3 Modelle

3.1 Vektorielle Zufallsvariable

a **Zufallsvariable und Stichproben.** In der (univariaten) Statistik stehen am Anfang Beobachtungen, die bis auf „Schwankungen“ das gleiche ergeben sollten: Die Länge eines Sepalblattes einer Iris setosa, gemessen an 50 Pflanzen, ergibt 50 „im Prinzip gleiche“, im Detail aber verschiedene Werte x_i. Um diese Situation zu beschreiben, gehen wir zu einem Wahrscheinlichkeitsmodell über. Die Einzelwerte sind „Realisierungen“ einer **Zufallsvariablen** X mit einer Verteilung. Die Länge des Sepalblattes ist die Zufallsvariable, die für jede Pflanze einen bestimmten Wert hat. Der Erwartungswert oder ein anderer Lageparameter der Verteilung der Zufallsvariablen ist der „eigentliche Wert“, die gemessenen Werte „weichen zufällig davon ab“. Um in der Schliessenden Statistik weiter zu kommen, müssen wir jeder Beobachtungseinheit (Pflanze) i eine eigene Zufallsvariable X_i zuordnen. Im einfachsten Fall der zufälligen Stichprobe haben alle X_i die gleiche Verteilung – die „Verteilung von X“ und sind stochastisch unabhängig. Mit einem solchen Modell können wir dann die Verteilungen von Schätzungen und Teststatistiken bestimmen.

b **Zufallsvektor.** Im vorhergehenden Kapitel haben wir n Beobachtungen von mehreren Variablen $X^{(j)}$, $j = 1, 2, ..., m$, betrachtet und sie als Vektoren $\underline{x}_i$ geschrieben. Statt einer einzelnen Zufallsvariablen X, wie vorher beschrieben, brauchen wir nun als Modell einen **Zufallsvektor**

$$\underline{X} = \begin{bmatrix} X^{(1)} \\ X^{(2)} \\ \vdots \\ X^{(m)} \end{bmatrix},$$

charakterisiert durch eine Verteilung, nämlich die **gemeinsame Verteilung der Zufallsvariablen** $X^{(1)}, X^{(2)}, ..., X^{(m)}$. Im nächsten Schritt wird eine **Stichprobe von Zufallsvektoren** $\underline{X}_i$ betrachtet, die alle die gleiche Verteilung zeigen und unabhängig von einander sind. Schliesslich kann man alle Daten der Stichprobe zu einer zufälligen **Datenmatrix**

$$\boldsymbol{X} = \begin{bmatrix} \underline{X}_1^T \\ \dots \\ \underline{X}_n^T \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & X_1^{(2)} & \dots & X_1^{(m)} \\ X_2^{(1)} & X_2^{(2)} & \dots & X_2^{(m)} \\ \vdots & \vdots & & \vdots \\ X_n^{(1)} & X_n^{(2)} & \dots & X_n^{(m)} \end{bmatrix}.$$

zusammenfassen. Achtung! Die Zufallsvektoren $\underline{X}_i$ werden (wie die $\underline{x}_i$) als Spaltenvektoren geschrieben, obwohl sie einer Zeile der Datenmatrix $\boldsymbol{X}$ entsprechen!

c **Erwartungswert und Varianz.** Mittelwertsvektor und Kovarianzmatrix für eine Matrix von Datenwerten haben wir bereits behandelt (2.5.d, 2.5.f). Der Übergang zu den entsprechenden theoretischen Kenngrössen geschieht genau wie in der univariaten Statistik: Alle Mittelwerte $\frac{1}{n}\sum ...$ und „Beinahe-Mittelwerte“ $\frac{1}{n-1}\sum ...$ werden durch Erwartungswerte ersetzt.

Fast unnötig zu sagen: Der Erwartungswert eines Zufallsvektors (oder einer zufälligen Matrix) ist natürlich einfach der Vektor der Erwartungswerte der Elemente, die ja gewöhnliche Zufallsvariable sind,

$$\underline{\mu} = \mathcal{E}\langle \underline{X} \rangle \;=\; \begin{bmatrix} \mathcal{E}\langle X^{(1)} \rangle \\ \mathcal{E}\langle X^{(2)} \rangle \\ \vdots \\ \mathcal{E}\langle X^{(m)} \rangle \end{bmatrix} .$$

Aus empirischen Varianzen $\widehat{\mathrm{var}}$ und Kovarianzen $\widehat{\mathrm{cov}}$ werden „theoretische“ Grössen var und cov,

$$\boldsymbol{\Sigma} = \mathrm{var}\langle \underline{X} \rangle = \begin{bmatrix} \mathrm{var}\langle X^{(1)} \rangle & \mathrm{cov}\langle X^{(1)}, X^{(2)} \rangle & ... & \mathrm{cov}\langle X^{(1)}, X^{(m)} \rangle \\ \mathrm{cov}\langle X^{(2)}, X^{(1)} \rangle & \mathrm{var}\langle X^{(2)} \rangle & ... & \mathrm{cov}\langle X^{(2)}, X^{(m)} \rangle \\ \vdots & \vdots & ... & \vdots \\ \mathrm{cov}\langle X^{(m)}, X^{(1)} \rangle & \mathrm{cov}\langle X^{(m)}, X^{(2)} \rangle & ... & \mathrm{var}\langle X^{(m)} \rangle \end{bmatrix} .$$

d **Kovarianzmatrix als Erwartungswert.** Für einfache Zufallsvariable ist $\mathrm{var}\langle X \rangle = \mathcal{E}\left\langle (X-\mu)^2 \right\rangle = \mathcal{E}\langle X^2 \rangle - \mu^2$. Für einen Zufallsvektor gilt ein entsprechendes Resultat:

$$\boldsymbol{\Sigma} = \mathrm{var}\langle \underline{X} \rangle = \mathcal{E}\left\langle \left(\underline{X} - \underline{\mu}\right)\left(\underline{X} - \underline{\mu}\right)^T \right\rangle = \mathcal{E}\langle \underline{X}\underline{X}^T \rangle - \underline{\mu}\underline{\mu}^T .$$

Es ist $\underline{X} - \underline{\mu}$ ein Spaltenvektor der Länge m und deshalb $(\underline{X} - \underline{\mu})\,(\underline{X} - \underline{\mu})^T$ eine $m \times m$-Matrix! Für den Beweis dieser Formeln pickt man sich einfach ein beliebiges Element $\boldsymbol{\Sigma}_{jk}$ heraus und stellt fest, dass das entsprechende Resultat aus der Theorie der gewöhnlichen Zufallsvariablen bekannt ist.

e **Lineare Transformationen.** Alle Formeln, die im letzten Kapitel für Mittelwerte und empirische Varianzen und Kovarianzmatrizen von Linearkombinationen, Projektionen und linear transformierten Vektoren von Variablen hergeleitet wurden, gelten auch für die entsprechenden theoretischen Grössen. Die wichtigsten sind

$$\mathcal{E}\langle \underline{a} + \boldsymbol{B}\underline{X} \rangle = \underline{a} + \boldsymbol{B}\,\mathcal{E}\langle \underline{X} \rangle \;, \qquad \mathrm{var}\langle \underline{a} + \boldsymbol{B}\underline{X} \rangle = \boldsymbol{B}\,\mathrm{var}\langle \underline{X} \rangle\,\boldsymbol{B}^T .$$

f **Summen von unabhängigen Zufallsvektoren.** Wenn man zwei unabhängige Zufallsvektoren $\underline{X}_1$ und $\underline{X}_2$ zusammenzählt, dann addieren sich wie in der univariaten Statistik Erwartungswert und Varianz,

$$\mathcal{E}\langle \underline{X}_1 + \underline{X}_2 \rangle = \mathcal{E}\langle \underline{X}_1 \rangle + \mathcal{E}\langle \underline{X}_2 \rangle \;, \quad \mathrm{var}\langle \underline{X}_1 + \underline{X}_2 \rangle = \mathrm{var}\langle \underline{X}_1 \rangle + \mathrm{var}\langle \underline{X}_2 \rangle \;.$$

* Zum Beweis kann man die beiden Vektoren zu einem einzigen, doppelt so langen $\underline{X}_{12}$ zusammenhängen und auf $\boldsymbol{B}\underline{X}_{12}$ mit $\boldsymbol{B} = [\boldsymbol{I} \quad \boldsymbol{I}]$ die Regeln der linearen Transformation anwenden.

3.2 Die mehrdimensionale Normalverteilung

a **Mehrdimensionale Verteilung.** Die Verteilung einer Zufallsvariablen ist durch die kumulative **Verteilungsfunktion** $F\langle x\rangle = P\langle X \le x\rangle$ festgelegt. Für eine mehrdimensionale Verteilung ist die Verteilungsfunktion analog definiert; $\underline{X} \le \underline{x}$ soll dabei heissen, dass für alle Variablen $X^{(j)} \le x^{(j)}$ gilt. Also ist

$$F\langle \underline{x}\rangle = P\Big\langle X^{(1)} \le x^{(1)}, X^{(2)} \le x^{(2)}, ..., X^{(m)} \le x^{(m)} \Big\rangle \ .$$

Wie im univariaten Fall haben gängige Verteilungen eine **Dichte** $f\langle \underline{x}\rangle$, die die Ableitung der kumulativen Verteilungsfunktion nach $\underline{x}$ – im Sinne der partiellen Ableitung nach allen Komponenten – darstellt. Aus ihr kann man Wahrscheinlichkeiten für Ereignisse durch Integration erhalten.

Ein Ereignis, das durch die Werte des Zufallsvektors festgelegt ist, ist eine Menge $\mathcal{A}$ im m-dimensionalen Raum; das Ereignis „tritt ein“, wenn $\underline{X} \in \mathcal{A}$ gilt oder, in Worten, wenn die Realisierung des Zufallsvektors einen Punkt in $\mathcal{A}$ liefert. Die Wahrscheinlichkeit des Ereignisses erhält man aus der Dichte als $P\langle\mathcal{A}\rangle = \int_{\underline{u}\in\mathcal{A}} f\langle \underline{u}\rangle\, du^{(1)}...du^{(m)}$. Speziell gilt für die Verteilungsfunktion

$$\begin{aligned} F\langle \underline{x}\rangle &= P\langle X^{(1)} \le x^{(1)}, ..., X^{(m)} \le x^{(m)}\rangle \\ &= \int_{u^{(1)}\le x^{(1)},...,u^{(m)}\le x^{(m)}} f\langle \underline{u}\rangle du^{(1)}...du^{(m)} \ . \end{aligned}$$

b **Mehrdimensionale Standard-Normalverteilung.** Das wirkt alles recht abstrakt. Eine der einfachsten mehrdimensionalen Verteilungen ist die m-dimensionale Standard-Normalverteilung, die wir mit Φ_m bezeichnen wollen. Sie ist dadurch festgelegt, dass die Komponenten unabhängig und standard-normalverteilt sind,

$$\underline{Z} \sim \Phi_m \quad \Longleftrightarrow \quad Z^{(j)} \sim \Phi_1 \ , \quad \text{unabhängig} \ .$$

Die Dichte dieser Verteilung ergibt sich aus der Tatsache, dass sich Dichten von unabhängigen Zufallsvariablen zur Dichte der gemeinsamen Verteilung multiplizieren,

$$f\langle \underline{z}\rangle = \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi}} \exp\Big\langle -z^{(j)2}/2 \Big\rangle = (2\pi)^{-m/2} \exp\big\langle -\|z\|^2/2 \big\rangle = \widetilde{f}\big\langle \|z\|^2 \big\rangle \ .$$

Die Dichte ist also nur eine Funktion der Länge $\|z\|$ des Vektors $\underline{z}$; die Konturen gleicher Dichte bilden für $m=2$ Kreise, in höheren Dimensionen (Hyper-) Kugel-Oberflächen (Abbildung 3.2.g).

Diese Verteilung ist natürlich selten ein sinnvolles Modell für reale Daten; sie lässt ja keine Abhängigkeit zwischen Variablen zu und beschreibt deshalb höchstens die uninteressanteste Situation.

c **Verteilung von Linearkombinationen.** Wir brauchen ein Modell, das korrelierte Variable beschreiben kann. Dazu zunächst eine Vorbetrachtung: Aus der Theorie der gewöhnlichen Normalverteilung ist bekannt, dass jede Linearkombination $b_1 Z^{(1)} + b_2 Z^{(2)} + ... + b_m Z^{(m)}$ von normalverteilten Zufallsvariablen $Z^{(j)}$ wieder normalverteilt ist. Erwartungswert und Varianz kennen wir von früher (3.1.e resp. 2.6.b und 3.1.f resp. 2.6.c): $b_1 \mathcal{E}\langle Z^{(1)}\rangle + b_2 \mathcal{E}\langle Z^{(2)}\rangle + ... + b_m \mathcal{E}\langle Z^{(m)}\rangle$ und $b_1^2 \operatorname{var}\langle Z^{(1)}\rangle + b_2^2 \operatorname{var}\langle Z^{(2)}\rangle + ... + b_m^2 \operatorname{var}\langle Z^{(m)}\rangle$. Also gilt für $X = \underline{b}^T \underline{Z}$

$$X = \underline{b}^T \underline{Z} \sim \mathcal{N}\langle 0, \textstyle\sum_j b_j^2\rangle = \mathcal{N}\langle 0, \|\underline{b}\|^2\rangle \ .$$

In 2.6.e haben wir mehrere Linearkombinationen gemeinsam betrachtet und den Begriff der linearen Transformation eingeführt. Betrachten wir nun eine lineare Transformation eines standard-normalverteilten Zufallsvektors $\underline{Z}$, $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$! Es gilt

$$\mathcal{E}\langle \underline{X}\rangle = \underline{\mu} \ , \qquad \operatorname{var}\langle \underline{X}\rangle = \boldsymbol{B}\,\boldsymbol{B}^T \ .$$

d **Multivariate Normalverteilung.** Die Familie der m-dimensionalen Normalverteilungen ist die Familie der Verteilungen aller Zufallsvektoren $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$, wobei $\underline{Z}$ m-dimensional standard-normalverteilt ist und $\boldsymbol{B}$ quadratisch (und $\underline{\mu}$ irgendein m-dimensionaler Vektor).

Es liegt nahe, die Grössen $\underline{\mu}$ und $\boldsymbol{B}$ als Parameter dieser Verteilungsfamilie anzusehen. Es zeigt sich aber, dass $\boldsymbol{B}$ dazu nicht geeignet ist: Falls $\boldsymbol{B}$ orthogonal ist – beispielsweise eine Drehung im zweidimensionalen Fall – dann haben $\underline{Z}$ und $\boldsymbol{B}\underline{Z}$ die gleiche Verteilung. Verschiedene Parameterwerte ($\boldsymbol{B}$ und $\boldsymbol{I}$), die die gleiche Verteilung bezeichnen, kann man nicht brauchen; jede Verteilung der Familie soll einen eindeutigen Satz von Parametern haben. Diese Forderung wird **Identifizierbarkeit** genannt.

Der geeignete Parameter neben dem Erwartungswert $\underline{\mu}$ ist die Kovarianzmatrix $\boldsymbol{\Sigma}$. Wenn zwei Matrizen $\boldsymbol{B}$ und $\boldsymbol{B'}$ zum gleichen $\boldsymbol{\Sigma}$ führen, also wenn $\boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{B'}\boldsymbol{B'}^T$ ist, dann (und nur dann) sind auch die Verteilungen der entsprechenden Zufallsvektoren $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$ und $\underline{X} = \underline{\mu} + \boldsymbol{B'}\underline{Z}$ gleich.

e* Für den Fall einer regulären Kovarianzmatrix $\boldsymbol{\Sigma}$ ist das nicht schwierig zu zeigen: Es sei $\boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{B'}\boldsymbol{B'}^T = \boldsymbol{\Sigma}$. Wenn $\boldsymbol{\Sigma}$ invertierbar ist, muss es auch $\boldsymbol{B}$ sein (wie man in der linearen Algebra beweist). Multipliziert man die letzte Gleichung mit $\boldsymbol{B}^{-1}$ von links und $(\boldsymbol{B}^{-1})^T$ von rechts, so erhält man $\boldsymbol{I} = \boldsymbol{B}^{-1}\boldsymbol{B'}\boldsymbol{B'}^T(\boldsymbol{B}^{-1})^T = (\boldsymbol{B}^{-1}\boldsymbol{B'})(\boldsymbol{B}^{-1}\boldsymbol{B'})^T$. Also ist $\boldsymbol{Q} = \boldsymbol{B}^{-1}\boldsymbol{B'}$ eine orthogonale Matrix. Es gilt $\boldsymbol{B'} = \boldsymbol{B}\,\boldsymbol{Q}$. Wir betrachten $\underline{X'} = \underline{\mu} + \boldsymbol{B'}\underline{Z} = \underline{\mu} + \boldsymbol{B}\,\boldsymbol{Q}\underline{Z}$. Da die Verteilung von $\underline{Z}$ und $\underline{Z'} = \boldsymbol{Q}\underline{Z}$ die gleiche ist – nämlich die Standard-Normalverteilung – hat $\underline{X'} = \underline{\mu} + \boldsymbol{B}\underline{Z'}$ die gleiche Verteilung wie $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$.

f Die **multivariate Normalverteilung** mit Erwartungswert $\underline{\mu}$ und Kovarianzmatrix $\boldsymbol{\Sigma}$, geschrieben als $\mathcal{N}_m\langle \underline{\mu}, \boldsymbol{\Sigma}\rangle$, ist die Verteilung von $\underline{X} = \underline{\mu} + \boldsymbol{B}\underline{Z}$, wobei $\underline{Z} \sim \Phi_m$ und $\boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{\Sigma}$ ist.

Als Parameterwerte für $\boldsymbol{\Sigma}$ kommen alle $m \times m$-Matrizen in Frage, die symmetrisch und „positiv semidefinit“ sind (2.6.m). Für jede solche Matrix lassen sich nämlich Matrizen $\boldsymbol{B}$ finden, für die $\boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{\Sigma}$ ist (2.6.m).

g **Dichte.** Falls $\boldsymbol{\Sigma}$ nicht singulär ist, hat die multivariate Normalverteilung die Dichte

$$f\langle \underline{x} \rangle = \tfrac{1}{c} \cdot \exp\big\langle -(\underline{x} - \underline{\mu})^T \boldsymbol{\Sigma}^{-1} (\underline{x} - \underline{\mu})/2 \big\rangle$$

mit der Normierungskonstanten $c = (2\pi)^{m/2} \det\langle \boldsymbol{\Sigma} \rangle^{1/2}$. Wenn $\boldsymbol{\Sigma}$ singulär ist, dann gibt es keine Dichte; die Verteilung ist dann auf einen Unterraum (oder, für Mathematiker/innen: „Nebenraum“) konzentriert.

