) = \frac{1}{2}$ über das Intervall $[-1, 1]$:

$$\int_{-1}^{1} \left(p_v(v) - \frac{1}{2} \right)^2 dv = \int_{-1}^{1} p_v^2(v) dv - \frac{1}{2} \quad . \tag{9.9}$$

Der Projektionsindex $I(\boldsymbol{a})$ sollte schnell berechenbar sein, sein Wert sollte stetig sein, sodass die ersten Ableitungen bezüglich der Parameter überall existieren. Die Ableitungen sollten ebenfalls schnell berechenbar sein. Diese Forderungen machen es plausibel, dass ein Projektionsindex basierend auf Polynomen am geeignetsten ist.

Abbildung 9.1: Illustration der Funktionsweise des Projektionsindex von Friedman



Die Dichte $p_v(v)$ von v wird durch Legendre-Polynome

$$\int_{-1}^{1} p_v^2(v)dv - \frac{1}{2} = \int_{-1}^{1} \left(\sum_{j=0}^{\infty} c_j P_j(v) \right) \; p_v(v)dv - \frac{1}{2} \tag{9.10}$$

approximiert, wobei die Legendre-Polynome durch

$$\begin{aligned} P_0(v) &= 1 \ , \\ P_1(v) &= v \ , \\ P_j(v) &= \frac{(2j-1)vP_{j-1}(v) - (j-1)P_{j-2}(v)}{j} \qquad \text{für } j \geq 2 \end{aligned}$$

definiert werden. Die Koeffizienten c_j sind durch

$$c_j = \frac{2j+1}{2} \int_{-1}^{1} P_j(v) p_v(v) dv = \frac{2j+1}{2} \; E_v[P_j(v)] \tag{9.11}$$

gegeben. Als Approximation für Gleichung (9.9) erhält man somit

$$\int_{-1}^{1} p_v^2(v) dv - \frac{1}{2} = \sum_{j=1}^{\infty} \frac{2j+1}{2} \; E_v^2[P_j(v)] - \frac{1}{2} \; . \tag{9.12}$$

Für eine Gleichverteilung im Intervall $(-1, 1)$ gilt $E[P_j(v)] = 0$ für $j > 0$.

Den Projektionsindex erhält man, indem die Summe in Gleichung (9.12) bei der Ordnung J abgeschnitten wird, d.h.

$$I(\boldsymbol{a}) = \sum_{j=1}^{J} \frac{2j+1}{2} \; E_v^2[P_j(v)] \; . \tag{9.13}$$

Für eine effiziente Optimierung ist es wichtig, dass die Ableitungen der Zielfunktion explizit berechenbar sind. Der Projektionsindex muss unter der Bedingung $\boldsymbol{a}^\top \boldsymbol{a} = 1$ in Bezug auf alle q Komponenten von $\boldsymbol{a}$ maximiert werden. Die Ableitungen von $I(\boldsymbol{a})$ mit $1 \leq k \leq q$ sind

$$\frac{\partial I}{\partial a_k} = \frac{2}{\sqrt{2\pi}} \sum_{j=1}^{J} (2j+1) \; E[P_j(v)] \; E\left[P_j'(v) e^{-\frac{y^2}{2}} (z_k - a_k y) \right] \tag{9.14}$$

mit

$$\begin{aligned} P_1'(v) &= 1 \; , \\ P_j'(v) &= v P_{j-1}'(v) + j P_{j-1}(v) \qquad \text{für } j \geq 1 \; . \end{aligned}$$

Für eine Datenmatrix mit n Beobachtungen,

$$\boldsymbol{Z} = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1q} \\ z_{21} & z_{22} & \dots & z_{2q} \\ \vdots & \vdots & & \vdots \\ z_{n1} & z_{n2} & \dots & z_{nq} \end{pmatrix} \quad ,$$

werden die Erwartungswerte in (9.13) durch die entsprechenden Stichprobenmittel geschätzt. Mit den Substitutionen (9.5) und (9.6) erhält man den geschätzten Projektionsindex

$$\hat{I}(\boldsymbol{a}) = \sum_{j=1}^{J} \frac{2j+1}{2} \left[\frac{1}{n} \sum_{i=1}^{n} P_j \left(2\Phi(\boldsymbol{a}^\top \boldsymbol{z_{i\cdot}}) - 1 \right) \right]^2 \quad , \tag{9.15}$$

wobei $\boldsymbol{z_{i\cdot}}$ die i-te Zeile (Spaltenvektor) von $\boldsymbol{Z}$ bezeichnet.