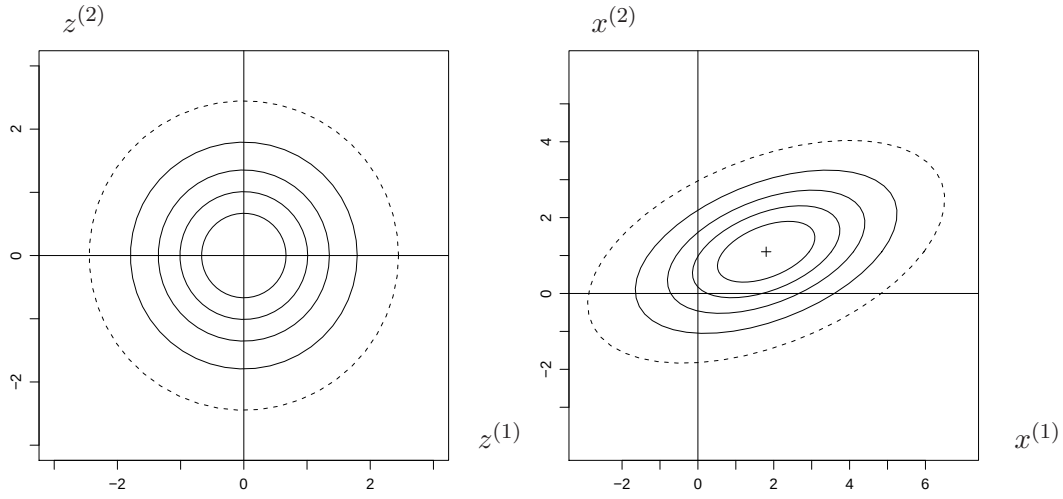


Abbildung 3.2.g: „Höhenlinien“ gleicher Dichten für die Standard- und eine allgemeine Normalverteilung

Die Dichte ist konstant für $(\underline{x}-\underline{\mu})^T \boldsymbol{\Sigma}^{-1} (\underline{x}-\underline{\mu}) =$ Konstante (siehe Abbildung 3.2.g). In zwei Dimensionen ist das die Gleichung einer Ellipse; in höheren Dimensionen spricht man von **Ellipsoid**. Für verschiedene Konstante entstehen konzentrische Ellipsen (Ellipsoide); das Zentrum ist immer der Erwartungswertsvektor $\underline{\mu}$.

h **Schätzung der Parameter.** Modelle sollen Daten beschreiben. Damit wir dies für das Beispiel der Iris-Blüten zeigen können, müssen wir zunächst die am besten passende Normalverteilung wählen, also die Parameter $\underline{\mu}$ und $\boldsymbol{\Sigma}$ schätzen. Es ist nahe liegend, dafür den Mittelwertsvektor $\widehat{\underline{\mu}} = \overline{\underline{X}}$ und die empirische Kovarianzmatrix $\widehat{\boldsymbol{\Sigma}}$ einzusetzen. Abbildung 3.2.h zeigt die Daten für die ersten beiden Variablen der Blüten von Iris setosa, zusammen mit den Ellipsen, die die angepasste Normalverteilung darstellen (siehe 3.2.m). Mit den Eigenschaften dieser Schätzungen befasst sich das nächste Kapitel.

i **Lineare Transformation.** Wenn ein multivariat normalverteilter Zufallsvektor $\underline{X} \sim \mathcal{N}_m\langle \underline{\mu}, \boldsymbol{\Sigma} \rangle$ linear transformiert wird zu $\underline{Y} = \underline{a} + \boldsymbol{B}\underline{X}$, dann ist auch der transformierte Vektor multivariat normalverteilt,

$$\underline{Y} \sim \mathcal{N}_m\big\langle \underline{a} + \boldsymbol{B}\underline{\mu},\, \boldsymbol{B}\,\boldsymbol{\Sigma}\,\boldsymbol{B}^T \big\rangle \ .$$

Beweis: $\underline{X}$ entsteht selbst durch lineare Transformation aus $\underline{Z}$. Setzt man die beiden linearen Transformationen zusammen, dann ist also $\underline{Y}$ selbst ein linear transformierter,

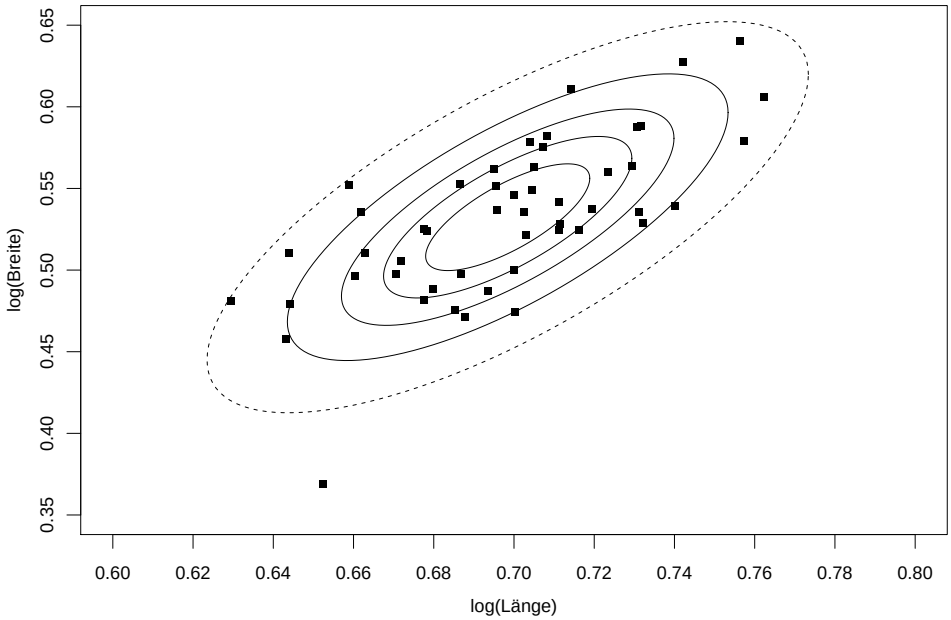


Abbildung 3.2.h: Daten und angepasste Normalverteilung im Beispiel der Irisblüten für die Art Iris setosa

standard-normalverteilter Zufallsvektor. Falls $\boldsymbol{B}$ quadratisch ist, entspricht $\underline{X}$ deshalb der Definition einer multivariaten Normalverteilung.

* Falls $\boldsymbol{B}$ weniger Zeilen als Spalten hat, ist das Resultat $\underline{Y}$ kürzer als der Ausgangsvektor $\underline{X}$. Man kann dann $\boldsymbol{B}$ beliebig zu einer quadratischen Matrix erweitern und erhält einen erweiterten Resultatvektor $\underline{\widetilde{Y}}$. Die gesuchte Verteilung ist die „Randverteilung" der ersten Komponenten, und das Ergebnis folgt aus 3.2.p. Falls schliesslich $\boldsymbol{B}$ weniger Spalten als Zeilen hat, ergänzt man sie durch Nullen zu einer quadratischen Matrix und erhält das Ergebnis wie für quadratische $\boldsymbol{B}$.

Die Formeln für die Bestimmung der Parameter der Normalverteilung von $\underline{Y}$ aus jenen von $\underline{X}$ kennen wir bereits (2.6.i, 3.1.e).

j* **Charakterisierung.** Die letzte Eigenschaft „charakterisiert" die multivariate Normalverteilung:

Wenn jede Linearkombination $\underline{b}^T\underline{X}$, $\underline{b} \in \mathbb{R}^m$, normalverteilt ist (oder allenfalls degeneriert), dann ist $\underline{X}$ multivariat normalverteilt.

k **Standardisierter Zufallsvektor.** Wir haben aus einem Zufallsvektor $\underline{Z}$ mit Erwartungswert $\underline{0}$ und Kovarianzmatrix $\boldsymbol{I}$ durch lineare Transformation einen Vektor $\underline{X}$ erzeugt, der einen Erwartungswert $\underline{\mu}$ und eine Kovarianzmatrix $\boldsymbol{\Sigma}$ aufweist. Nun soll umgekehrt aus $\underline{X}$ ein standardisierter Zufallsvektor erzeugt werden, wie wir dies für Stichproben bereits getan haben (2.6.m). Man muss dazu in den erwähnten Formeln einfach den Erwartungswert $\underline{\mu}$ statt des Mittelwertsvektors $\underline{\overline{x}}$ und die theoretische Kovarianzmatrix $\boldsymbol{\Sigma}$ statt der empirischen $\widehat{\boldsymbol{\Sigma}}$ einsetzen und man erhält

$$\underline{Z} = \boldsymbol{B}^{-1}(\underline{X} - \underline{\mu}) \quad \text{mit} \quad \boldsymbol{B}\,\boldsymbol{B}^T = \boldsymbol{\Sigma}\,.$$

Dieser Zufallsvektor hat Erwartungswert $\underline{0}$ und Kovarianzmatrix $\boldsymbol{I}$, auch wenn $\underline{X}$ nicht normalverteilt war. Wenn die Normalverteilung vorausgesetzt wird, ist $\underline{Z}$ natürlich standard-normalverteilt.

l **Chiquadrat-Verteilung.** Die χ^2-Verteilung mit m Freiheitsgraden ist *definiert* als Verteilung der Summe von m unabhängigen, quadrierten standard-normalverteilten Zufallsvariablen $Z^{(j)}$, also von $U = \sum_{j=1}^{m} Z^{(j)2}$, $Z^{(j)} \sim \mathcal{N}\langle 0,1\rangle$, oder

$$U = \sum_{j=1}^{m} Z^{(j)2} = \|\underline{Z}\|^2 , \qquad \underline{Z} \sim \Phi_m .$$

Sie ist also die Verteilung der quadrierten Länge eines standard-normalverteilten Zufallsvektors.

Die χ^2_m-Verteilung hat die Dichte

$$f_m\langle u\rangle = \frac{1}{2^{m/2}\Gamma\langle m/2\rangle} \cdot u^{m/2-1} e^{-u/2} .$$

(Für die Normierungskonstante wird die so genannte Gamma-Funktion Γ benötigt.)

m **Mahalanobis-Distanz.** Setzen wir die beiden letzten Gedanken zusammen: Ausgehend von einem multivariaten Zufallsvektor $\underline{X}$ bilden wir die quadrierte Länge des zugehörigen *standardisierten* Vektors $\underline{Z}$,

$$d^2\langle \underline{X}, \underline{\mu}; \boldsymbol{\Sigma}\rangle = \|\underline{Z}\|^2 = \underline{Z}^T\underline{Z} = (\underline{X}-\underline{\mu})^T \boldsymbol{C}^T \boldsymbol{C} (\underline{X}-\underline{\mu}) = (\underline{X}-\underline{\mu})^T \boldsymbol{\Sigma}^{-1} (\underline{X}-\underline{\mu}) .$$

Diese Grösse heisst nach einem indischen Statistiker quadrierte „Mahalanobis-Distanz“ von $\underline{X}$ zu seinem Erwartungswert $\underline{\mu}$.

Wenn $\underline{X}$ normalverteilt ist, dann hat d^2, wie gesagt, eine χ^2-Verteilung mit m Freiheitsgraden.

n* **Konturen der Dichte zeichnen.** Die Distanz bestimmt gerade auch die Wahrscheinlichkeits-Dichte der Normalverteilung, wie man aus 3.2.g sehen kann; die Vektoren $\underline{x}$, die $d^2\langle \underline{x}, \underline{\mu}; \boldsymbol{\Sigma}\rangle =$ konstant erfüllen, bilden eine Kontur gleicher Dichte. Um sie konkret für eine Zeichnung wie Abbildung 3.2.h bestimmen zu können, muss man also eine zu $\widehat{\boldsymbol{\Sigma}}$ gehörige $\boldsymbol{B}$-Matrix $\widehat{\boldsymbol{B}}$ finden – beispielsweise mit Cholesky-Zerlegung –, die Punkte $\underline{z}_k = c \cdot [\cos\langle k\Delta\rangle, \sin\langle k\Delta\rangle]^T$ eines Kreises zu $\underline{x}_k = \underline{\widehat{\mu}} + \underline{z}_k \widehat{\boldsymbol{B}}$ transformieren und in die Grafik einzeichnen. Den Faktor c bestimmt man so, dass der Kreis eine gewünschte Wahrscheinlichkeit π einschliesst; dazu muss c^2 das π-Quantil der χ^2-Verteilung mit m Freiheitsgraden sein.

o **Q-Q-Diagramm.** Wenn man Verteilungs-Annahmen für eine einzelne Zufallsvariable überprüfen will, kann man ein Histogramm mit einer angepassten Dichtekurve vergleichen oder ein Quantil-Quantil-Diagramm betrachten (siehe beispielsweise Abschnitt 11.2 in Stahel (2002)). Die gemeinsame Verteilung von zwei Variablen lässt sich allenfalls noch anhand einer gemeinsamen Darstellung der Daten und der Konturen der Modellverteilung wie in Abbildung 3.2.h beurteilen. Für höhere Dimensionen sind solche grafischen Mittel nicht möglich.

Immerhin können wir jetzt einen Aspekt der gemeinsamen Verteilung prüfen; die Verteilung der Mahalanobis-Distanzen $d^2\langle \underline{X}_i, \underline{\widehat{\mu}}; \widehat{\boldsymbol{\Sigma}}\rangle$ müsste ungefähr eine χ^2-Verteilung mit m Freiheitsgraden sein. Damit die Figur (für kleine Dimensionszahl m) nicht allzu

sehr durch die grossen Werte dominiert wird, vergleichen wir die Werte von d, also die Wurzeln aus den d^2-Werte, mit den Wurzeln aus den Quantilen der χ^2-Verteilung.

Abbildung 3.2.o zeigt für die Sepal-Blätter der Iris setosa eine gute Übereinstimmung der Daten mit der theoretischen Vorstellung, bis auf einen Ausreisser, der in Abbildung 3.2.h schon sichtbar ist.

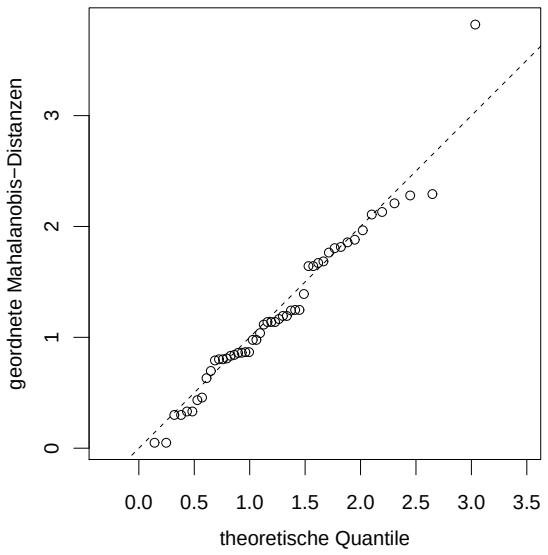


Abbildung 3.2.o: Q-Q-Diagramm der Mahalanobis-Distanzen für die logarithmierten Längen und Breiten der Sepalblätter der 50 Iris setosa-Pflanzen

p **Randverteilungen.** Wenn die gemeinsame Verteilung von mehreren Zufallsvariablen festgelegt ist, dann ist auch die Verteilung der einzelnen Zufallsvariablen, für sich allein betrachtet, bestimmt. Diese Verteilungen heissen **Randverteilungen** oder Marginalverteilungen. (Der Ausdruck ist vor allem bei diskreten Variablen anschaulich: Die gemeinsame Verteilung von zwei diskreten Zufallsvariablen kann in einer Tabelle angegeben werden. Die „Randsummen“ in einer solchen Tabelle geben die Verteilungen der beiden Variablen wieder.)

Bei mehr als zwei Zufallsvariablen können auch „mehrdimensionale Ränder“ betrachtet werden: Bei drei Variablen kann man sich für die (gemeinsame) Verteilung der ersten beiden interessieren, ohne Berücksichtigung des Wertes der dritten. Das ist die Randverteilung von $[X^{(1)}, X^{(2)}]^T$.

Der Einfachheit halber nehmen wir an, dass es die ersten p Variablen sind, die uns interessierten. (Der allgemeine Fall wird nur in der Notation schwerfälliger, ist aber sonst nicht schwieriger.) Wir schreiben die ersten p Komponenten von $\underline{X}$ als $\underline{X}^{[1:p]}$. Wir brauchen, gelegentlich auch später, eine Notation für **aufgeteilte Vektoren und Matrizen** *(partitioned vectors and matrices):*

$$\underline{a} = \begin{bmatrix} \underline{a}^{[1:p]} \\ \underline{a}^{[(p+1):m]} \end{bmatrix} = \left[\begin{array}{c} a^{(1)} \\ \dots \\ a^{(p)} \\ \hline a^{(p+1)} \\ \dots \\ a^{(m)} \end{array}\right]$$

$$\boldsymbol{C} = \begin{bmatrix} C^{[1:p]}_{[1:p]} & C^{[1:p]}_{[(p+1):m]} \\ C^{[(p+1):m]}_{[1:p]} & C^{[(p+1):m]}_{[(p+1):m]} \end{bmatrix} = \left[\begin{array}{ccc|ccc} C_{11} & \dots & C_{1p} & C_{1,p+1} & \dots & C_{1m} \\ & \dots & & & \dots & \\ C_{p1} & \dots & C_{pp} & C_{p,p+1} & \dots & C_{pm} \\ \hline C_{p+1,1} & \dots & C_{p+1,p} & C_{p+1,p+1} & \dots & C_{p+1,m} \\ & \dots & & & \dots & \\ C_{m1} & \dots & C_{mp} & C_{m,p+1} & \dots & C_{mm} \end{array}\right]$$

Wenn die Verteilung von $\underline{X}$ eine Dichte hat, dann erhält man für die Dichte der Randverteilung von $\underline{X}^{[1:p]}$ formal durch Integration über die „überflüssigen“ Variablen,

$$f^{[1:p]}\left\langle \underline{x}^{[1:p]} \right\rangle = \int_{u^{(m)}} \dots \int_{u^{(p+1)}} f\left\langle x^{(1)}, ..., x^{(p)}, u^{(p+1)}, ..., u^{(m)} \right\rangle du^{(p+1)} ... du^{(m)} \ .$$

q Die mehrdimensionale Normalverteilung erfüllt auch hier, was man sich von ihr erhoffen kann: Jede Randverteilung ist selbst eine Normalverteilung. Die Parameter sind gegeben durch das oder die entsprechende(n) Element(e) des Erwartungswerts-Vektors μ und die entsprechende(n) Zeile(n) und Spalte(n) von $\boldsymbol{\Sigma}$. Also gilt

$$\underline{X}^{[1:p]} \sim \mathcal{N}_p\Big\langle \underline{\mu}^{[1:p]}, \boldsymbol{\Sigma}^{[1:p]}_{[1:p]} \Big\rangle \ .$$

r* **Korrelation und Unabhängigkeit.** Zwei unkorrelierte Zufallsvariable müssen nicht unabhängig sein. Dafür gibt es einfache Beispiele (vergleiche Stahel (2002), 3.2.i). Eines davon beschreibt eine quadratische Regression und besteht aus einer um 0 symmetrischen Zufallsvariablen $X^{(1)}$ und einer zweiten $X^{(2)} = (X^{(2)})^2 + E$ mit einer von $X^{(1)}$ unabhängigen Variablen E. Aus Symmetriegründen kann man leicht einsehen, dass die Korrelation von $X^{(1)}$ und $X^{(2)}$ null ist.

Man könnte hoffen, dass aus Korrelation null die Unabhängigkeit wenigstens dann folgt, wenn die beiden Variablen normalverteilt sind. Leider gilt auch das nicht, wie das folgende Beispiel (Th.3.2.8 in Flury (1997)) zeigt: Sei $X_1 \sim \mathcal{N}\langle 0, 1\rangle$ und $X_2 = X_1$ mit Wahrscheinlichkeit 0.5, $X_2 = -X_1$ sonst. Dann ist $\mathrm{cov}\langle X_1, X_2\rangle = 0$, aber die gemeinsame Verteilung liegt auf den beiden „Diagonalen“ – den Geraden durch den Nullpunkt mit Steigung 1 respektive -1 –, also sind X_1 und X_2 nicht unabhängig.

Wenn schliesslich die gemeinsame Verteilung von X_1 und X_2 eine zweidimensionale Normalverteilung ist, dann ist der Schluss zulässig: Korrelation null kann dann nur für unabhängige Zufallsvariable auftreten. Das zeigt man wie folgt: Wenn $\boldsymbol{\Sigma}_{12} = 0$ ist, dann ist $\boldsymbol{\Sigma}^{-1} = \mathrm{diag}\langle 1/\boldsymbol{\Sigma}_{11}, 1/\boldsymbol{\Sigma}_{22}\rangle$. Setzt man das in die Formel für die Dichte ein, dann sieht man, dass diese „faktorisiert“, also als Produkt $f\langle x_1, x_2\rangle = f_1\langle x_1\rangle \cdot f_2\langle x_2\rangle$ geschrieben werden kann. Also sind die beiden Zufallsvariablen unabhängig.

s **Andere Verteilungen.** In der univariaten Statistik spielt die Normalverteilung eine entscheidende Rolle, aber für die Verteilung von Daten gibt es noch eine Reihe anderer gebräuchlicher Modelle: Lognormal-, Exponential-, Weibull-, Gamma-Verteilungen und andere.

In der multivariaten Statistik gäbe es einerseits noch viel mehr Möglichkeiten, Modelle zu definieren, andererseits gibt es kaum brauchbare Alternativen zur multivariaten Normalverteilung. Das liegt an zwei Schwierigkeiten: Einerseits ist man viel stärker auf die Vereinfachungen angewiesen, die sich durch die Eigenschaften der Normalverteilung ergeben, da im höher dimensionalen Raum unsere Vorstellungskraft und die Möglichkeiten, alle Probleme mit „brutaler Rechengewalt“ zu lösen, rasch schwinden. Andererseits führen die vielen Möglichkeiten von gemeinsamen Verteilungen rasch in eine Beliebigkeit, die es für einzelne Modelle schwierig macht, sich durchzusetzen.

Immerhin ist eine weitere Verteilung erwähnenswert: Die **multivariate Lognormal-Verteilung** ist analog zum univariaten Fall dadurch definiert, dass der Vektor der logarithmierten Komponenten eine Normalverteilung zeigt.

t* **Elliptische Verteilungen.** Einigermassen bekannt sind die so genannten elliptischen Verteilungen. Sie entstehen so, wie wir die mulitvariate Normalverteilung eingeführt haben. Man ersetzt in 3.2.f einfach die Standard-Normalverteilung durch eine andere Verteilung, deren Dichte nur von $\|\underline{z}\|$ abhängt.

Der grosse Nachteil dieser Modelle ist, dass sie keine unabhängigen Variablen beschreiben können; bereits für die „Standard-Verteilung“ mit $\boldsymbol{\Sigma} = \boldsymbol{I}$ sind die Variablen abhängig – wenn auch unkorreliert!

3.3 Theoretische Resultate für andere Gebiete

a **Verteilung der geschätzten Koeffizienten in der multiplen linearen Regression.** Das Modell der multiplen linearen Regression lautete in Matrix-Schreibweise

$$\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{E} \,.$$

Die Annahmen $E_i \sim \mathcal{N}\langle 0, \sigma^2\rangle$, unabhängig, kann man mit den jetzt bekannten Begriffen schreiben als

$$\underline{E} \sim \mathcal{N}_n\langle \underline{0}, \sigma^2 \boldsymbol{I}\rangle \,.$$

Die Kleinste-Quadrate-Schätzung des Koeffizientenvektors $\underline{\beta}$ war

$$\underline{\widehat{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\underline{Y} \,.$$

Setzen wir das Modell ein! Das ergibt

$$\begin{aligned} \underline{\widehat{\beta}} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T(\boldsymbol{X}\underline{\beta} + \underline{E}) = \beta + \boldsymbol{C}\underline{E} \;\text{ mit} \\ \boldsymbol{C} &= \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T \,. \end{aligned}$$

Das ist eine lineare Transformation des Zufallsvektors $\underline{E}$. Also ist $\underline{\widehat{\beta}}$ multivariat normalverteilt. Der Erwartungswert ist $= \underline{\beta}$, da $\mathcal{E}\langle\underline{E}\rangle = \underline{0}$ ist. Die Kovarianzmatrix ist

$$\operatorname{var}\langle\underline{\widehat{\beta}}\rangle = \boldsymbol{C}(\sigma^2\boldsymbol{I})\boldsymbol{C}^T = \sigma^2 \cdot \boldsymbol{C}\boldsymbol{C}^T$$

$$
\begin{aligned}
&= \sigma^2 \cdot \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \cdot \boldsymbol{X} \left(\left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right)^T \\
&= \sigma^2 \cdot \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} .
\end{aligned}
$$

Ein wichtiges Resultat, das man ohne Zufallsvektoren und Kovarianzmatrizen nur sehr mühsam herleiten kann!

b* **Unabhängigkeit von Mittelwert und empirischer Standardabweichung einer Zufallsvariablen.** In der univariaten Statistik ist die Tatsache nützlich, dass der Mittelwert und die empirische Standardabweichung stochastisch unabhängig sind, wenn die Daten normalverteilt sind. Dies lässt sich mit den vorgängig eingeführten Resultaten recht einfach beweisen.

Es gilt $\underline{X} = [X_1, X_2, ..., X_n]^T \sim \mathcal{N}\langle \mu \underline{1}, \sigma^2 \boldsymbol{I} \rangle$. Der Mittelwert ist gleich $\underline{1}^T \underline{X}/n$ und deshalb bis auf einen Faktor $\sqrt{n}$ gleich $\underline{q}_1^T \underline{X}$ mit $\underline{q}_1 = \underline{1}/\sqrt{n}$. Wir schreiben das so kompliziert, damit wir mit $\underline{q}_1$ ein Vektor der Länge 1 erhalten.

Jetzt ergänzen wir den Vektor $\underline{q}_1^T$ mit $n-1$ weiteren Zeilen zu einer orthogonalen Matrix $\boldsymbol{Q}$. Der Vektor $\underline{Y} = \boldsymbol{Q}(\underline{X} - \mu \underline{1})$ hat die Verteilung $\underline{Y} \sim \mathcal{N}\langle \underline{0}, \sigma^2 \boldsymbol{I} \rangle$, denn $\text{var}\langle \underline{Y} \rangle = \boldsymbol{Q} \sigma^2 \boldsymbol{I} \boldsymbol{Q}^T = \sigma^2 \boldsymbol{I}$. Also sind die Y_k von einander unabhängig. Zudem gilt $\|\underline{Y}\| = \|\underline{X}\|$ und $Y_1 = \sqrt{n}\overline{X}$.

Die empirische Varianz, multipliziert mit $n-1$, lässt sich schreiben als

$$
\sum\nolimits_i X_i^2 - n\overline{X}^2 = \|\underline{X}\|^2 - n\overline{X}^2 = \|\underline{Y}\|^2 - Y_1^2 = \sum\nolimits_{k=2}^{n} Y_k^2
$$

und hängt also nur von $Y_2, ..., Y_n$ ab. Sie ist somit unabhängig von $Y_1 = \sqrt{n}\overline{X}$, also auch von $\overline{X}$.

4 Statistik normalverteilter Daten

4.1 Eine Stichprobe

a **Die drei Grundfragen.** Die schliessende Statistik bildet die Brücke zwischen den Wahrscheinlichkeitsmodellen, die unser Denken strukturieren, und der Wirklichkeit, die wir mit Zahlen oder anderen Daten einzufangen versuchen. Wenn parametrische Modelle zur Beschreibung der Wirklichkeit benützt werden, dann lauten die drei Grundfragen der schliessenden Statistik so:

Drei Grundfragen der Schliessenden Statistik

1. Welcher Wert ist ist für den (jeden) Parameter **am plausibelsten?** Die Antwort führt zur **Schätzung** der Parameter.
2. Ist **ein bestimmter Wert** plausibel? Diese Frage wird durch einen statistischen **Test** beantwortet.
3. **Welche Werte** sind insgesamt plausibel? Die Menge aller Parameterwerte, die plausibel sind (im Sinne eines bestimmten Tests) bildet, wenn nur ein Parameter betrachtet wird, üblicherweise ein Intervall, das **Vertrauens-** oder **Konfidenzintervall**. Im Fall mehrerer Parameter entsteht eine allgemeinere Menge, die üblicherweise zusammenhängend ist und **Vertrauensbereich** genannt wird.

b **Schätzungen** der Parameter $\underline{\mu}$ und $\boldsymbol{\Sigma}$ der Normalverteilung wurden bereits in der beschreibenden Statistik eingeführt: der Mittelwertsvektor $\overline{\underline{X}}$ und die empirische Kovarianzmatrix $\boldsymbol{S}$. Wie im univariaten Fall sind das die gebräuchlichsten Schätzungen und für multivariat normalverteilte Daten die optimalen – aber auch ebenso wenig robust wie diese. Robuste Schätzungen gibt's, aber hier ist nicht der Platz, sie zu behandeln.

Wenden wir uns also den Tests zu!

c **Test für den Erwartungswert.** In der univariaten Statistik lautete eine einfache grundlegende Frage, ob eine Behandlung eine Änderung bewirke, ob also beispielsweise ein Schlafmittel wirklich eine Verlängerung des Schlafes herbeiführe.

Wenn wir jetzt zwei oder mehr Zielgrössen betrachten, also zum Beispiel Blutdruck und Puls, dann können wir wieder fragen, ob eine Veränderung durch ein Medikament statistisch nachweisbar sei. Die Veränderung $\underline{X}$ ist jetzt eine zweidimensionale Grösse, für die wir die Hypothese „keine Veränderung" als $\mathcal{E}\langle\underline{X}\rangle = \underline{0}$ testen können.

Wenn eine Normalverteilung für die Daten vorausgesetzt wird, gilt also $\underline{X}_i \sim \mathcal{N}_m\langle\underline{\mu}, \boldsymbol{\Sigma}\rangle$, und es soll die **Nullhypothese** $\underline{\mu} = \underline{0}$ **getestet** werden.

d Die nahe liegende Teststatistik ist zunächst die Schätzung des Erwartungswertes $\mathcal{E}\langle\underline{X}\rangle$, also der **Mittelwertsvektor** $\overline{\underline{X}}$. Um beurteilen zu können, ob er „zu gross" ist und deshalb die Nullhypothese abgelehnt werden soll, müssen wir seine **Verteilung** unter der Nullhypothese kennen. Das ist nicht schwierig: Wenn wir wie früher $\mathcal{E}\langle\underline{X}_i\rangle = \underline{\mu}$ und $\mathrm{var}\langle\underline{X}_i\rangle = \boldsymbol{\Sigma}$ schreiben, wird nach 3.1.f

$$\begin{aligned}\mathcal{E}\langle\overline{\underline{X}}\rangle &= \tfrac{1}{n}\sum_{i=1}^{n}\mathcal{E}\langle\underline{X}_i\rangle = \tfrac{1}{n}\,n\underline{\mu} = \underline{\mu}\\ \mathrm{var}\langle\overline{\underline{X}}\rangle &= (\tfrac{1}{n})^2\sum_{i=1}^{n}\mathrm{var}\langle\underline{X}_i\rangle = \tfrac{1}{n^2}\,n\mathfrak{\Sigma} = \tfrac{1}{n}\mathfrak{\Sigma}\ .\end{aligned}$$

Setzen wir nun noch die Normalverteilung für die $\underline{X}_i$ voraus, so wird

$$\overline{\underline{X}} \sim \mathcal{N}_m\langle\underline{\mu}, \tfrac{1}{n}\mathfrak{\Sigma}\rangle\ ,$$

wie das in Analogie zum univariaten Fall zu vermuten war.

e **Standardisierung.** Die Analogie legt auch nahe, die Teststatistik durch Standardisierung von den Parametern unabhängig zu machen. Nach 3.1.e und 2.6.m bestimmt man also eine Matrix $\boldsymbol{B}$ mit $\boldsymbol{B}\,\boldsymbol{B}^T = \mathfrak{\Sigma}$, mit der sich sowohl die einzelnen Beobachtungen als auch ihr Mittelwert standardisieren lassen,

$$\underline{Z}_i = \boldsymbol{B}^{-1}(\underline{X}_i - \underline{\mu})\ , \qquad \overline{\underline{Z}} = \boldsymbol{B}^{-1}(\overline{\underline{X}} - \underline{\mu})\ .$$

Jetzt ist $\sqrt{n}\overline{\underline{Z}}$ multivariat standard-normalverteilt.

Nun erhält die Frage, welche Werte von $\overline{\underline{Z}}$ als „zu gross“ gelten sollen, eine natürliche Antwort: Der Annahmebereich wird für $m = 2$ ein Kreis sein, für grössere Dimensionen eine (Hyper-) Kugel, und allgemein gegeben durch $\sqrt{n}\|\overline{\underline{Z}}\| \le c$, wobei c so gewählt wird, dass der Annahmebereich unter der Nullhypothese die Wahrscheinlichkeit 95% erhält. Das gilt gemäss 3.2.m dann, wenn c^2 das 95%-Quantil der Chiquadrat-Verteilung mit m Freiheitsgraden ist.

f Zurück zum Test! Wir wollen die Nullhypothese $\underline{\mu} = \underline{0}$ oder, gleich allgemeiner, eine Nullhypothese der Form $\underline{\mu} = \underline{\mu}_0$ testen. Setzen wir zunächst voraus, dass die Kovarianzmatrix $\mathfrak{\Sigma}$ bekannt sei! Dann kann man das zu den Beobachtungen $\underline{X}_1, ..., \underline{X}_n$ gehörige $\overline{\underline{Z}}$ (mit $\underline{\mu} = \underline{\mu}_0$) berechnen und erhält den Annahmebereich $\sqrt{n}\|\overline{\underline{Z}}\| \le c$.

Es ist für die Anschaulichkeit auch nützlich, diesen Bereich durch $\overline{\underline{X}}$ ausdrücken. Es ist

$$\begin{aligned}\|\overline{\underline{Z}}\|^2 &= \overline{\underline{Z}}^T\overline{\underline{Z}} = (\overline{\underline{X}} - \underline{\mu})^T(\boldsymbol{B}^{-1})^T\boldsymbol{B}^{-1}(\overline{\underline{X}} - \underline{\mu})\\ &= (\overline{\underline{X}} - \underline{\mu})^T\left(\boldsymbol{B}\boldsymbol{B}^T\right)^{-1}(\overline{\underline{X}} - \underline{\mu}) = (\overline{\underline{X}} - \underline{\mu})^T\mathfrak{\Sigma}^{-1}(\overline{\underline{X}} - \underline{\mu})\\ &= d^2\langle\overline{\underline{X}}, \underline{\mu}; \mathfrak{\Sigma}\rangle \le c^2/n\ .\end{aligned}$$

Der Mittelwertsvektor $\overline{\underline{X}}$ soll also, gemessen mit der Mahalanobis-Distanz d, nahe beim zu testenden Erwartungswert $\underline{\mu}_0$ liegen.

g **Studentisierung.** Da man $\mathfrak{\Sigma}$ (normalerweise) nicht kennt, muss man es aus den Daten schätzen. Dadurch wird im univariaten Fall aus der Normalverteilung eine Student'sche t-Verteilung. Im multivariaten Fall wird die Teststatistik

$$T^2 = n\,d^2\langle\overline{\underline{X}}, \underline{\mu}; \boldsymbol{S}\rangle = n\,(\overline{\underline{X}} - \underline{\mu})^T\boldsymbol{S}^{-1}(\overline{\underline{X}} - \underline{\mu})$$

benützt und als T^2 von **Hotelling** bezeichnet, da dieser Mann auch ihre Verteilung gefunden hat: Ein Vielfaches von T^2 zeigt eine F-Verteilung,

$$\frac{n(n-m)}{(n-1)m}\,d^2\langle\overline{\underline{X}}, \underline{\mu}; \boldsymbol{S}\rangle \sim \mathcal{F}\langle m, n-m\rangle\ .$$

h ▷ Als **Beispiel** nehmen wir die vier Exemplare von Iris setosa-Blüten wieder auf (2.5.b). Die Nullhypothese laute $\underline{\mu} = [5,\, 2.5]^T$. Die Mahalanobis-Distanz wird

$$d^2 = [-0.175,\, 0.7] \begin{bmatrix} 0.0492 & 0.0300 \\ 0.0300 & 0.0467 \end{bmatrix}^{-1} \begin{bmatrix} -0.175 \\ 0.7 \end{bmatrix} = 23.6\,.$$

Vergleicht man $4(4-2)/((4-1)2)\,23.6 = 31.4$ mit einer $\mathcal{F}\langle 2,2\rangle$-Verteilung, so erhält man einen P-Wert von 3.1%, also knappe Signifikanz. Die beiden univariaten t-Tests geben keine signifikanten Abweichungen. (Die Nullhypothese wurde zugegebenermassen so gewählt, dass dieser Effekt auftritt, und es versteht sich, dass man nur im äussersten Notfall eine solche Nullhypothese mit nur vier Beobachtungen testen würde.) ◁

i **Vertrauensbereich.** Nachdem bekannt ist, wie man testen soll, kann man die Menge aller Parametervektoren μ^* bestimmen, für die der Test die Nullhypothese $\mu = \mu^*$ nicht verwirft. Dieser Fall tritt ein, wenn

$$d^2\langle \overline{\underline{X}}, \underline{\mu}^*; \boldsymbol{S}\rangle \leq \frac{m(n-1)}{n(n-m)}\, q$$

ist, wobei q das 95%-Quantil der F-Verteilung mit m und $n-m$ Freiheitsgraden ist. Die Vektoren μ^*, die dieser Ungleichung genügen, füllen für $m = 2$ eine **Ellipse** (Abbildung 4.1.i), in höheren Dimensionen ein „Ellipsoid“.

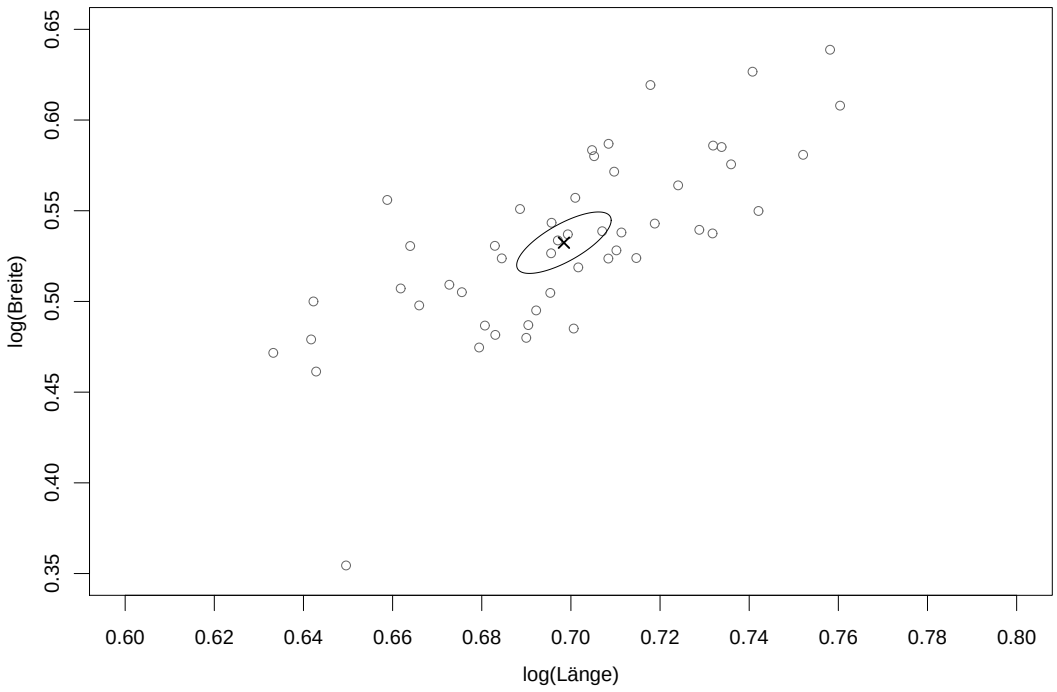


Abbildung 4.1.i: Vertrauensbereich für den Erwartungswert der logarithmierten Länge und Breite der Sepalblätter von Iris setosa-Pflanzen

4.2 Statistik der Kovarianzmatrix

a **Test für Unkorreliertheit.** Eine Hypothese, die in der Praxis oft interessiert, ist diejenige der Unabhängigkeit von zwei Variablen. Unter dieser Hypothese ist die Kovarianz und damit auch die Korrelation gleich null, also $H_0: \; \mathbf{\Sigma}_{jk} = 0$ (mit $j \neq k$) oder $H_0: \; \rho\langle X^{(j)}, X^{(k)}\rangle = 0$.