Der Projektionsindex für die Projektion auf eine Ebene wird analog zum eindimensionalen Fall definiert. Es werden zwei Linearkombinationen

$$y_1 = \boldsymbol{a}^\top \boldsymbol{z} \ , \qquad y_2 = \boldsymbol{b}^\top \boldsymbol{z} \ , \tag{9.16}$$

gesucht, sodass die gemeinsame Wahrscheinlichkeitsdichte $p_{ab}(y_1, y_2)$ von y_1 und y_2 relativ starke Struktur aufweist. Da man an nichtlinearen Strukturen interessiert ist, werden die Linearkombinationen y_1 und y_2 als unkorreliert vorausgesetzt. Weil die Daten skaliert wurden, ist diese Bedingung äquivalent dazu, dass $\boldsymbol{a}$ und $\boldsymbol{b}$ orthogonal sind, d.h. $\boldsymbol{a}^\top \boldsymbol{b} = 0$. Analog zum eindimensionalen Fall sind die weiteren Bedingungen $\boldsymbol{a}^\top \boldsymbol{a} = \boldsymbol{b}^\top \boldsymbol{b} = 1$.

Die bivariate Normalverteilung wird als die am wenigsten interessierende Verteilung betrachtet. Abweichungen von der Normalverteilung werden im Zentrum stärker gewichtet als in den Schwänzen, man führt daher wieder eine Transformation

$$v_1 = 2\ \Phi(y_1) - 1 \ , \qquad v_2 = 2\ \Phi(y_2) - 1 \tag{9.17}$$

der projizierten Daten auf das Quadrat $(-1, 1) \times (-1, 1)$ durch. Dadurch sind v_1 und v_2 gleichverteilt, wenn y_1 und y_2 einer bivariaten Normalverteilung folgen. Als Abweichungsmaß kann wieder das Integral über das Quadrat der Differenz zwischen $p_v(v_1, v_2)$ und der Dichte der Gleichverteilung herangezogen werden,

$$\int_{-1}^{1}\int_{-1}^{1}\left[p_v(v_1, v_2) - \frac{1}{4}\right]^2 dv_1 dv_2 = \int_{-1}^{1}\int_{-1}^{1} p_v^2(v_1, v_2) dv_1 dv_2 - \frac{1}{4} \quad , \tag{9.18}$$

was einem Maß für die Abweichung von der Gleichverteilung entspricht.

Approximation von $p_v(v_1, v_2)$ mit Legendre-Polynomen liefert

$$\int_{-1}^{1}\int_{-1}^{1} p_v^2(v_1, v_2) dv_1 dv_2 - \frac{1}{4} = \frac{1}{4}\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}(2j+1)(2l+1)\ E^2\left[P_j(v_1)P_l(v_2)\right] - \frac{1}{4}\ . \tag{9.19}$$

Es werden dieselben Legendre-Polynome wie im eindimensionalen Fall verwendet.

Den bivariaten Projektionsindex erhält man durch Abschneiden der Summe der Ordnung J,

$$\begin{aligned} I(a, b) = &\sum_{j=1}^{J} \frac{2j+1}{4}\ E^2[P_j(v_1)] + \sum_{l=1}^{J} \frac{2l+1}{4}\ E^2[P_l(v_2)] \\ &+ \sum_{j=1}^{J}\sum_{l=1}^{J-j} \frac{(2j+1)(2l+1)}{4}\ E^2[P_j(v_1)P_l(v_2)]\ . \end{aligned}$$

Die Ableitungen des bivariaten Projektionsindex unter den Bedingungen $\boldsymbol{a}^\top \boldsymbol{a} = \boldsymbol{b}^\top \boldsymbol{b} = 1$ und $\boldsymbol{a}^\top \boldsymbol{b} = 0$ mit $1 \le k \le q$ und $1 \le m \le q$ sind

$$\frac{\partial I}{\partial a_k} = \frac{1}{\sqrt{2\pi}}\sum_{j=1}^{J}(2j+1)E[P_j(v_1)]\ E\left[P_j'(v_1)e^{-\frac{y_1^2}{2}}(z_k - a_k y_1 - b_k y_2)\right]$$