Als Teststatistik dient die geschätzte Korrelation $\widehat{\rho}_{jk} = S_{jk} \Big/ \sqrt{S_{jj}\, S_{kk}}$ (siehe 2.4.b und 2.5.g). Es kann gezeigt werden, dass $T = \widehat{\rho}_{jk}\sqrt{n-2} \Big/ \sqrt{1-\widehat{\rho}_{jk}^2}$ t-verteilt ist mit $n-2$ Freiheitsgraden, wenn die Beobachtungen gemeinsam normalverteilt sind.

b **Vertrauensintervall für ρ.** Die Verteilung der geschätzten Korrelation ist auch für Korrelationen $\rho_{jk} \neq 0$ bekannt. Sie lässt sich einfacher angeben, wenn man die „z-Transformation“ anwendet: Es sei

$$\zeta = \frac{1}{2}\log_e\left\langle \frac{1+\rho}{1-\rho} \right\rangle \; .$$

Die entsprechend transformierte geschätzte Grösse ist näherungsweise normalverteilt mit konstanter Varianz,

$$\widehat{\zeta} \approx\sim \mathcal{N}\left\langle \zeta, 1/(n-3) \right\rangle \; .$$

Das führt zum Vertrauensintervall $[\, \widehat{\zeta} - 1.96/\sqrt{n-3},\; \widehat{\zeta} + 1.96/\sqrt{n-3}\,]$. Das Intervall für ρ erhält man durch Rücktransformation dieser Grenzen mittels der Formel $\rho = (\alpha-1)/(\alpha+1)$ mit $\alpha = \exp\langle 2\zeta\rangle$. Es ist oft deutlich asymmetrisch bezüglich $\widehat{\rho}$.

c ▷ Im **Beispiel** der **Iris-Blüten** der Art setosa erhält man für die ersten beiden Variablen $S_{11} = 0.1242$, $S_{22} = 0.1437$, $S_{12} = 0.0992$ und daraus $\widehat{\rho} = 0.0992/\sqrt{0.1242 \cdot 0.1437} = 0.743$ und $\widehat{\zeta} = 0.956$. Das Vertrauensintervall für ζ wird $[0.670,\; 1.24]$. Aus den zugehörigen α-Werten 3.82 und 11.99 resultieren für die Korrelation zwischen Länge und Breite der Blütenblätter die Grenzen 0.585 und 0.846, also ein bezüglich $\widehat{\rho} = 0.743$ asymmetrisches Intervall. Dass die Korrelation nicht null sein kann, schliesst man daraus, dass null nicht in diesem Intervall liegt – oder aus dem direkten Test (es wird $T = 0.743 \cdot \sqrt{48}/\sqrt{1-0.743^2} = 7.68$ mit P-Wert 0.000).

All diese Werte sind jedoch etwas vorsichtig zu interpretieren, da die Voraussetzung der Normalverteilung wegen des in Abbildung 2.1.a ersichtlichen Ausreissers kaum als erfüllt betrachtet werden kann. ◁

d **Tests für die Kovarianzmatrix.** Für die Kovarianzmatrix $\mathbf{\Sigma}$ oder Teile von ihr kann man auch allgemeinere Nullhypothesen formulieren. Um die entsprechenden Tests herzuleiten, braucht man die Verteilung der geschätzten Kovarianzmatrix. Setzt man die Normalverteilung für die Beobachtungen voraus, dann ist diese Verteilung unter dem Namen **Wishart-Verteilung** wohlbekannt. Sie hängt, wie man sich leicht überlegt, nur von der wahren Kovarianzmatrix $\mathbf{\Sigma}$ und dem Stichprobenumfang n ab. Statt n verwendet man die Anzahl Freiheitsgrade $n-1$ als Parameter und schreibt

$$\boldsymbol{S} \sim \mathcal{W}\langle \mathbf{\Sigma}, n-1\rangle \; .$$

e* Die Herleitung dieser Verteilung lässt sich vereinfachen, indem man zunächst den Fall $\boldsymbol{\Sigma} = \boldsymbol{I}$ betrachtet. Dann sind die $X^{(j)}$ unabhängig und standard-normalverteilt. Die geschätzten Varianzen S_{jj} sind dann also unabhängig und jedes $(n-1)S_{jj}$ ist chiquadrat-verteilt mit $n-1$ Freiheitsgraden. Über die Verteilung der geschätzten Kovarianzen S_{jk} soll hier nichts weiter ausgeführt werden. (Vielleicht ändert sich das in einer späteren Version.)

So ergibt sich die Standard-Wishart-Verteilung $\mathcal{W}\langle \boldsymbol{I}, n-1\rangle$.

Die Verteilung für allgemeines $\boldsymbol{\Sigma}$ erhält man dadurch, dass man $\boldsymbol{S}$ als lineare Funktion einer Matrix $\boldsymbol{S}^{(0)}$ auffasst, die diese Standard-Wishart-Verteilung aufweist.

4.3 Zwei Stichproben

a Im **Beispiel der Iris-Blüten** ist die Art setosa in den gemessenen Grössen klar von den anderen beiden verschieden. Gibt es auch Unterschiede zwischen versicolor und virginica? Zur Vereinfachung der Darstellung wollen wir uns auf Länge und Breite der Sepalblätter beschränken. Aus den Abbildungen 1.2.b (i) und (ii) kann man vermuten, dass sich die Verteilung der Länge signifikant unterscheidet, während dies für die Breite nicht offensichtlich ist. Wir wollen hier aber nicht die Frage nach einem Unterschied für jede der beiden Variablen getrennt stellen, sondern fragen, ob die **gemeinsame Verteilung** von Länge und Breite für die beiden Arten verschieden sei.

b **Modell.** Für die gemeinsame Verteilung der Variablen innerhalb jeder Gruppe (Art) h bewährt sich als einfachstes Modell wieder die Normalverteilung,

$$\underline{X}_{hi} \sim \mathcal{N}_m\langle \underline{\mu}_h, \boldsymbol{\Sigma}_h\rangle \ , \quad h = 1, 2 \ , \quad i = 1, 2, ..., n_h \ ,$$

und alle *Beobachtungen* sollen stochastisch unabhängig sein – die Variablen hängen so zusammen, wie es die Kovarianzmatrizen $\boldsymbol{\Sigma}_h$ ausdrücken.

Die Unterschiede der Erwartungswerte fassen wir im Vektor $\underline{\Delta}$ zusammen,

$$\underline{\Delta} = \underline{\mu}_2 - \underline{\mu}_1$$

c **Schätzung der Unterschiede.** Natürlich wird $\underline{\Delta}$ durch die Differenz der Mittelwerte $\widehat{\underline{\Delta}} = \overline{\underline{X}}_{2\cdot} - \overline{\underline{X}}_{1\cdot}$ geschätzt. Wie ist diese Grösse verteilt?

Der Erwartungswert ist gleich $\underline{\Delta}$, da ja die Gruppenmittelwerte $\overline{\underline{X}}_{h\cdot}$ erwartungstreu sind für die Erwartungswerte $\underline{\mu}_h$. Da die $\overline{\underline{X}}_{h\cdot}$ unabhängig sind, addieren sich ihre Kovarianzmatrizen

$$\mathrm{var}\langle \widehat{\underline{\Delta}}\rangle = \tfrac{1}{n_1}\boldsymbol{\Sigma}_1 + \tfrac{1}{n_2}\boldsymbol{\Sigma}_2 \ .$$

Die genaue Verteilung ist nur dann einfach herzuleiten und zu beschreiben, wenn wir voraussetzen, dass die Beobachtungen normalverteilt und die Kovarianzmatrizen innerhalb der Gruppen gleich sind, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 =: \boldsymbol{\Sigma}$. Dann gilt

$$\widehat{\underline{\Delta}} \sim \mathcal{N}_m\Big\langle \underline{\Delta} \ , \ \Big(\tfrac{1}{n_1} + \tfrac{1}{n_2}\Big)\,\boldsymbol{\Sigma} \Big\rangle \ .$$

Diese Überlegung liefert die Grundlage für den folgenden Test.

d **Test.** Die Nullhypothese lautet, dass $\underline{\Delta} = \underline{\mu}_2 - \underline{\mu}_1 = \underline{0}$ sei. Man verfährt ganz analog zum univariaten t-Test einerseits und zum Ein-Stichproben-Test andererseits.

Die nahe liegende Teststatistik ist zunächst die Schätzung $\widehat{\underline{\Delta}}$. Sie ist aber mehrdimensional und deshalb als Teststatistik wenig geeignet. Als eindimensionale Teststatistik dient wie im Ein-Stichproben-Fall die Länge der Mittelwertsdifferenz für standardisierte Beobachtungen. Sie kann geschrieben werden als

$$d^2\langle \overline{\underline{X}}_{1.}, \overline{\underline{X}}_{2.}; \mathbb{\Sigma} \rangle = (\overline{\underline{X}}_{2.} - \overline{\underline{X}}_{1.})^T \mathbb{\Sigma}^{-1} (\overline{\underline{X}}_{2.} - \overline{\underline{X}}_{1.})$$

e **Schätzung von $\mathbb{\Sigma}$.** Wieder ist die Kovarianzmatrix $\mathbb{\Sigma}$ normalerweise unbekannt, und man muss sie aus den Beobachtungen schätzen – über den Mittelwert der beiden für die Gruppen geschätzten Kovarianzmatrizen. Wenn die Gruppen verschieden gross sind, gewichtet man am besten mit der Anzahl der Freiheitsgrade, $n_h - 1$, und erhält so

$$\boldsymbol{S} = \tfrac{1}{n-2} \left(\sum\nolimits_{i=1}^{n_1} (\underline{X}_{1i} - \overline{\underline{X}}_{1.})(\underline{X}_{1i} - \overline{\underline{X}}_{1.})^T + \sum\nolimits_{i=1}^{n_2} (\underline{X}_{2i} - \overline{\underline{X}}_{2.})(\underline{X}_{2i} - \overline{\underline{X}}_{2.})^T \right) .$$

Einsetzen führt zur „studentisierten“ Differenz

$$d^2\langle \overline{\underline{X}}_{1.}, \overline{\underline{X}}_{2.}; \boldsymbol{S} \rangle = (\overline{\underline{X}}_{2.} - \overline{\underline{X}}_{1.})^T \boldsymbol{S}^{-1} (\overline{\underline{X}}_{2.} - \overline{\underline{X}}_{1.})$$

als sinnvolle Teststatistik. Sie misst den Abstand der Mittelwerte „in der Metrik der gemeinsamen Kovarianzmatrix“. Wegen ihrer grundlegenden Bedeutung wird diese Grösse auch **Standard-Distanz** zwischen den beiden Stichproben genannt.

f **T^2-Test.** Die zum univariaten t-Test analoge Teststatistik ist $T^2 = d^2\langle \overline{\underline{X}}_{2.}, \overline{\underline{X}}_{1.}; \boldsymbol{S} \rangle / (1/n_1 + 1/n_2)$. Die Verteilung lässt sich aber einfacher angeben für ein Vielfaches von T^2; es gilt

$$\frac{(n-m-1)}{m(n-2)(1/n_1 + 1/n_2)}\, d^2\langle \overline{\underline{X}}_{2.}, \overline{\underline{X}}_{1.}; \boldsymbol{S} \rangle \sim \mathcal{F}\langle m, n-m-1 \rangle ,$$

falls die Nullhypothese $\underline{\mu}_1 = \underline{\mu}_2$ erfüllt ist – und die Beobachtungen dem Normalverteilungsmodell mit gleichen Kovarianzmatrizen folgen.

g ▷ **D**ie beiden Arten virginica und versicolor im **Beispiel der Irisblüten** sind sehr ähnlich in den gemessenen Variablen. Sind Unterschiede vorhanden? Man erhält die folgenden unterschiedlichen Mittelwerte:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
versicolor	5.94	2.77	4.26	1.33
virginica	6.59	2.97	5.55	2.03

Die Unterschiede erweisen sich als über alle Zweifel gesichert, mit einem P-Wert, der als $< 2 \cdot 10^{-16}$ angegeben wird. ◁

4.S S-Funktionen

a Das multivariate **Zwei-Stichproben-Problem** kann, wie das univariate, als Spezialfall einer Regression erkannt werden, mit der zweiwertigen Variablen, die die Gruppenzugehörigkeit wiedergibt, als Eingangsvariable.

b **Regression und Varianzanalyse** mit mehreren Zielgrössen kann mit den gleichen Funktionen wie für eine univariate Zielgrösse durchgeführt werden, indem links vom Tilde-Zeichen $\sim$ eine Matrix angegeben wird. Für das Zwei-Stichproben-Problem sieht das so aus:

```
> t.y  <- as.matrix(iris[,1:4])
> t.r  <- lm(t.y~Species, data=iris, subset=Species!="setosa")
```

Die Funktion `summary(t.r)` zeigt für ein mit `lm` erzeugtes `t.r` aber nur die Resultate der separaten Regressionen der einzelnen Zielvariablen.

c Den multivariaten Test von Hotelling für die Unterschiede erhält man von der Funktion `manova`

```
> t.r  <- manova(t.y Species, data=iris, subset=Species!="setosa")
> summary(t.r)
```

(Die Funktion `manova` ruft im Wesentlichen `aov` auf und setzt als Klasse des Resultats `manova`, damit `summary` dann die geeigneten Resultate zusammenstellt.)

5 Diskriminanz-Analyse

5.1 Einleitung

a ▷ **Beipiel Iris-Arten.** In Kapitel 3.3.b wurde nachgewiesen, dass zwischen den beiden ähnlichen Arten versicolor und virginica signifikante Unterschiede bestehen. Das Ziel ging aber weiter (1.2.a): Es wird nach einer Regel gesucht, die es erlaubt, die Pflanzen möglichst fehlerfrei den drei Arten zuzuordnen. Für den vorhandenen Datensatz ist die Zuordnung bekannt. Die Regel wird also nicht für diese Pflanzen gebraucht, sondern soll es erlauben, weitere **Pflanzen allein auf Grund der Messungen richtig zuzuordnen.** ◁

b **Das allgemeine Modell.** Jede Beobachtungseinheit i ist charakterisiert durch ihre Zugehörigkeit zu einer Klasse k_i und durch die Werte $X_i^{(j)}$ der Variablen. Die Klasse k_i bestimmt die Verteilung des Zufallsvektors $\underline{X}_i$, die wir allgemein mit $\mathcal{F}_{k_i}$ bezeichnen wollen,

$$\underline{X}_i \sim \mathcal{F}_{k_i} \ .$$

(Im Gegensatz zu früher wird die Klassenzugehörigkeit hier nicht mit einem doppelten Index hi für die $i$te Beobachtung in der Gruppe h angegeben, sondern mit der kategoriellen Variablen k_i.) Die Verteilungen $\mathcal{F}_k$, die die Klassen charakterisieren, werden üblicherweise als parametrische Verteilungen angesetzt; wir werden den Fall von Normalverteilungen näher betrachten.

c In einer ersten Variante des Modells wird die **Klassenzugehörigkeit** k_i als **feste, unbekannte Zahl** aufgefasst. Solche Zahlen haben wir bisher jeweils als Parameter bezeichnet. Die Situation ist aber eine andere als bei parametrischen Modellen im üblichen Sinn, da für jede Beobachtung ein neuer Parameter dazukommt. Man nennt solche Grössen englisch **incidental parameters**.

d In einer zweiten Version des Modells ist die **Klassenzugehörigkeit** selbst eine **Zufallsvariable** K_i, und das Modell legt auch die g Wahrscheinlichkeiten $P\langle K_i = k\rangle =: \pi_k$ der Zugehörigkeiten zu den g Gruppen fest. Die Verteilungen $\mathcal{F}_k$ sind dann die bedingten Verteilungen von $\underline{X}_i$, gegeben $K_i = k$.

e Wenn die Verteilungen $\mathcal{F}_k$ und, in der zweiten Variante, die Wahrscheinlichkeiten π_k bekannt sind, kann man daraus Regeln ableiten, um **aus den Werten von $\underline{X}$ auf die Klassenzugehörigkeit k zurückschliessen zu können.** Diese Aufgabe wird auch als **Identifikations-Analyse** bezeichnet. In der Anwendung wird man die Parameter der Verteilungen $\mathcal{F}_k$ meistens aus Daten schätzen müssen. Dazu braucht es einen Datensatz, für den neben den Variablenwerten $\underline{x}_i$ auch die Klassenzugehörigkeiten k_i bekannt sind. Dieser Datensatz wird **Trainings-Datensatz** genannt.

f Wir wollen hier die Ideen wieder anhand des einfachsten üblichen Modells entwickeln, das Klassen mit multivariat normalverteilten Daten $\underline{X}_i$ umfasst, die sich nur im Erwartungswert $\underline{\mu}_k$ unterscheiden, also

$$\underline{X}_i \sim \mathcal{N}_m\langle \underline{\mu}_{k_i}, \boldsymbol{\Sigma} \rangle \ .$$

Wie die Parameter $\underline{\mu}_k$ und $\boldsymbol{\Sigma}$ aus den Trainingsdaten geschätzt werden sollen, wurde für $g = 2$ im Zusammenhang mit dem Zwei-Stichproben-Problem schon gesagt (4.3.c und 4.3.e). Allgemein werden die Erwartungswerte $\underline{\mu}_k$ durch die Mittelwerte

$$\overline{\underline{X}}_k = \tfrac{1}{n_k} \sum\nolimits_{\{i|k_i=k\}} \underline{X}_i$$

geschätzt. Die Schätzung von $\boldsymbol{\Sigma}$ wird

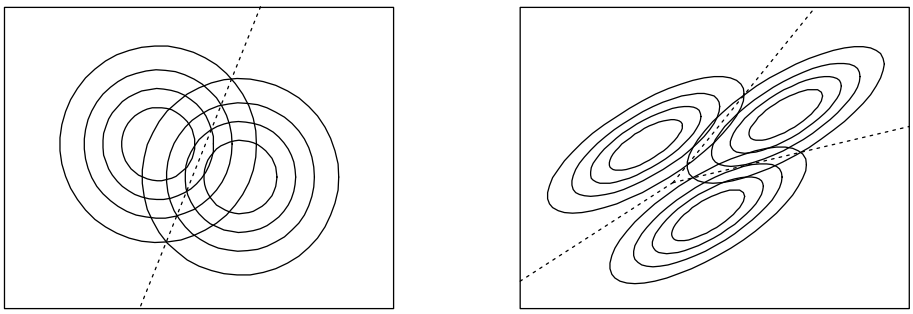
$$\widehat{\boldsymbol{\Sigma}} = \tfrac{1}{n-g} \sum\nolimits_{k=1}^{g} \sum\nolimits_{\{i|k_i=k\}} (\underline{X}_i - \overline{\underline{X}}_{k.})(\underline{X}_i - \overline{\underline{X}}_{k.})^T = \tfrac{1}{n-g} \sum\nolimits_{i} (\underline{X}_i - \overline{\underline{X}}_{k_i.})(\underline{X}_i - \overline{\underline{X}}_{k_i.})^T \ .$$

Wir können uns also sofort um die Identifikations-Analyse kümmern. Für diese Betrachtung nehmen wir die Parameter als fest vorgegeben an. Welche Konsequenzen die Ungenauigkeit der Schätzung der Parameter haben, studieren wir später.

5.2 Klassierung bei bekannten Verteilungen

a Gegeben sei eine Beobachtung $\underline{x}_0$, für die die Klassenzugehörigkeit k_0 nicht bekannt, sondern zu bestimmen sei. Auf Grund der beobachteten Merkmale soll also eine **Entscheidung** zwischen g möglichen Zuordnungen getroffen werden. Man kann auch sagen, dass der „diskrete Parameter“ k_0 aus der Beobachtung $\underline{x}_0$ **zu schätzen** sei.

b Im einfachsten Fall liegen Klassen mit multivariater Normalverteilung und gleichen Kovarianzmatrizen, aber verschiedenen Erwartungswerten vor, $\underline{X}_0 \sim \mathcal{N}\langle \underline{\mu}_k, \boldsymbol{\Sigma} \rangle$. Abbildung 5.2.b veranschaulicht dieses Modell für $\boldsymbol{\Sigma} = \boldsymbol{I}$ und allgemeines $\boldsymbol{\Sigma}$.



(i) (ii)

Abbildung 5.2.b: Modelle (i) für zwei Gruppen mit $\boldsymbol{\Sigma} = \boldsymbol{I}$ und (ii) für drei Gruppen mit allgemeiner Kovarianz-Matrix

Im ersten Fall gibt es eine „natürlichste“ Zuordnungsregel: Die Beobachtung $\underline{x}_0$ wird der Klasse zugeordnet, für die der Abstand zum „Klassenzentrum“ $\underline{\mu}_k$ am kleinsten ist. Für allgemeines $\mathbf{\Sigma}$ liegt es nahe, den Abstand durch die Mahalanobis-Distanz $d\langle \underline{x}_0, \underline{\mu}_k; \mathbf{\Sigma} \rangle$ zu messen.

c **Zwei Klassen, gleiche $\mathbf{\Sigma}$.** Im Falle von zwei Klassen wird die Regel besonders einfach. Eine Beobachtung wird dann der zweiten Klasse zugeordnet, wenn die Differenz der quadrierten Abstände $d^2\langle \underline{x}_0, \underline{\mu}_1; \mathbf{\Sigma} \rangle - d^2\langle \underline{x}_0, \underline{\mu}_2; \mathbf{\Sigma} \rangle$ positiv ist. Die Differenz lässt sich schreiben als

$$\begin{aligned} d^2\langle \underline{x}_0, \underline{\mu}_1; \mathbf{\Sigma} \rangle - d^2\langle \underline{x}_0, \underline{\mu}_2; \mathbf{\Sigma} \rangle &= (\underline{x}_0 - \underline{\mu}_1)^T \mathbf{\Sigma}^{-1} (\underline{x}_0 - \underline{\mu}_1) - (\underline{x}_0 - \underline{\mu}_2)^T \mathbf{\Sigma}^{-1} (\underline{x}_0 - \underline{\mu}_2) \\ &= 2(\underline{\mu}_2 - \underline{\mu}_1)^T \mathbf{\Sigma}^{-1} \underline{x}_0 + \underline{\mu}_1^T \mathbf{\Sigma}^{-1} \underline{\mu}_1 - \underline{\mu}_2^T \mathbf{\Sigma}^{-1} \underline{\mu}_2 \end{aligned}$$

Die letzten beiden Terme hängen nicht von $\underline{x}_0$ ab, bilden also zusammen eine Konstante. Der erste ist von der Form $2\underline{\beta}^T \underline{x}_0$.

Zusammen bilden diese beiden Teile also eine lineare Funktion

$$h\langle \underline{x}_0 \rangle = \alpha + \underline{\beta}^T \underline{x}_0 \ ,$$

die zur Unterscheidung zwischen den Klassen dient nach der Regel

$$\widehat{k}\langle \underline{x}_0 \rangle = \begin{cases} 1 & \text{falls } \ h\langle \underline{x}_0 \rangle < 0 \\ 2 & \text{falls } \ h\langle \underline{x}_0 \rangle > 0 \end{cases} \ .$$

Die Funktion h wird **Diskriminanz-Funktion** genannt – genauer: die **lineare** Diskriminanz-Funktion von Fisher, da sie von diesem Gentleman gefunden wurde.

d Für die praktische Anwendung muss man nun noch die Parameter schätzen, und zwar aus Trainingsdaten, für die die Klasse bekannt ist (5.1.e). Verwendet man die oben (5.1.f) genannten Schätzungen, dann erhält man **Fishers lineare Diskriminanzanalyse** für zwei Gruppen.

e ▷ Abbildung 5.2.e zeigt, dass die Werte der Diskriminanzfunktion mit geschätzten Parametern eine recht gute Trennung der beiden ähnlichen Arten im Iris-Beispiel erlaubt. Auf die Fehl-Klassierungen kommen wir gleich zurück. ◁

f **Logistische Regression.** Die lineare Diskriminanzfunktion erinnert an ein lineares Regressionsmodell. In der Regression wollte man eine kontinuierliche Zielgrösse Y aus den Werten der erklärenden Variablen $\underline{x}$ „vorhersagen“ können. Hier möchten wir die Klassenzugehörigkeit k bestimmen. Der Unterschied besteht nur darin, dass die Zielgrösse nun nicht mehr kontinuierlich, sondern zweiwertig oder binär ist. Genau genommen haben wir sie in diesem Abschnitt bisher auch nicht als Zufallsvariable behandelt. Das wollen wir jetzt nachholen.

Eine binäre Zufallsvariable Y mit den möglichen Werten 0 und 1 ist charakterisiert durch die Wahrscheinlichkeit $\pi = P\langle Y = 1 \rangle$. In unserem Zusammenhang soll diese Wahrscheinlichkeit mit $\underline{x}$ zusammenhängen. Die Wahrscheinlichkeit, dass eine Beobachtung mit Merkmalswerten $\underline{x}$ zur Klasse 2 gehört – was $Y = 1$ bedeuten soll – wird grösser, wenn die Beobachtung näher beim Mittelpunkt $\underline{\mu}_2$ der Gruppe 2 und weiter vom Zentrum $\underline{\mu}_1$ der Gruppe 1 entfernt liegt. Es ist naheliegend, die Wahrscheinlich-

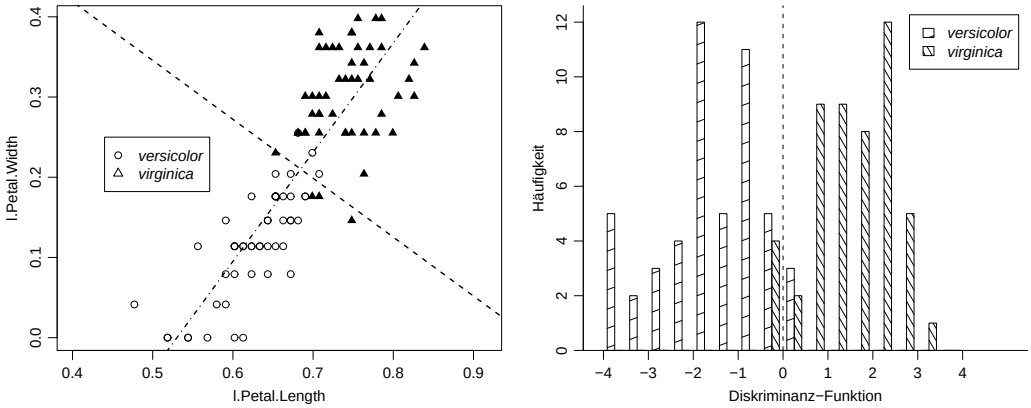


Abbildung 5.2.e: Werte der Diskriminanzfunktion mit geschätzten Parametern für zwei Iris-Arten

keiten für $Y=1$ und $Y=0$ proportional zu den entsprechenden Wahrscheinlichkeits-Dichten $f_k\langle \underline{x} \rangle$ ($k=2$ resp. $=1$) zu setzen. Dann wird

$$\log \left\langle \frac{P\langle Y=1\rangle}{P\langle Y=0\rangle} \right\rangle = \log \left\langle \frac{f_2(\underline{x}\rangle}{f_1(\underline{x}\rangle} \right\rangle = h\langle \underline{x} \rangle \ .$$

Im Fall der Normalverteilung mit gleichen Kovarianzmatrizen erhält man gerade die Differenz der Mahalanobis-Abstände, die gemäss 5.2.c eine lineare Funktion $\alpha + \underline{\beta}^T \underline{x}$ von $\underline{x}$ ist.

Das ist das Modell der **logistischen Regression**. Ein Verhältnis „Wahrscheinlichkeit : Gegen-Wahrscheinlichkeit“ wird auch als **Wett-Verhältnis** bezeichnet; im Englischen gehört das entsprechende Wort **odds** zum täglichen Sprachgebrauch – und vor allem zum sonntäglichen der Leute, die an Pferdewetten teilnehmen. Das Modell der logistischen Regression drückt die „log odds“ als lineare Funktion mit den Ausgangsgrössen oder Regressoren $X^{(j)}$ aus.

g Die Methodik der logistischen Regression führt zu einer direkten Schätzung von α und $\underline{\beta}$, die nicht auf der Schätzung der Parameter $\underline{\mu}_k$ und $\mathbf{\Sigma}$ beruht. Im Modell werden nämlich wie in der klassischen multiplen linearen Regression **keine Annahmen über die Regressoren** $X^{(j)}$ gemacht. Die starke Annahme, dass die Daten der beiden Klassen einer multivariaten Normalverteilung, gar noch mit gleichen Kovarianzmatrizen, folgen sollten, wird also nicht verwendet. Gebraucht wird lediglich die Annahme, dass die „log odds“ eine lineare Funktion der erklärenden Variablen $X^{(j)}$ seien; aus der linearen Regression wissen wir aber, dass mit diesem Ansatz viele Zusammenhänge modelliert werden können, die zunächst gar nicht linear aussehen. Stichworte waren **Transformationen** und Wechselwirkungen.

In der Praxis ist deshalb **die logistische Regression der linearen Diskriminanzanalyse vorzuziehen.**

h **Mehrere Klassen.** Im Falle mehrerer Klassen mit multivariat normalverteilten Merkmalen $X^{(j)}$ und gleicher Kovarianzmatrix $\mathbf{\Sigma}$ wird man, wie erwähnt (5.2.b) sucht man die Klasse mit dem minimalen quadrierten Mahalanobis-Abstand
$d^2\langle \underline{x}_0, \underline{\mu}_k; \mathbf{\Sigma}\rangle = (\underline{x}_0 - \underline{\mu}_k)^T \mathbf{\Sigma}^{-1}(\underline{x}_0 - \underline{\mu}_k)$. Bei zwei Klassen liess sich die Klassierungsregel mit der linearen Diskriminanzfunktion h ausdrücken. Lässt sich bei mehr Klassen etwas Analoges finden?

Am einfachsten liegt der Fall, wie früher, wenn $\mathbf{\Sigma} = \boldsymbol{I}$ gilt. Der Mahalanobis-Abstand ist dann die gewöhnliche (Euklidische) Distanz. Bei dreidimensionalen Daten und drei Gruppen hilft die geometrische Vorstellung zur Veranschaulichung. Um festzustellen, zu welcher von drei Klassenzentren ein Punkt $\underline{x}_0$ am nächsten liegt, können wir den Punkt auf die Ebene projizieren, die durch die drei Zentren geht. Wie weit der Punkt von der Ebene weg ist, hat keinen Einfluss darauf, zu welcher Klasse er zugeordnet wird – ganz bedeutungslos ist er für die Zuordnung trotzdem nicht, wie wir anschliessend besprechen werden. – Es genügen also bei drei Gruppen zwei Dimensionen, um die Zuordnung zu bestimmen. Das bleibt auch so, wenn die Beobachtungen mehr als dreidimensional sind. Bei g Klassen liegen die g Zentren in einem (höchstens) $g-1$-dimensionalen Unterraum (* Nebenraum), und dieser Raum genügt für die Zuordnung. (Wenn weniger als g Variable vorliegen, wird der ganze m-dimensionale Raum benötigt.)

Ein solcher Unterraum ist durch $g-1$ Vektoren festgelegt. Man kann beispielsweise die Vektoren $\underline{\mu}_k - \underline{\mu}_1$, $k = 2, 3, ..., g$ wählen. Besser geeignet sind Vektoren, die senkrecht aufeinander stehen (orthogonal sind), da dann die Distanzen, die für die Klassierung entscheidend sind, erhalten bleiben. Wenn man sie als Basis des Raumes benützt, das heisst, wenn man die Beobachtungen auf solche Vektoren projiziert und die Projektionen als Koordinaten benützt, dann kann deshalb für die Klassen-Zuordnung wieder die gewöhnliche Distanz im Unterraum benützt werden.

Den Fall einer allgemeinen Kovarianzmatrix $\mathbf{\Sigma} \neq \boldsymbol{I}$ führen wir wie früher auf den gerade besprochenen Spezialfall zurück, indem wir Beobachtungen $\underline{X}$ zuerst mit einer Matrix $\boldsymbol{B}$ so transformieren, dass für die transformierten Beobachtungen die Kovarianzmatrix $= \boldsymbol{I}$ wird.

i Zusammenfassend gibt es bei g Klassen $g-1$ **Diskriminanz-Funktionen**, die den Raum festlegen, in dem die Klassen-Zugehörigkeit bestimmt wird.

Sie sind durch diese wesentliche Eigenschaft allerdings noch nicht eindeutig festgelegt, denn man kann sie innerhalb des Unterraumes beliebig orthogonal transformieren. Die Konvention legt sie so fest, dass sie die Hauptkomponenten der Klassen-Zentren im Unterraum bilden.

j ▷ Im **Beispiel der Iris-Arten** sind gesamthaft 3 Arten zu unterscheiden. Das führt zu 2 Diskriminanzfunktionen, deren Werte in Abbildung 5.2.j gezeigt werden. Die Diskriminanzfunktionen werden so gewählt, dass die geschätzte Kovarianzmatrix innerhalb der Klassen (analog zu 5.1.f) die Einheitsmatrix wird. Deshalb sind die Trennlinien zwischen den Klassen durch die Mittelsenkrechten zwischen den Zentren gegeben. ◁

Die Koeffizienten $\underline{\widehat{\beta}}$ der Diskriminanzfunktionen sind

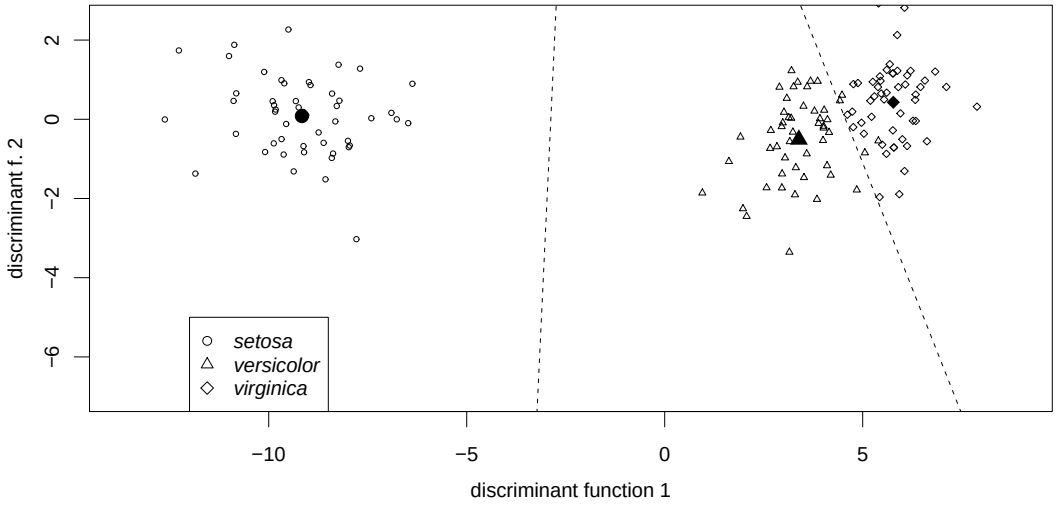


Abbildung 5.2.j: Die beiden Diskriminanzfunktionen im Beispiel der Iris-Arten. Die Mittelwerte der Gruppen sind mit ausgefüllten Symbolen gezeichnet.

	Sepalblätter		Petalblätter	
D.f. 1	8.70	9.07	-20.779	-3.529
D.f. 2	-9.85	-15.18	-0.713	0.313

Die erste Diskriminanzfunktion bildet im Wesentlichen einen Kontrast zwischen den Längenmassen der Sepalblätter und der Breite der Petalblätter, die zweite zwischen der Breite der Sepalblätter und den beiden Längenmassen der Petalblätter.

k **Multinomiale Regression.** Wie der Fall von zwei Klassen auf die logistische, so führt die Diskriminanz-Analyse für mehrere Klassen auf die multinomiale Regression, die wieder ohne Annahmen über die Verteilung der Variablen $X^{(j)}$ verzichtet und deshalb in der Praxis vorzuziehen ist.

l **Modellwahl.** In der linearen Regression sind folgende Schritte unter dem Stichwort **Residuenanalyse** wichtig:

- Modellannahmen prüfen,
- Ausgangsvariable (oder erklärende Variable) „auslesen“,
- nichtlineare Abhängigkeiten und Wechselwirkungen modellieren.

In der hier besprochenen linearen Diskriminanzanalyse muss die Annahme der multivariaten Normalverteilung geprüft werden. Transformationen können helfen, Abweichungen möglichst zu vermeiden. Es ist auch für die Diskriminanzanalyse sinnvoll, Variable auszulesen. Generell ist also Modellwahl auch hier angesagt.

Noch ähnlicher zur gewöhnlichen linearen Regression werden diese Fragestellungen, wenn statt der linearen Diskriminanzanalyse die **logistische Regression** respektive die multinomiale verwendet wird. Man braucht dann, wie erwähnt, keine multivariate Normalverteilung der Variablen $X^{(j)}$ und ist damit in der Modellwahl so frei wie in der gewöhnlichen linearen Regression. Wir gehen deshalb hier auf dieses Kapitel nicht näher ein.

m **Allgemeine Verteilung.** Wie soll die Idee der linearen Diskriminanzanalyse für **andere Verteilungsannahmen** verallgemeinert werden? Um k_0 zu schätzen, können wir auf die Idee der maximalen Likelihood zurückgreifen. Wenn also die $\mathcal{F}_k$ eine Dichte f_k haben, soll $\widehat{k}_0$ das k mit der maximalen Dichte $f_k\langle \underline{x}_0\rangle$ sein,

$$\widehat{k}_0 = \text{argmax}_k \left\langle f_k\langle \underline{x}_0\rangle \right\rangle \ .$$

n Für den Fall multivariat normalverteilter Daten mit **ungleichen Kovarianzmatrizen** $\mathbf{\Sigma}_k$ ergibt sich daraus eine einfache Regel: Die Grösse $-2\log\langle f_k\rangle$ ist gleich dem quadrierten Mahalanobis-Abstand plus einem Term, der von $\mathbf{\Sigma}_k$ abhängt. Statt f_k zu maximieren, bestimmt man

$$\arg\min_k \left\langle d^2\langle \underline{x}_0, \underline{\mu}_k; \mathbf{\Sigma}_k\rangle + \log\langle\det\langle \mathbf{\Sigma}_k\rangle\rangle \right\rangle \ .$$

Wie im Fall gleicher Kovarianzmatrizen ist es nützlich, zu veranschaulichen, wie im Fall von $g = 2$ Klassen die Gebiete aussehen, in denen eine Beobachtung $\underline{x}_0$ der einen oder der anderen Klasse zugeordnet wird. Die Grenze ist gegeben durch

$$\begin{aligned} d^2\langle \underline{x}, \underline{\mu}_2; \mathbf{\Sigma}_2\rangle - d^2\langle \underline{x}, \underline{\mu}_1; \mathbf{\Sigma}_1\rangle - c &= \\ (\underline{x} - \underline{\mu}_2)^T \mathbf{\Sigma}_2^{-1}(\underline{x} - \underline{\mu}_2) - (\underline{x} - \underline{\mu}_1)^T \mathbf{\Sigma}_1^{-1}(\underline{x} - \underline{\mu}_1) - c &= 0 \end{aligned}$$

mit $2c = \log\langle\det\langle \mathbf{\Sigma}_2\rangle\rangle - \log\langle\det\langle \mathbf{\Sigma}_1\rangle\rangle$. Dies kann man analog zur Rechnung in 5.2.f ausdrücken in der Form $\underline{x}^T(\mathbf{\Sigma}_1 - \mathbf{\Sigma}_2)\underline{x} + \underline{\beta}^T\underline{x} + \alpha = 0$, also als quadratische Gleichung in $\underline{x}$. Man nennt daher diese Zuordnungs-Methode **quadratische Diskriminanzanalyse**.