$$
\begin{aligned}
&+\frac{1}{\sqrt{2\pi}}\sum_{j=1}^{J}\sum_{l=1}^{J-j}(2j+1)(2l+1)\ E[P_j(v_1)P_l(v_2)] \\
&E\left[P_j'(v_1)P_l(v_2)e^{-\frac{y_1^2}{2}}(z_k-a_ky_1-b_ky_2)\right]\ , \\
\frac{\partial I}{\partial b_m} &= \frac{1}{\sqrt{2\pi}}\sum_{l=1}^{J}(2l+1)E[P_l(v_2)]\ E\left[P_l'(v_2)e^{-\frac{y_2^2}{2}}(z_m-a_my_1-b_my_2)\right] \\
&+\frac{1}{\sqrt{2\pi}}\sum_{j=1}^{J}\sum_{l=1}^{J-j}(2j+1)(2l+1)\ E[P_j(v_1)P_l(v_2)] \\
&E\left[P_j(v_1)P_l'(v_2)e^{-\frac{y_2^2}{2}}(z_m-a_my_1-b_my_2)\right]\ .
\end{aligned}
$$

y_1 und y_2 wurden in (9.16) definiert, v_1 und v_2 sind durch die Gleichungen (9.17) gegeben. Für eine konkrete Stichprobe werden zur Berechnung des Projektionsindex die Erwartungswerte durch die Stichprobenmittel geschätzt.

Die Berechnungszeit des eindimensionalen Projektionsindex steigt linear mit wachsender Ordnung J der Legendre-Polynome, im zweidimensionalen Fall steigt sie quadratisch.

9.2.3 Entfernung einer Struktur

Der Sinn von PP ist, möglichst viele Projektionen zu finden, die interessante Information enthalten. Wenn man eine solche Projektion gefunden hat, muss die gefundene Struktur entfernt werden, sonst würde immer die selbe Projektion gefunden werden. Es gibt verschiedene Möglichkeiten zum Entfernen von Strukturen (Huber, 1985), ein Weg ist eine rekursive Annäherung wie bei *Projection Pursuit Density Estimation* (Friedman, Stuetzle, und Schroeder, 1984). Die interessante Struktur wird entfernt und der Projektionsindex wird wieder maximiert. Der dabei verwendete Algorithmus ist sehr rechenintensiv.

Friedman (1987) schlägt zum Entfernen einer Struktur einen Algorithmus vor, der schneller berechenbar ist und auch für zweidimensionale Projektionen entwickelt wurde.

Nach der oben aufgestellten Hypothese ist man an Projektionen, deren Dichte der Dichtefunktion der Standardnormalverteilung nahekommt, nicht interessiert. Die Grundidee für die Entfernung einer Struktur ist daher, durch eine Transformation die Dichte der Projektion in eine Standardnormalverteilung überzuführen. Diese Transformation darf allerdings nur im Unterraum der Projektion wirksam sein und muss alle orthogonalen Richtungen unverändert lassen.

Die Prozedur wird erst für eindimensionale Projektionen beschrieben. Sei $\boldsymbol{a}$ die Projektionsrichtung, die durch Maximierung des Projektionsindex gefunden wurde, und $\boldsymbol{U}$ eine orthonormale $(q \times q)$-Matrix mit $\boldsymbol{a}$ in der ersten Zeile. Durch die Lineartransformation

$$\boldsymbol{T} = (T_1, \ldots, T_q)^\top = \boldsymbol{U}\boldsymbol{z} \tag{9.20}$$

erhält man eine Rotation, sodass die erste Koordinate

$$T_1 = \boldsymbol{a}^\top \boldsymbol{z} = y \tag{9.21}$$

ist. Sei weiters $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_q)^\top$ eine Transformation, deren Komponenten durch

$$\begin{aligned} \theta_1(T_1) &= \Phi^{-1}\left(F_a(T_1)\right) \ , \\ \theta_j(T_j) &= T_j \qquad \text{für } 2 \leq j \leq q \end{aligned} \tag{9.22}$$

definiert sind. $F_a(T_1)$ ist die kummulative Verteilungsfunktion der eindimensionalen Projektion (9.21). Durch Anwendung einer Transformation mit der Inversen der kummulative Standardnormalverteilungsfunktion erhält man als Resultat eine Standardnormalverteilung. Die Transformation $\boldsymbol{\Theta}$ wandelt also T_1 in eine Standardnormalverteilung um und lässt die Komponenten $T_2, \ldots, T_q$ unverändert.

Mit der Transformation

$$\boldsymbol{z}^* = \boldsymbol{U}^\top \boldsymbol{\Theta}(\boldsymbol{U}\boldsymbol{z}) \tag{9.23}$$

ist die durch die Projektion $y = \boldsymbol{a}^\top \boldsymbol{z}$ gefundene Struktur entfernt und die Maximierung des Projektionsindex kann auf $\boldsymbol{z}^*$ wieder angewandt werden.