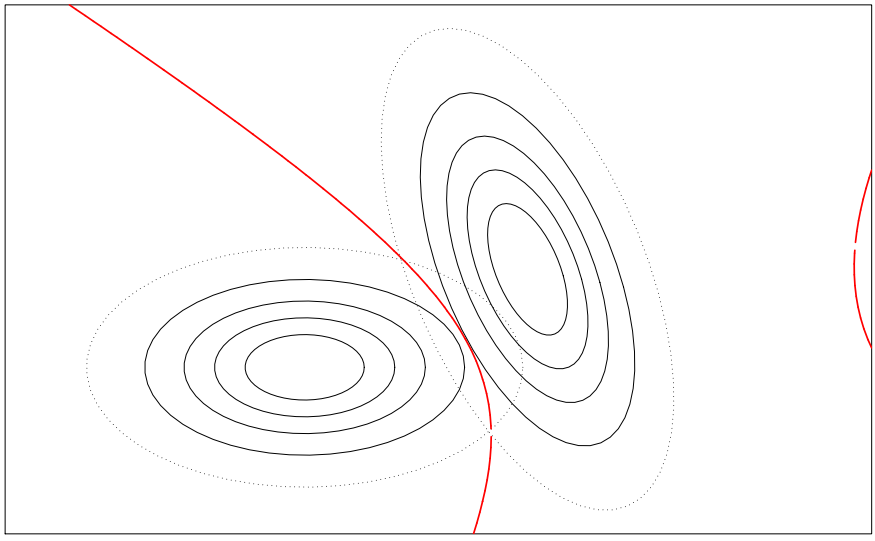


Abbildung 5.2.n: Zwei Gruppen mit ungleichen Kovarianzmatrizen, mit (nur näherungsweise berechneter) quadratischer Trennlinie der Zuordnung

In zwei Dimensionen ist die Trennlinie zwischen den Gebieten der Zuordnung zu den beiden Gruppen eine Hyperbel (Abbildung 5.2.n). Der Kurvenast am rechten Rand der Abbildung zeigt, dass weit rechts die Dichte für die „Gruppe links“ grösser wird als für diejenige mit näher gelegenem Zentrum. Für eine sinnvolle Zuordnung ist das von fraglichem Wert. Beide Dichten sind dort so klein, dass eine Beobachtung, die dort verzeichnet würde, für beide Gruppen sehr unwahrscheinlich ist.

o In der Praxis muss man, wie vorher, die Parameter schätzen. Da es nun für jede Klasse eine ganze Kovarianzmatrix $\boldsymbol{\Sigma}_k$ zu schätzen gibt, ist dies nur sinnvoll, wenn man viele Trainingsdaten hat. Deshalb ist die lineare Diskriminanzanalyse oft besser, auch wenn die Annahme der gleichen Kovarianzmatrizen nicht gerechtfertigt ist, sofern sich die Kovarianzmatrizen nicht allzu sehr unterscheiden.

5.3 Fehlerraten

a In der Einleitung (1.2.c) wurde von diagnostischen Tests in der Medizin gesprochen, die dazu dienen, die Patienten im Hinblick auf eine bestimmte Krankheit in Kranke und Gesunde einzuteilen. Es gibt in diesem Fall zwei Arten von Fehlern mit klar unterschiedlicher Tragweite: Wenn Gesunde vom Test als krank klassiert werden, führt dies zu unnötiger Verunsicherung und, falls keine präziseren Abklärungen den Irrtum zeigen können, zu einer unnötigen Behandlung. Dies ist meistens weniger schlimm als der umgekehrte Fehler: Wenn ein Kranker für gesund erklärt wird, kann eine lebenswichtige Behandlung verpasst werden. Es macht also wenig Sinn, die beiden Fehler zusammenzuzählen und eine einzige „Fehlerrate“ zu bestimmen.

b **Sensitivität und Spezifizität.** Sie kennen vermutlich den verwirrenden medizinischen Jargon: Wenn ein Test Sie als krank klassiert, dann wird das als *positives* Testresultat bezeichnet. Die Gesunden, die als krank eingestuft werden, heissen deshalb **fälschlich Positive**, und die Kranken, die für gesund erklärt werden, sind die **fälschlich Negativen**. Es gibt zwei sinnvolle Arten, ihre Anzahlen als Anteile auszudrücken: Bezieht man sich auf die Gesamtzahl der Positiven respektive Negativen, dann erhält man die

$$\begin{aligned}\text{Rate der fälschlich Positiven} &= \frac{\text{Anzahl der fälschlich Positiven}}{\text{Anzahl der Positiven}}\\ \text{Rate der fälschlich Negativen} &= \frac{\text{Anzahl der fälschlich Negativen}}{\text{Anzahl der Negativen}}\end{aligned}$$

Für Patienten, die ein „positives“ Resultat erhalten, bedeutet die erste Rate die **bedingte Wahrscheinlichkeit**, trotzdem gesund zu sein.

Für die Beurteilung einer Methode ist es aber aussagekräftiger, die Ergebnisse auf den wahren Status der Patienten zu beziehen. Man betont die richtigen Ergebnisse und wählt die Bezeichnungen

$$\begin{aligned}\text{Sensitivität} &= \frac{\text{Anzahl der kranken Positiven}}{\text{Anzahl der Kranken}}\\ \text{Spezifizität} &= \frac{\text{Anzahl der gesunden Negativen}}{\text{Anzahl der Gesunden}}\end{aligned}$$

Die Sensitivität misst die (bedingte) Wahrscheinlichkeit, dass ein Kranker als solcher klassiert wird, während die Spezifizität die (bedingte) Wahrscheinlichkeit angibt, dass ein Gesunder keinen Fehlalarm erhält. Diese beiden Werte charakterisieren die „Trennschärfe“ des medizinischen Tests. Sie werden durch den Anteil der Kranken in der betrachteten Grundgesamtheit, die „**Prävalenz**“, nicht beeinflusst. Dagegen wird die Rate der fälschlich Positiven kleiner, wenn die Prävalenz steigt, während die Rate der fälschlich Negativen steigt.

c **Variable Grenze.** Für den Fall von 2 Klassen bestimmen die lineare Diskriminanzanalyse und die logistische Regression eine lineare Diskriminanz-Funktion $h\langle \underline{x} \rangle = \widehat{\alpha} + \underline{\widehat{\beta}}^T \underline{x}$. Die Klassierung erfolgt durch Vergleich mit einer Konstanten c, die gemäss 5.2.c gleich 0 ist. Es gibt gute Gründe, die Grenze anders zu setzen,

- wenn eine der Klassen häufiger ist als die andere und
- wenn Kosten, die mit einer falschen Klassifizierung verbunden sind, verschieden ausfallen: Einen Kranken irrtümlicherweise als gesund zu erklären, kann fatal sein, wogegen die Verunsicherung von Gesunden, die man (bis zu einer genaueren Untersuchung) als krank diagnostiziert, weniger ins Gewicht fällt. (Wenn Therapien so oder so nicht sicher fruchten, kann die Beurteilung auch umgekehrt ausfallen.)

Solche Überlegungen werden in der **Entscheidungs-Theorie** präzisiert, siehe Abschnitt 5.4.

Oft ist es am einfachsten, nur die Werte der Diskriminanz-Funktion zu bestimmen und die Grenze nach pragmatischen Gesichtspunkten der Anwendung festzulegen. Beispielsweise kann bei einer Werbekampagne der Umfang eines Versandes zum vornherein festliegen, und man wird die entsprechende Anzahl der Adressaten mit den höchsten Scores bedienen (oder belästigen). Bei der Wahl der Grenze wird man oft auf die Anzahl zu erwartender Fehl-Klassierungen abstellen, die wir im Folgenden studieren wollen.

d **Sensitivitäts- und Spezifizitäts-Kurven.** Durch die Wahl des Grenzwertes c für die Diskriminanz-Funktion kann die Sensitivität beliebig erhöhen auf Kosten der Spezifizität erhöht werden – im Extremfall wird man alle als krank erklären, um sicher keinen zu verpassen! Umgekehrt kann man durch Erhöhung von c die Spezifizität erhöhen. Man wird dann die Verunsicherung von Gesunden vermeiden, wird aber die Erkennung von Kranken verpassen. Es ist sinnvoll, die beiden Grössen als Funktion der Grenze c zu betrachten. Abbildung 5.3.d zeigt dies für das Beispiel der Ader-Verengung.

Die Sensitivitäts- und Spezifizitäts-Kurven bilden eine sehr informative Charakterisierung einer Methode. Sie erlauben es,

- die Grenze auf Grund eines sinnvollen Kompromisses zwischen den beiden Kriterien zu wählen,
- die Entscheidung zu verfeinern: Man kann zwei Grenzen c_0 und c_1 einführen, zwischen denen ein Bereich liegt, in dem weitere Abklärungen vorgenommen werden müssen. Es wird dann nur für Werte $< c_0$ auf „gesund“ und für Werte $> c_1$ auf „krank“ geschlossen.

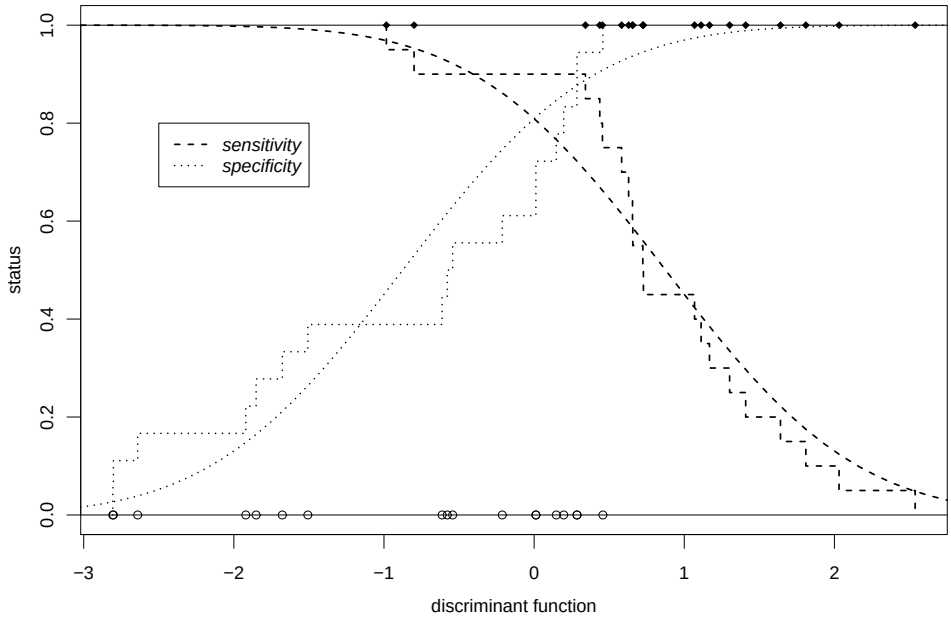


Abbildung 5.3.d: Sensitivität und Spezifizität im Beispiel der Ader-Verengung. Die Treppenkurven stellen die empirischen Fehlerraten dar, die glatten Kurven stehen für die dem geschätzten Modell entsprechenden Raten.

e **Fehlerraten.** Zur Bestimmung der Sensitivität und der Spezifizität – oder der entsprechenden Fehlerraten – gibt es mehrere Wege, die im Folgenden erörtert werden. Bleiben wir beim Fall von zwei Klassen, und betrachten wir eine feste Zuordnungsregel $\widehat{k}\langle \underline{x} \rangle$, also beispielsweise die lineare Diskriminanzfunktion mit einer festen Grenze c. Die Fehler-Wahrscheinlichkeit, die der Sensitivität entspricht, ist die Wahrscheinlichkeit

$$Q_{k\ell} = P\Big\langle \widehat{k}\langle \underline{X} \rangle = \ell \mid \underline{X} \sim F\langle \underline{\theta}_k \rangle \Big\rangle$$

mit $k = 1$ und $\ell = 2$, und umgekehrt für die Spezifizität.

f **Empirische Fehlerrate** *(Appearant error rate).* Die einfachste Schätzung der Fehlerrate besteht aus der relativen Fehlerhäufigkeit in den Trainingsdaten,

$$Q_{k\ell}^{app} = \#\{i \mid \widehat{k}\langle \underline{X}_i \rangle = \ell \ , \ k_i = k\} \ / \ n_k \ ,$$

wobei n_k die Anzahl Beobachtungen in der Gruppe k, also mit $k_i = k$, ist (und $\#\{i \mid ...\}$ die *Anzahl* i ist, für die ... gilt).

Man kann für eine gesamthafte Beurteilung der Regel die beiden Fehlerraten Q_{12}^{app} und Q_{21}^{app} zusammenzählen. Tut man das mit Gewichten, die der Anzahl Beobachtungen in den beiden Klassen entsprechen, so erhält man die einfache Gesamtfehlerrate

$$Q^{app} = (\#\{i \mid \widehat{k}\langle \underline{X}_i \rangle = 2 \ , \ k_i = 1\} + \#\{i \mid \widehat{k}\langle \underline{X}_i \rangle = 1 \ , \ k_i = 2\})/n \ .$$

g **Theoretische Fehlerrate.** Wenn man die Verteilungen $F\langle\underline{\theta}_k\rangle$ kennt – inklusive Parameter θ_k – dann kann man diese theoretischen Fehlerraten ausrechnen.

Ein denkbar einfacher Fall ist das Modell $\underline{X}_i \sim \mathcal{N}_m\langle\underline{\mu}_{K_i}, \boldsymbol{I}\rangle$ mit $\underline{\mu}_1 = [0,0]^T$, $\underline{\mu}_2 = [\Delta,0]^T$, vergleiche Abbildung 5.2.b (i). Die naheliegende Klassierungsregel sagt $\widehat{k}\langle\underline{x}\rangle = 1$, falls $x^{(1)} < 0$ ist, und sonst $\widehat{k}\langle\underline{x}\rangle = 2$. Die Fehlerraten für diese Regel lassen sich für das Modell einfach ausrechnen: Sie sind beide gleich $Q_{12} = Q_{21} = Q = \Phi\langle-\Delta/2\rangle$. Das Resultat gilt auch für allgemeine $\underline{\mu}_k$ und $\boldsymbol{\Sigma}$, wenn man $\Delta^2 = (\underline{\mu}_2 - \underline{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\underline{\mu}_2 - \underline{\mu}_1)$ setzt.

h Wenn man die Parameter θ_k nicht kennt, so kann man Schätzungen aus den Trainingsdaten einsetzen und erhält man eine „**parametrisch geschätzte Fehlerrate**“ $\widehat{Q} = \Phi\langle-\widehat{\Delta}/2\rangle$.

i Es ist plausibel, dass diese Fehlerrate ein **zu optimistisches Resultat** gibt, da die Regel ja an die Trainingsdaten optimal angepasst wurde. (Eine analoge Feststellung macht man in der multiplen Regression, wo man feststellt, dass die Residuen eine etwas kleinere Streuung haben als die Zufallsabweichungen.) Das gleiche gilt für die empirische Fehlerrate.

Diese Schwierigkeit kann man mit zwei Ideen umgehen, die auch in anderen Situation anwendbar und deshalb von grundlegender Bedeutung sind.

j **Testdaten.** Die erste ist grundsätzlich einfach: Man bestimme die Fehlerrate mit Hilfe von „neuen“ Daten, die für die Schätzung der Klassierungsregel nicht gebraucht wurden. Man nennt solche Beobachtungen zur Unterscheidung von den Trainingsdaten „Testdaten“. Man kann sich am Anfang einer Analyse dafür entscheiden, die vorhandenen Daten in zufälliger Weise in Trainingdaten und Testdaten aufzuteilen, um die Nützlichkeit der Ergebnisse am Schluss realistisch einschätzen zu können. Das ist sicher eine gute Strategie, wenn man eine Datenquelle hat, die grosse Anzahlen von Beobachtungen liefert, wie dies typischerweise im data mining der Fall ist.

k **Kreuz-Validierung.** In vielen Anwendungen ist der Datensatz begrenzt, und es wäre unvernünftig, für die Schätzung der Regel nur einen Teil zu benützen – nur, um ihre Genauigkeit nachher richtig einschätzen zu können. (* Richtig heisst hier „ohne systematischen Fehler“. Wenn der Testdatensatz klein ist, überwiegen aber die Zufallsfehler in der Schätzung der Genauigkeit!) Da hilft eine raffiniertere Version der vorhergehenden Idee weiter:

Wenn wir eine Beobachtung X_i nicht für die Schätzung der Entscheidungsregel benützen, bleibt die Wahrscheinlichkeit einer Fehlklassifikation intakt. Wir lassen also die $i$te Beobachtung weg und leiten die Regel mit den verbleibenden $n-1$ Trainingdaten her. Nun stellen wir fest, ob die Regel die $i$te Beobachtung richtig einteilt. Wenn wir den Aufwand nicht scheuen, können wir das für jede Beobachtung tun. Wenn wir jetzt die Anzahl Fehlklassifikationen zählen, gibt es keinen systematischen Optimismus mehr.

Die Klassierungsregel, die ohne die $i$te Beobachtung bestimmt wird, sei mit $\widehat{k}_{[-i]}$ bezeichnet. Dann wird die geschätzte Fehlerrate

$$Q_{cv1} = (\#\{i \mid \widehat{k}_{[-i]}\langle\underline{X}_{1i}\rangle = 2\} + \#\{i \mid \widehat{k}_{[-i]}\langle\underline{X}_{2i}\rangle = 1\})/n \ .$$

Die Methode heisst englisch *cross validation.*

* Für die Berechnung der Regeln ohne die $i$te Beobachtung muss man meistens nicht den ganzen Rechenaufwand wiederholen. Es gibt so genannte „update"-Formeln.

Die Idee lässt sich auch so ändern, dass jeweils nicht nur eine, sondern mehrere Beobachtungen aufs Mal weggelassen werden und dann die Fehlklassifikationen für alle Weggelassenen festgestellt werden.

L **Literatur:** Rencher (1998), Sec. 6.4

5.4 * Entscheidungstheorie

a In einigen Situationen, in denen Beobachtungen klassiert werden sollen, weiss man etwas über die Wahrscheinlichkeit der einzelnen Klassen. Im Beispiel der Iris-Blüten kann es sein, dass man die Häufigkeiten des Vorkommens der drei Arten in der untersuchten Gegend kennt. Man wird dann im Zweifelsfall eine Pflanze eher der häufigeren von zwei in Frage kommenden Arten zuordnen.

Diese Situation wird durch ein Modell beschrieben, in dem die **Klassenzugehörigkeit eine Zufallsvariable** K ist (5.1.d). Die Verteilungen $\mathcal{F}_k$ für die Klassen k müssen wir nun als **bedingte Verteilung** der Beobachtung $\underline{X}_i$, gegeben, dass sie zur Klasse k gehört, bezeichnen. Für ein vollständiges Modell brauchen wir noch die Wahrscheinlichkeiten $P\langle K_i = k\rangle$ der Klassen. In Formeln zusammengefasst wird das zu

$$(\underline{X}_i \mid K_i = k) \sim \mathcal{F}_k \;, \qquad P\langle K_i = k\rangle = \pi_k \;.$$

Die Verteilung von $\underline{X}_i$ bei unbekannter Klasse ist dann eine so genannt **Misch-Verteilung**. Wenn die Verteilungen $\mathcal{F}_k$ der Klassen Dichten f_k haben, dann ist die Dichte der Misch-Verteilung gleich

$$f\langle \underline{x}\rangle = \sum\nolimits_k \pi_k f_k\langle \underline{x}\rangle \;.$$

b **Bayes'scher Ansatz.** Welcher Klasse soll nun eine Beobachtung mit Merkmalswerten $\underline{x}_0$ zugeordnet werden? In der Sprache des Modells fragen wir nach der Zufallsvariablen K_0, wenn $\underline{X}_0$ gegeben ist. Die bedingten Wahrscheinlichkeiten von $K_0 = k$, gegeben die Merkmale $\underline{X}_0 = \underline{x}_0$, lassen sich berechnen,

$$P\langle K_0 = k \mid \underline{X}_0 = \underline{x}_0\rangle = \frac{dP\langle K_0 = k \cap \underline{X}_0 = \underline{x}_0\rangle}{dP\langle \underline{X}_0 = \underline{x}_0\rangle} = \frac{\pi_k f_k\langle \underline{x}_0\rangle}{\sum_\ell \pi_\ell f_\ell\langle \underline{x}_0\rangle} \;.$$

(Die Bezeichnung dP bedeutet, dass es sich da nicht um Wahrscheinlichkeiten, sondern eigentlich um Wahrscheinlichkeits-Dichten handelt.) Eine solche Formel haben wir im Einführungsteil schon angetroffen. Aus bedingten Wahrscheinlichkeiten eines Ereignisses B, gegeben eines der Ereignisse A_k, wurden die Wahrscheinlichkeiten der A_k, gegeben B berechnet. Die Formel entsprach genau dem hier angeführten Resultat und hiess **Satz von Bayes**. Die Wahrscheinlichkeiten π_k der Klassen werden in diesem Zusammenhang als **apriori-Wahrscheinlichkeiten** und die bedingten Wahrscheinlichkeiten $P\langle K_0 = k \mid \underline{X}_0 = \underline{x}_0\rangle$ als **aposteriori-Wahrscheinlichkeiten** bezeichnet – die ersteren gelten, *bevor* wir die Merkmalswerte $\underline{x}_0$ kennen, die letzteren *nachher.*

c Damit haben wir das **Grundschema der Bayes'schen Statistik** skizziert: Man geht davon aus, dass man über die unbekannte Grösse K ein Vorwissen hat, das sich in einer Wahrscheinlichkeitsverteilung, der apriori-Verteilung, ausdrücken lässt. Dieses Vorwissen kann aus früheren Studien stammen oder aus einer subjektiven Einschätzung bestehen. Es wiederspiegelt den Stand des Wissens, bevor die Beobachtung von $\underline{X}$ bekannt ist. Durch die Beobachtung wird das Wissen vemehrt, und man gelangt auf Grund des Satzes von Bayes zu einer neuen Stufe des Wissens, der aposteriori-Verteilung von K.