Für konkrete Stichproben werden die theoretischen Größen von (9.22) durch die geschätzten Größen ersetzt und man erhält

$$y_i^* = \Phi^{-1}\left(\hat{F}_a(y_i)\right) = \Phi^{-1}\left(\frac{r(y_i) - \frac{1}{2}}{n}\right) \qquad \text{für } i = 1, \ldots, n \quad . \tag{9.24}$$

$\hat{F}_a(y_i)$ ist die empirische Verteilungsfunktion und $r(y_i)$ ist der Rang von y_i über die n projizierten Beobachtungen. Die Transformation (9.24) wird mit *Gaussianisierung* bezeichnet, sie ersetzt jede Beobachtung durch ihren normalverteilten Wert in der Projektion.

Es muss beachtet werden, dass wiederholte Gaussianisierung die Normalverteilung von nichtorthogonalen früher gefundenen Projektionen verändert, sodass der Projektionsindex größer als Null wird. Die so fälschlicherweise erhaltene Struktur ist aber in den meisten Fällen gering.

Zur Entfernung von zweidimensionalen Strukturen benötigt man eine Transformation, die eine allgemeine bivariate Dichtefunktion $p_{ab}(y_1, y_2)$ in die Dichte der bivariaten Normalverteilung

$$\phi(y_1, y_2) = \frac{1}{2\pi} e^{-\frac{y_1^2 + y_2^2}{2}} \tag{9.25}$$

umwandelt. Es ist bekannt, dass jede eindimensionale Projektion der bivariaten Normalverteilung normalverteilt ist. Daher muss $p_{ab}(y_1, y_2)$ in alle Richtungen des zweidimensionalen Unterraums auf Normalverteilung transformiert werden. Praktisch wäre das nicht durchführbar, daher normalisiert man nur in ein paar Richtungen, die man durch Rotation erhält.

Sei

$$\begin{aligned} y_1^* &= y_1 \cos\gamma + y_2 \sin\gamma \\ y_2^* &= y_2 \cos\gamma - y_1 \sin\gamma \end{aligned} \tag{9.26}$$

eine Rotation um den Winkel γ. Die Verteilungen von y_1^* und y_2^* werden auf die selbe Weise wie im eindimensionalen Fall zur Normalverteilung transformiert. Dieser Vorgang kann für verschiedene Werte von γ $(0, \frac{\pi}{8}, \frac{\pi}{4}, \frac{3\pi}{8})$ so lange wiederholt werden, bis die Verteilungen hinreichend genau die Normalverteilung approximieren.

Seien $\boldsymbol{a}$ und $\boldsymbol{b}$ die Projektionsrichtungen, die den Projektionsindex maximieren, und sei $\boldsymbol{U}$ eine orthonormale $(q \times q)$-Matrix mit $\boldsymbol{a}$ und $\boldsymbol{b}$ in den ersten beiden Zeilen. Mit der Transformation

$$\boldsymbol{T} = \boldsymbol{U}\boldsymbol{z} \tag{9.27}$$

erhält man als die ersten beiden Koordinaten von $\boldsymbol{T}$ die Koordinaten der Projektionsebene. Mit einer weiteren Transformation $\boldsymbol{\Theta} = (\theta_1, \ldots, \theta_q)^\top$, definiert durch

$$\begin{aligned} \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \end{pmatrix} &= \Phi^{-1}\left(F_{ab}(T_1, T_2)\right) \quad , \\ \theta_j(T_j) &= T_j \qquad \text{für } 3 \leq j \leq q \quad , \end{aligned} \tag{9.28}$$

wird erreicht, dass durch

$$\boldsymbol{z}^* = \boldsymbol{U}^\top \boldsymbol{\Theta}(\boldsymbol{U}\boldsymbol{z}) \tag{9.29}$$

die Lösungsebene $\{\boldsymbol{a}, \boldsymbol{b}\}$ von PP in eine bivariate Standardnormalverteilung umgewandelt wird und alle orthogonalen Projektionen unverändert bleiben.

9.2.4 Robustheit

Der ein- und zweidimensionale Projektionsindex ist robust gegenüber extremen Ausreißern, weil das Maß für die Abweichung von der Normalverteilung hauptsächlich durch das Zentrum der Verteilung beeinflusst wird und nur gering durch die Schwänze. Multivariate Ausreißer, die keine extremen Werte in irgendeine Richtung annehmen, beeinflussen den Projektionsindex allerdings schon. Solche Ausreißer können aber möglicherweise mit PP entdeckt werden, wenn sie eine starke Struktur liefern, sodass der Projektionsindex maximiert wird. Die Ausreißer können für die weitere Analyse ausgeschlossen werden.