Dieses Denkschema kann auch auf die Parameter eines beliebigen parametrischen Modells angewandt werden. Der Parameter wird nicht mehr als feste, unbekannte Grösse angesehen, sondern als Zufallsvariable modelliert, für die eine apriori-Verteilung postuliert wird. Man erhält mit einer Beobachtung oder einer ganzen Stichprobe auf Grund des Satzes von Bayes eine aposteriori-Verteilung für den Parameter.

Die Bayes'sche Statistik hat in einigen Anwendungsgebieten eine starke Verbreitung erlangt.

d **Bayes'sche Zuordnungsregel.** Wahrscheinlichkeiten für die einzelnen Klassen ergeben ein genaues Bild des Wissens über die Klassenzugehörigkeit. Eine natürliche Zuordnungsregel oder Entscheidungsregel besteht darin, die Beobachtung $\underline{x}_0$ der Klasse mit der höchsten aposteriori-Wahrscheinlichkeit zuzuordnen, also

$$\widehat{k}_0 = \arg\max_k \langle P\langle K_0 = k \mid \underline{X}_0 = \underline{x}_0\rangle\rangle = \arg\max_k \left\langle \frac{\pi_k f_k\langle \underline{x}_0\rangle}{\sum_\ell \pi_\ell f_\ell\langle \underline{x}_0\rangle} \right\rangle \ .$$

Da der Nenner für alle Klassen gleich ist, wird die Regel einfach zu

$$\widehat{k}_0 = \arg\max_k \langle a_k\langle \underline{x}_0\rangle\rangle \quad \text{mit} \quad a_k\langle \underline{x}\rangle = \pi_k f_k\langle \underline{x}\rangle \ .$$

Für den Fall gleicher apriori-Wahrscheinlichkeiten π_k ist das die Regel, die bei der „nicht-Bayes'schen" Betrachtungsweise als naheliegend erschien. Im allgemeinen Fall zeigt sie die einfachste Art, das oben erwähnte Vorwissen über die Wahrscheinlichkeiten der Klassen in die Entscheidung einzubeziehen.

e* **Erwartete Fehlerrate.** Bisher haben wir die Regeln als „naheliegend" oder „natürlich" bezeichnet. Mathematiker und Mathematikerinnen mögen's präziser – und mit ihnen viele andere. Allgemein ist für eine Klassierung sicher anzustreben, dass man möglichst wenige Fehler macht. Diese Idee lässt sich formalisieren als Wahrscheinlichkeit einer falschen Klassierung oder erwartete Fehlerrate

$$P\langle \widehat{k}\langle \underline{X}_0\rangle \neq K_0\rangle = 1 - P\langle \widehat{k}\langle \underline{X}_0\rangle = K_0\rangle = 1 - \sum\nolimits_k \pi_k P\langle \widehat{k}\langle \underline{X}\rangle = k \mid K = k\rangle \ .$$

Man kann beweisen, dass die Bayes'sche Zuordnungsregel dieses Kriterium minimiert, dass sie also unter den gemachten Annahmen die optimale Regel ist.

f* **Kosten einer Fehlklassierung.** Der Fehler, einen Kranken als gesund zu erklären, ist oft viel gravierender als der umgekehrte Fall, dass ein Gesunder als krank erklärt wird – vor allem, wenn einige genauere Untersuchungen diesen Irrtum rasch klären können. Wenn gute und schlechte Schuldner einer Bank eruiert werden sollen, ist es weniger gravierend, einen guten Schuldner als schlecht einzustufen und dementsprechend sein Verhalten genauer zu verfolgen, als bei einem Schuldner, der schliesslich zahlungsunfähig wird, die notwendigen Massnahmen zur rechtzeitigen Verminderung des Risikos zu verpassen. Allgemein ist ein verpasster Alarm meistens teurer als ein Fehlalarm.

In solchen Fällen kann man versuchen, für die verschiedenen möglichen Fehler Kosten anzugeben. Es seien $c_{k\ell}$ die Kosten, falls die geschätzte Klasse $\widehat{k} = k$ und die wahre Klasse $K = \ell$ ist. Es können auch verschiedene „Gewinne“ c_{kk} für die richtigen Entscheidungen für die verschiedenen Klassen angegeben werden.

Nun liegt es nahe, als optimale Entscheidungsregel diejenige zu bezeichnen, für die die erwarteten Kosten minimiert werden. Als optimale Regel erhält man wieder $\widehat{k}_0 = \arg\max_k \langle a_k\langle \underline{x}_0\rangle\rangle$, diesmal mit $a_k\langle \underline{x}\rangle = -\sum_\ell c_{k\ell}\pi_\ell f_\ell\langle \underline{x}\rangle$. Falls $c_{k\ell} = 0$ für $k = \ell$ und $= 1$ sonst gilt, erhält man den vorhergehenden Fall.

Diese Überlegung bildet den Ausgangspunkt der **Entscheidungstheorie**, die von Wald ... begründet wurde.

g **Grundschema.** Die besprochenen Regeln beruhen alle darauf, aus $\underline{x}$ „Affinitäten“ $a_k\langle \underline{x}\rangle$ zu bestimmen und $\underline{x}$ der Gruppe mit der höchsten Affinität zuzuordnen.

Für die Praxis ist es häufig nützlich, nicht nur die Entscheidung $\widehat{k}\langle \underline{x}\rangle = \arg\max_k \langle a_k\langle \underline{x}\rangle\rangle$ anzugeben, sondern auch die $a_k\langle \underline{x}\rangle$ selbst. Man erhält damit ein differenzierteres Bild:

- Wenn es ähnlich grosse Affinitäten gibt, $a_\ell\langle \underline{x}_i\rangle \approx a_{\hat{k}}\langle \underline{x}_i\rangle$, dann ist die Entscheidung unsicher,
- Wenn auch die maximale Affinität klein ist, passt die Beobachtung zu keiner Klasse.

h* Die $a_k\langle \underline{x}\rangle$ müssen nicht als Wahrscheinlichkeiten interpretierbar sein – sie sind es in 5.4.f nicht. Wenn man das unbedingt – auch für a_k, die negativ werden können – formal ermöglichen will, kann man $p_k\langle \underline{x}\rangle = c^{a_k\langle \underline{x}\rangle} / \sum_{k'} c^{a_{k'}\langle \underline{x}\rangle}$ bilden. Für $c = e$ wird diese Regel „**softmax**“ genannt. Achtung: Wenn die Regel auf Wahrscheinlichkeiten angewendet wird, verändert sie diese. Die Basis c spielt eine Rolle.

5.5 * Weitere Methoden der Diskriminanz-Analyse

a Das bisher Besprochene hängt an der Annahme der multivariaten Normalverteilung. Für grosse Datensätze sollte man Verfahren finden können, die besser auf die Feinheiten der Verteilungen reagieren. Hier sollen einige solche Verfahren kurz erwähnt werden.

b **Nächste Nachbarn.** Ein Verfahren, das prinzipiell beliebig unförmige „Hohheitsgebiete“ der einzelnen Klassen zulässt, beruht darauf, für eine neu zu klassierende Beobachtung die $\ell \geq 1$ nächsten Nachbarn aus den Trainingsdaten zu suchen und ihre Klassenzugehörigkeit festzustellen. Die Zugehörigkeit der neuen Beobachtung wird dann durch „Mehrheits-Abstimmung“ dieser Nachbarn bestimmt. Als Verfeinerung könnte man ihre „Stimmen“ noch entsprechend ihrer Nähe zur neuen Beobachtung gewichten.

```
R> library(class) ; knn(...) ; knn1(...)
```

c **Neur(on)ale Netzwerke** eignen sich, um ein allgemeines Regressionsproblem mit Ausgangsgrössen oder „Input"-Variablen $X^{(j)}$ und einer oder mehrerer Zielgrössen oder „Output"-Variablen $\underline{Y}^{(k)}$ zu modellieren. Sie tun dies, indem sie eine Art von Schaltung abbilden, in der noch weitere „Schaltknoten" – oder, statistisch gesprochen, latente Variable – eingeführt werden.

Die üblichste Variante ist das „One hidden layer feed-forward neural network", das in Abbildung 5.5.c schematisch dargestellt ist. Das Modell hat die Form

$$Y^{(k)} = g_k \left\langle \alpha_k + \textstyle\sum_\ell w_{\ell k}\, \widetilde{g}_\ell \left\langle \widetilde{\alpha}_\ell + \sum_j \widetilde{w}_{j\ell} X^{(j)} \right\rangle \right\rangle \ .$$

Für g und $\widetilde{g}$ wird üblicherweise die logistische Funktion verwendet.

Für die Diskriminanz-Analyse muss der Output $\underline{Y}$ noch in eine Klassierung verwandelt werden. Bei 2 Klassen kann ein einzelnes Y mit einer Grenze verglichen werden wie bei der linearen Diskriminanz-Analyse; bei mehreren Klassen können ebenso viele $Y^{(k)}$ eingeführt und $K = \arg\max_k \langle Y^{(k)} \rangle$ gesetzt werden.

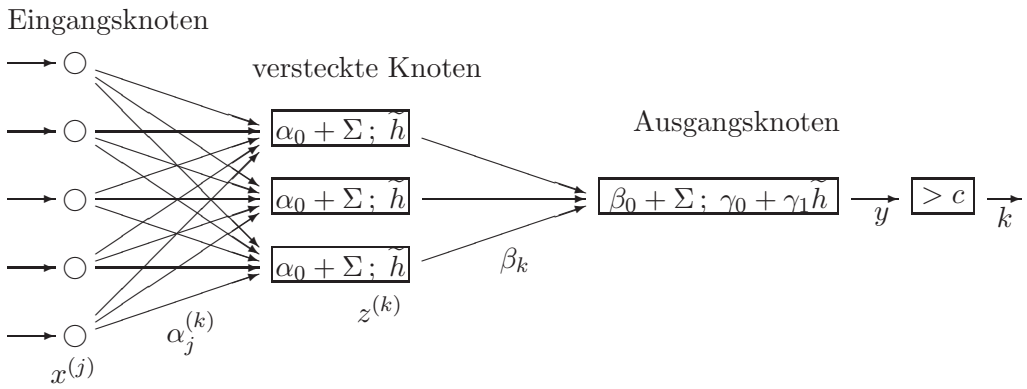


Abbildung 5.5.c: Schema eines Neuronalen Netzes mit einer „versteckten Schicht" von Knoten

Neuronale Netzwerke gelten bei Ingenieuren als Universalwerkzeug zur flexiblen Modellierung irgendwelcher Input-Output-Beziehungen. Statistiker geben zwei Punkte zu bedenken:

- Die Gefahr der Überanpassung an die Daten ist gross. Man muss darauf achten, dass die Anzahl der geschätzten Parameter klein bleibt gegenüber der Anzahl Beobachtungen, die zu ihrer Schätzung benützt werden.
- Das Modell liefert keine direkte anschauliche Darstellung und Interpretation, es bleibt eine „black box".
- Wenn neue Input-Daten etwas ausserhalb des Bereiches der Input-Daten des Trainings-Datensatzes liegen, wird die Klassierung völlig unzuverlässig sein.

```
R> library(nnet) ; nnet(...)
```

d **Classification and Regression Trees (CART).** Die Idee eines Klassierungs-Baumes ist einfach und entspricht den klassischen Bestimmungs-Schlüsseln für Pflanzenarten: Man teilt die Beobachtungen auf Grund einer geeigneten Variablen in 2 Gruppen ein, möglichst so, dass keine der Klassen in beiden Gruppen vertreten ist. Dann spaltet man jede Gruppe weiter in zwei auf mit Hilfe einer geeigneten Variablen und fährt so fort, bis in jeder Gruppe, soweit möglich, nur noch eine Klasse erscheint. So entsteht ein „Entscheidungsbaum“ *(decision tree)*.

`R> library(tree) ; tree(...)` oder `R> library(rpart) ; rpart(...)`

e **Boosting** heisst ein Rezept, das aus einer (zu) einfachen Klassierungsmethode durch „recycling“ ein besseres Verfahren macht:

1. Schätze die Regel mit der einfachen Methode. Das ergibt die Zuordnung $K^{(0)}\langle \underline{x}_i \rangle$
2. Bestimme die falsch klassierten Beobachtungen. Schätze die Regel nochmals, mit grösseren Gewichten für diese Beobachtungen. Bestimme so die Zuordnung $K^{(1)}\langle \underline{x}_i \rangle$

Wiederhole diesen Schritt einige Male, bis sich die Fehlklassierungen nicht mehr ändern. Bilde dann als neue Regel die Zuordnung, die sich aus einer „gewichteten Mehrheitsabstimmung“ ergibt.

Ein Nachteil des Verfahrens besteht darin, dass die erreichte Klassierungs-Regel nicht direkt interpretierbar ist, da nicht mehr eindeutig ist, welche Variablen zur Klassierung wirksam sind. Das gilt auch für das folgende Verfahren, das auf einer ähnlichen Idee beruht.

Literatur: Friedman, Hastie and Tibshirani (2000)

f **Bagging** ist eine zweite Idee, die eine einfache Klassierungsmethode verbessert. Wie der längere Name „Bootstrap Aggregating“ ausdrückt, wird die einfache Regel mittels boostrap vielfach bestimmt. Das verbesserte Verfahren wird wieder durch Mehrheits-Abstimmung festgelegt.

L **Literatur:** Ripley (1996) Behandelt alle ausser den letzten 2 Methoden recht ausführlich und angewandt. Manchmal nicht präzis.

5.S S-Funktionen

a **Lineare Diskriminanzanalyse.** Eine lineare Diskriminanzanalyse wird mit der Funktion `lda` aus dem package `MASS` durchgeführt,

```
> library(MASS)
> t.r  <- lda(Species~.,data=iris)    oder
> t.r  <- lda(x=iris[,1:4], grouping=iris[,5])
```

Im ersten Fall gibt man eine Formel `groups ~ x1 + x2 + ...` an, wobei `groups` der Gruppierungsfaktor ist und x_i die kontinuierlichen X-Variablen. (Die verwendete kurze Formel `Species~.` ist eine abgekürzte Schreibweise für den Fall, dass man alle Variablen ausser der links von $\sim$ stehenden als X-Variable verwenden will.) Das Argument `data` gibt wie üblich an, in welchem Data Frame diese Variablen zu finden sind. In der zweiten Variante ist `x` ein Data Frame oder eine Matrix und enthält die X-Variablen, und `grouping` ist die Gruppierungsvariable.

Mit weiteren Argumenten kann man Varianten wählen:

`prior`	a-priori Wahrscheinlichkeiten π_i
`CV = T`	Schätzung von Fehlerraten durch cross validation
`method`	robuste Schätzmethoden

Das Resultat, ein `lda`-Objekt, enthält die Komponenten

`$counts`	Anzahl Beobachtungen in den Gruppen
`$means`	Mittelwerte
`$scaling`	β-Koeffizienten der Diskriminanzfunktion(en); α erhält man leider nicht.
Falls `CV = TRUE`:	
`$class`	Klassen-Zugehörigkeiten gemäss Kreuzvalidierung
`$posterior`	a-posteriori Wahrscheinlichkeiten.

b **Grafische Ausgabe.** `plot(t.r)` stellt die Werte der Diskriminanzfunktion(en) der Beobachtungen dar. Für zwei Gruppen wird ein Histogramm gezeigt, für drei ein Streudiagramm der beiden Diskriminanzfunktionen, für mehr Gruppen eine Streudiagramm-Matrix.

c Die **Identifikation**, also die Bestimmung der plausibelsten Klassen-Zugehörigkeit für beliebige Beobachtungen liefert

```
> predict(object=t.r, newdata)
```

Dabei ist `object` das Resultat von `lda` (ein `lda`-Objekt) und `newdata` sind die zu klassierenden Daten; beim Weglassen werden die Trainingsdaten verwendet.

d **Logistische Regression.** Die logistische Regression erhält man mittels der Funktion `glm` mit dem Argument `family=binomial`,
`> t.r <- glm(Species~., data=iris[51:150,], family=binomial) t.p <- predict(t.r, newdata, type="response")` liefert dann a-posteriori-Wahrscheinlichkeiten, und die Zuordnung erhält man, indem man die Beobachtungen mit `t.p<0.5` der ersten und die anderen der zweiten Gruppe zuordnet.

e* **Weitere Methoden.** Mit `qda` („q“ für quadratic) lässt sich eine quadratische Diskriminanzanalyse durchführen. `predict` wählt dann die entsprechende Methode für die Identifikation.
Es gibt im package `mda` auch `mda` (mixed) und `fda` (flexible discriminant analysis).

6 Multivariate Regression

6.1 Das Modell

a In der multiplen linearen Regression wird der Zusammenhang von mehreren Eingangsvariablen oder Regressoren mit einer kontinuierlichen Zielgrösse untersucht. Hier sollen **mehrere Zielgrössen** gleichzeitig betrachtet werden.

▷ **Beispiel Fossilien.** Aus Fossilien, die man in verschiedenen Schichten von Meeres-Ablagerungen findet, will man auf Umweltbedingungen (Temperatur, Salzgehalt) der entsprechenden Zeitperioden schliessen.

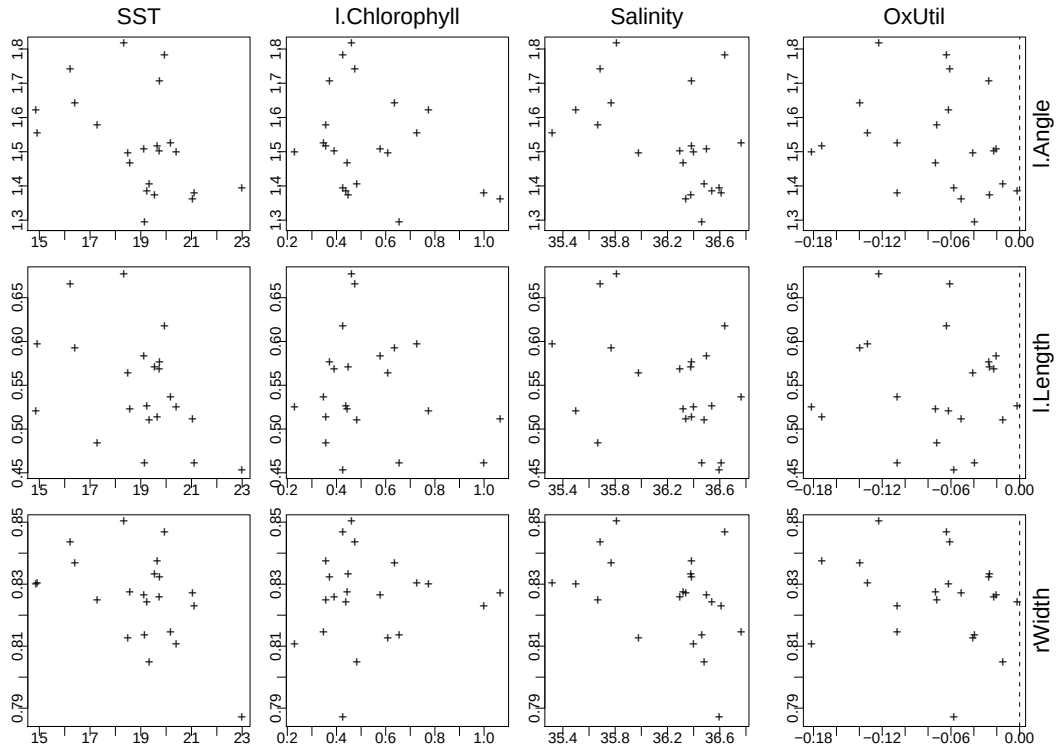


Abbildung 6.1.a: Umweltvariable und Form-Merkmale im Beispiel der Fossilien

Dazu werden an verschiedenen Stellen des Atlantischen Ozeans Messungen an „coccoliths“ der Art *Gephyrocapsa* der obersten Ablagerungen vorgenommen und mit den heutigen Umweltbedingungen in Beziehung gesetzt. In Abbildung 6.1.a sind die Beziehungen zwischen den einzelnen Umweltvariablen und den Form-Merkmalen der Fossilien dargestellt.