Die Entfernung einer Struktur ist durch Ausreißer nicht beeinflusst. Nur die Skalierung der Daten ist nicht robust, da sie auf dem Stichprobenmittel und der Stichprobenkovarianzmatrix basiert und beide stark von Ausreißern beeinflusst werden. Allerdings gibt es bereits Verfahren für robuste Skalierung wie den MVE-Schätzer (*minimum volume ellipsoid*) und den MCD-Schätzer (*minimum covariance determinant*).

9.3 Andere Projektionsindices

9.3.1 Der Projektionsindex von Friedman und Tukey

Friedman und Tukey (1974) entwickelten einen Projektionsindex, um Cluster in den Daten zu finden. Dabei wird ein Kriterium maximiert, sodass möglichst viele Punkte

in einem Cluster liegen, während die Cluster möglichst separiert sein sollten. Der Projektionsindex hat somit die Form

$$I(\boldsymbol{a}) = s(\boldsymbol{a})\ d(\boldsymbol{a})\ ,$$

$\boldsymbol{a}$ ist die Projektionsachse, $s(\boldsymbol{a})$ misst die Streuung der Daten und $d(\boldsymbol{a})$ beschreibt eine „lokale Dichte“ der Punkte nach erfolgter Projektion auf $\boldsymbol{a}$. Für $s(\boldsymbol{a})$ wurde (aus Gründen der Robustheit) eine getrimmte Standardabweichung und für $d(\boldsymbol{a})$ eine „durchschnittliche Nähefunktion“ herangezogen.

9.3.2 Der Entropieindex

Sei f eine Dichtefunktion und $\int f \ln f$ ein Entropiemaß. Es kann gezeigt werden, dass jene Dichte mit Mittel Null und Varianz Eins, die dieses Maß minimiert, die Dichte der Standardnormalverteilung ist. Nimmt man also $\int f \ln f$ als Projektionsindex, so wird dadurch die Abweichung von der Dichte der Standardnormalverteilung gemessen. Diese Darstellung wurde von Huber (1985) eingeführt. Er zeigte auch, dass die *Fisher-Information* $\int \frac{(f')^2}{f}$ eine andere Möglichkeit für einen Projektionsindex ist, der durch die Dichte der Normalverteilung optimiert wird.

Wenn man einen Entropieindex auf eine konkrete Stichprobe anwenden möchte, muss vorher die Dichte f geschätzt werden. Dies kann durch eine *Kerndichteschätzung* (siehe z.B. Scott, 1992)

$$\hat{f}(y) = \frac{1}{nw} \sum_{i=1}^{n} K\left(\frac{y - y_i}{w}\right) \qquad y \in I\!R \tag{9.30}$$

geschehen. K ist eine Kernfunktion, $w \geq 0$ ein *Smoothing*-Parameter und y_i ist der i-te Datenpunkt, der auf $\boldsymbol{a}$ projiziert wurde. K ist typischerweise eine symmetrische nichtnegative Funktion mit Integral Eins, die monoton gegen Null geht, wenn ihr Argument gegen unendlich strebt. Ein Beispiel dafür ist die Dichte der Standardnormalverteilung.

9.3.3 Der Momentindex

Jones und Sibson (1987) zeigten, dass das Entropiemaß $\int f \ln f$ durch das dritte und vierte Moment, μ_3 und μ_4 , in der Form

$$\int f \ln f \approx \frac{\mu_3^2 + \frac{(\mu_4 - 3)^2}{4}}{12} \tag{9.31}$$

approximiert werden kann. Für zentrierte und skalierte normalverteilte Daten ist dieser Ausdruck Null, und er kann daher als Projektionsindex verwendet werden.

Ein Index, der auf dem dritten und vierten Moment basiert, wird möglicherweise auch durch andere Verteilungen optimiert. Allerdings sind bei konkreten Daten die höheren Momente in der Regel nicht sehr klein.

Literatur

J.H. Friedman. Exploratory Projection Pursuit. *J. Amer. Statist. Assoc.*, 82(397):249–266, 1987.

J.H. Friedman and W. Stuetzle. Projection Pursuit Regression. *J. Amer. Statist. Assoc.*, 76(376):817–823, 1981.

J.H. Friedman and J.W. Tukey. A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers*, c-23(9):881–890, 1974.

J.H. Friedman, W. Stuetzle, and A. Schroeder. Projection Pursuit Density Estimation. *J. Amer. Statist. Assoc.*, 79(387):599–608, 1984.

P.J. Huber. Projection Pursuit. *The Annals of Statistics*, 13(2):435–475, 1985.

M.C. Jones and R. Sibson. What is Projection Pursuit? *J. Roy. Statist. Soc. A*, 150(1):1–36, 1987.

D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, New York, 1992.