Das entsprechende Modell soll nachher dazu benützt werden, anhand von coccoliths aus tieferen Schichten auf die Umweltbedingungen in den entsprechenden Zeitperioden zurückzuschliessen. Für diesen Schluss muss man von der Annahme ausgehen, dass sich diese Beziehungen seither nicht geändert haben. Genaueres steht in Bollmann, Henderiks and Brabec (2002). ◁

b **Modell.** Das Modell der multiplen linearen Regression mit einer einzigen Zielgrösse war $Y_i = \beta_0 + \sum_k \beta_k x_i^{(k)} + E_i$. Wenn nun der Zusammenhang mehrerer Zielgrössen $Y^{(j)}$, $j = 1, ..., m$, von den Eingangsgrössen (oder erklärenden Variablen) $X^{(k)}$ untersucht werden soll, dann können wir zunächst für jede ein solches Modell aufstellen, also

$$Y_i^{(j)} = \beta_0^{(j)} + \sum_k \beta_k^{(j)} x_i^{(k)} + E_i^{(j)}$$

Das soll wieder mit Matrizen zusammengefasst werden,

$$\boldsymbol{Y} = \boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{E}\ .$$

Die einzelnen Modelle, in Matrix-Schreibweise, erhalten wir, wenn wir jeweils die j-te Spalte von $\boldsymbol{Y}$, $\boldsymbol{\beta}$ und $\boldsymbol{E}$ auswählen: $\underline{Y}^{(j)} = \boldsymbol{X}\,\underline{\beta}^{(j)} + \underline{E}^{(j)}$. Wie in der Matrixschreibweise der **univariaten Regression**, also der Regression mit einer einzigen Zielgrösse, erscheint der **Achsenabschnitt** β_0 in der Matrixform nicht mehr; er wird dadurch berücksichtigt, dass eine Spalte mit lauter Einsen in die Design-Matrix $\boldsymbol{X}$ eingeschlossen wird.

c **Eingangsgrössen, Regressoren und Terme.** Mit Regressionsmodellen wird allgemein ein Zusammenhang zwischen Zielgrössen und Eingangsgrössen untersucht. Die Eingangsgrössen werden oft als erklärende Variable bezeichnet, was sicher gerechtfertigt ist, wenn ein Ursache-Wirkungs-Zusammenhang besteht. Da Regression auch sinnvoll ist, wenn das nicht postuliert werden kann, soll der neutrale Ausdruck **Eingangsgrösse statt erklärende Variable** benützt werden. Die ebenfalls übliche Bezeichnung „unabhängige Variable“ wird vermieden, da das Adjektiv „unabhängig“ nur verwirrt: Die X-Variablen müssen in keiner Weise unabhängig voneinander sein.

Eingangsgrössen gehen oft nicht in der ursprünglichen Form ins Regressionsmodell ein, sondern werden zunächst transformiert – einzeln, beispielsweise mit einer Logarithmus-Transformation, oder gemeinsam, indem beispielsweise die eine als Prozentzahl einer anderen ausgedrückt wird. Diese transformierten Grössen, die als X-Variable ins Regressionsmodell eingehen, nennen wir **Regressoren**. Analoges kann mit den Zielgrössen geschehen. Man könnte dann die transformierten Zielgrössen als „Regressanden“ bezeichnen. Die Unterscheidung zu den Zielgrössen ist weniger wichtig: in der Residuenanalyse spielen die untransformierten Eingangsgrössen eine Rolle, untransformierte Zielgrössen dagegen nicht. Da zudem „Regressand“ zu ähnlich tönt wie „Regressor“, bleiben wir beim Ausdruck „Zielgrösse“, der auch für transformierte Zielgrössen gelten soll.

Bei der Festlegung des Modells kommt zusätzlich der Begriff „**Term**“ ins Spiel. Die dummy-Variablen, die zu einem Faktor (s. unten, 6.1.f) oder einer Interaktion zwischen Variablen gehören, bilden jeweils einen Term. Bei der Modellwahl wird jeder Term ins Modell gesamthaft einbezogen oder weggelassen.

d **Zufallsabweichungen.** Die Annahmen über die Verteilung der Zufallsabweichungen $E_i^{(j)}$ bilden die naheliegende Verallgemeinerung der Annahmen im Fall einer einzigen Zielgrösse. Es sei $\underline{E}_i$ die $i$te Zeile von $\boldsymbol{E}$, also der Vektor der Zufallsabweichungen aller Zielgrössen für die Beobachtung i. Die Annahmen sind:

- Erwartungswert $\mathcal{E}\langle\underline{E}_i\rangle = \underline{0}$. Diese Festlegung ist für die Identifizierbarkeit von $\underline{\beta}$ nötig und sagt auch, dass die (lineare) Regressionsfunktion richtig ist.
- Die Zufallsabweichungen $E_i^{(j)}$ haben Varianzen σ_j^2, die für sich für die Zielgrössen j unterscheiden. Ausserdem können die $E_i^{(j)}$ für verschiedene Zielgrössen zusammenhängen. Beides zusammen wird durch die Kovarianzmatrix $\mathrm{var}\langle\underline{E}_i\rangle = \mathbf{\Sigma}$ charakterisiert, von der wir annehmen, dass sie gleich ist für alle Beobachtungen i.
- Die Zufallsabweichungen der *verschiedenen Beobachtungen* sind unabhängig (oder wenigstens unkorreliert), $\mathcal{E}\langle\underline{E}_h\underline{E}_i^T\rangle = \mathbf{0}$, falls $h \neq i$.
- Die Zufallsabweichungen sind gemeinsam normalverteilt.

Man kann das alles zusammenfassen zu

$$\underline{E}_i \sim \mathcal{N}_m\langle\underline{0}, \mathbf{\Sigma}\rangle\,, \quad \text{unabhängig}\,.$$

Das Modell mit der gemeinsamen Verteilung der Fehlerterme ist das Modell der **multivariaten Regression**. Sie ist auch eine *multiple Regression*, soweit sie mehrere Regressoren $X^{(j)}$ umfasst.

e Die Modelle für die einzelnen Zielgrössen haben wir zunächst einfach formal in eine einzige Matrizen-Formel geschrieben. Durch die Annahme einer gemeinsamen Normalverteilung der Fehlerterme erhalten sie jetzt auch inhaltlich eine Verbindung.

Die Tatsache, dass die Design-Matrix $\boldsymbol{X}$ für alle Zielgrössen die gleiche ist, muss nicht zwingend einen inhaltlichen Zusammenhang angeben: Wenn die Koeffizienten-Matrix $\boldsymbol{\beta}$ in jeder Zeile nur ein einziges von Null verschiedenes Element enthält, dann reagieren die Zielgrössen eben auf verschiedene Regressoren, die man nur formell zu einer Matrix zusammengefasst hat.

f **Varianzanalyse.** Die Eingangsgrössen können, wie in der univariaten Regression, auch Faktoren (nominale oder kategorielle Variable) sein, die als „dummy variables“ in die Design-Matrix $\boldsymbol{X}$ eingehen.
Die multivariate Varianzanalyse mit festen Effekten (MANOVA) kann deshalb als Spezialfall der Regression behandelt werden. Wie bei einer einzigen Zielgrösse gibt es aber interessante zusätzliche methodische Aspekte.

6.2 Schätzungen und Tests

a **Schätzung der Koeffizienten.** Die Spalten von $\boldsymbol{\beta}$ können separat durch Kleinste Quadrate, also mit je einer (univariate) multiplen Regressionrechnung geschätzt werden. Das lässt sich aber auch zusammengefasst schreiben als

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Die angepassten Werte und die Residuen sind auch wie früher definiert und werden zusammengefasst zur Matrix $\widehat{\boldsymbol{Y}} = \boldsymbol{X}\,\widehat{\boldsymbol{\beta}}$ und zur Residuen-Matrix $\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$.

Die **Schätzung der Kovarianzmatrix** der Zufallsabweichungen erfolgt durch die empirische Kovarianzmatrix der Residuen unter Berücksichtigung der Anzahl $n-p$ der Freiheitsgrade,

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-p}\boldsymbol{R}^T\boldsymbol{R}$$

b **Verteilung der geschätzten Koeffizienten.** Wie zu erwarten ist, sind die Koeffizienten erwartungstreu und normalverteilt. Die Kovarianzmatrix der geschätzten Koeffizienten wird schon irgendwie zu berechnen sein; überlassen wir das getrost den Programmen!

* Wollen Sie es etwas genauer wissen? Da stossen wir auf eine Schwierigkeit in der Notation: Die geschätzten Koeffizienten bilden eine zufällige Matrix. Wer brauchen nicht nur die Verteilung jedes einzelnen Elementes dieser Matrix, sondern auch die gemeinsame Verteilung der Elemente. Insbesondere interessieren uns auch die Kovarianzen zwischen den Elementen. Sie werden $\operatorname{cov}\langle\widehat{\beta}_h^{(j)},\widehat{\beta}_k^{(\ell)}\rangle = ((\boldsymbol{X}^T\boldsymbol{X})^{-1})_{hk}\,\boldsymbol{\Sigma}_{j\ell}$. Das lässt sich nicht direkt als Matrix schreiben, denn es variieren vier Indices!

Im Übrigen ist die Herleitung nicht schwierig: Man setzt $\boldsymbol{C} = \boldsymbol{X}^T\boldsymbol{X}$ und rechnet

$$\begin{aligned}\operatorname{cov}\langle\underline{\widehat{\beta}}^{(j)},\underline{\widehat{\beta}}^{(k)}\rangle &= \operatorname{cov}\langle\boldsymbol{C}^{-1}\boldsymbol{X}^T\underline{Y}^{(j)},\boldsymbol{C}^{-1}\boldsymbol{X}^T\underline{Y}^{(k)}\rangle = \boldsymbol{C}^{-1}\boldsymbol{X}^T\operatorname{cov}\langle\underline{Y}^{(j)},\underline{Y}^{(k)}\rangle(\boldsymbol{C}^{-1}\boldsymbol{X}^T)^T\\ &= \boldsymbol{C}^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}_{j\ell}\boldsymbol{X}(\boldsymbol{C}^{-1})^T = \boldsymbol{\Sigma}_{j\ell}\boldsymbol{C}^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{C}^{-1} = \boldsymbol{\Sigma}_{j\ell}\boldsymbol{C}^{-1}\end{aligned}$$

c ▷ **Im Beispiel der Fossilien** sind die Ergebnisse für die einzelnen Zielgrössen in Tabelle 6.2.c zusammengestellt. Sie sind nicht ermutigend, liefert doch der Gesamttest für keine Zielgrösse ein signifikantes Resultat! ◁

	l.Angle		l.Length		rWidth	
	coef	p-value	coef	p-value	coef	p-value
(Intercept)	447.787	0.347	1.6590	0.471	0.59265	0.283
SST	-0.721	0.800	-0.0102	0.463	-0.00493	0.147
l.Chlorophyll	-19.202	0.155	-0.0765	0.238	0.00208	0.890
Salinity	-10.756	0.452	-0.0242	0.726	0.00888	0.590
OxUtil	-23.770	0.662	0.0433	0.869	-0.04368	0.489
R^2	0.285	0.260	0.2571	0.198	0.24889	0.255

Tabelle 6.2.c: Regressionskoeffizienten und Bestimmtheitsmasse mit p-Werten für die einzelnen Form-Variablen als Zielgrössen im Beispiel der Fossilien

d **Gemeinsame Tests.** Wir wissen, wie wir für jede einzelne Zielgrösse $Y^{(j)}$ testen, ob sie von den Eingangsgrössen abhängt. Aus der gemeinsamen Betrachtung ergibt sich auch die **gemeinsame Nullhypothese, dass keine der Zielgrössen von einem Regressor $X^{(k)}$ abhängt**, dass also $\beta_k^{(j)} = 0$ ist für alle j oder, noch umfassender, dass zwischen keiner Zielgrösse und keinem Regressor ein Zusammenhang besteht, dass also alle $\beta_k^{(j)} = 0$ sind. Dazwischen liegen, wie in der univariaten linearen Regression, die Vergleiche von hierarchisch geschachtelten Modellen.

Die naheliegendste Art, eine solche Hypothese zu testen, besteht in der Verwendung des entsprechenden Likelihood-Ratio-Tests. Diese Teststatistik wird als Wilks' Λ (grosses griechisches Lambda) bezeichnet.

In der univariaten Regression setzt die Teststatistik des F-Tests im Wesentlichen die „between group sum of squares" ins Verhältnis zur „within group sum of squares". Im multivariaten Fall werden beide Grössen zu „sum of squares and cross products" *Matrizen*, bezeichnet mit $\boldsymbol{B}$ und $\boldsymbol{W}$. Entscheidend ist wieder die Grösse des Quotienten. Teststatistiken sind deshalb Funktionen der Eigenwerte λ_k von $\boldsymbol{W}^{-1}\boldsymbol{B}$.

- Wilks: $\prod_\ell 1/(1+\lambda_\ell)$
- Pillai: $\sum_\ell \lambda_\ell/(1+\lambda_\ell)$
- Lawley-Hotelling: $\sum_\ell \lambda_\ell$
- Roy (union-intersection): λ_1 (resp. $\lambda_1/(1+\lambda_1)$)

Im multivariaten Fall gibt es also **mehrere gebräuchliche Tests,** die für den Fall einer einzigen Zielgrösse in den üblichen F-Test (oder t-Test) übergehen. Wenn der Einfluss einer einzigen kontinuierlichen Variablen getestet wird, liefern alle diese Tests das gleiche Ergebnis (* da $\boldsymbol{B}$ nur einen Freiheitsgrad hat und deshalb nur der erste Eigenwert λ_1 von 0 verschieden ist).

e ▷ **I**m **Beispiel der Fossilien** zeigt der globale Test, der die Nullhypothese prüft, dass kein Zusammenhang zwischen den Form- und den Umwelt-Variablen bestehe, keine Signifikanz! Die Sache scheint also hoffnungslos. – Genauere Analysen ergaben die Möglichkeit, aus der Verteilung des Winkels (Angle) und der Länge (l.Length) drei Gruppen zu identifizieren und die Anteile dieser Gruppen in den Stichproben als neue Zielvariable einzuführen. Tabelle 6.2.e gibt in der mit „.total." bezeichneten Zeile an, dass die Umweltvariablen gesamthaft auf diese Gruppen einen signifikanten Einfluss haben. Die anderen Teststatistiken führen zu P-Werten von 0.0388 (Pillai), 0.0163 (Hotelling-Lawley) und 0.00381 (Roy).

	Df	Wilks	approx F	num Df	den Df	p value
SST	1	0.564	5.405	2	14	0.0182
l.Chlorophyll	1	0.886	0.905	2	14	0.4271
Salinity	1	0.847	1.267	2	14	0.3122
OxUtil	1	0.890	0.863	2	14	0.4431
.total.	4	0.417	1.922	8	28	0.0961
Residuals	15					

Tabelle 6.2.e: Gesamttests für den Einfluss der einzelnen Regressoren auf die Gruppenanteile, sowie für alle Regressoren zusammen im Beispiel der Fossilien

Gemäss Tabelle haben die Variablen SST und OxUtil einen signifikanten Einfluss. Wie in der univariaten Regression ist es aber denkbar, dass die höheren P-Werte der anderen beiden Variablen von Kollinearität verursacht sind. Um dies zu überprüfen, rechnen wir wie früher das Modell ohne die am wenigsten signifikante Variable durch, also ohne `l.Chlorophyll`. Die Variable `Salinity` erhält dann einen P-Wert von 0.177, während die andern beiden mit 0.018 und 0.053 ähnliche P-Werte behalten. ◁

f **Bedeutung der multivariaten Regression.** Von der Interpretation her sind meistens die Koeffizienten $\boldsymbol{\beta}$ von Interesse.
Schätzung und Vertrauensintervall für ein $\beta_k^{(j)}$ aus der multivariaten Regression sind identisch mit denen aus der multiplen Regression von $Y^{(j)}$ auf die Regressoren – die anderen Zielgrössen haben keinen Einfluss.
Wenn man einen Lauf mit einem Programm für multivariate Regression macht, erhält man also als Hauptsache das, was auch m Läufe eines Programms für multiple Regression liefert. Zusätzlich erhält man:

- die Kovarianzmatrix der Zufallsfehler. Die Korrelation zwischen den Zufallsabweichungen $E^{(j)}$ und $E^{(\ell)}$ von den linearen Regressionen von $Y^{(j)}$ und $Y^{(\ell)}$ auf die Regressoren $X^{(k)}$, $k = 1, ..., p$ nennt man auch **partielle Korrelation** zwischen $Y^{(j)}$ und $Y^{(\ell)}$, gegeben die X-Variablen.
- gemeinsame Tests für die vorher genannten Fragen, ob die Zielgrössen alle von mit gewissen Regressoren lineare Zusammenhänge zeigen.

g **Residuen-Analyse.** Residuen-Analyse zur Prüfung der Modellannahmen ist, wie in allen Regressionsmodellen, ein unverzichtbarer Bestandteil einer seriösen Datenanalyse. Zuerst sollen die Regressionen für alle einzelnen Zielgrössen mit den bekannten Methoden überprüft werden.

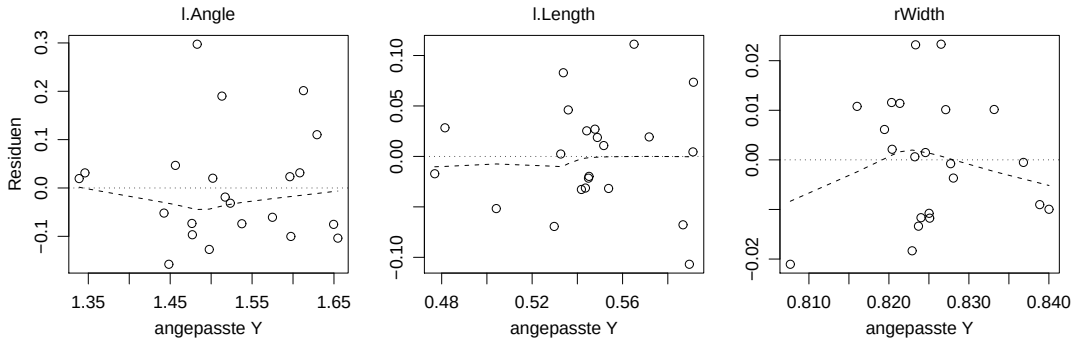


Abbildung 6.2.g (i): Tukey-Anscombe-Diagramme für das Beispiel der Fossilien

Abbildung 6.2.g (i) zeigt für das Beispiel die Zusammenstellung der Tukey-Anscombe-Diagramme, die zur Gesamtüberprüfung des Modells und insbesondere für Hinweise auf die Nützlichkeit einer Transformation der Zielgrössen dienen. Die Streudiagramme der Residuen gegen die Hebelarm-Werte *(leverages,* Abbildung 6.2.g (ii)), aus denen man einflussreiche Beobachtungen erkennt. Die Streudiagramme der Residuen gegen die Eingangsgrössen (Abbildung 6.2.g (iii)) sollen vor allem Hinweise auf Nichtlinearitäten in den Eingangsgrössen geben. – Ausser einer schiefen Fehler-Verteilung für `l.Angle` zeigt sich im Beispiel kaum etwas ernst zu Nehmendes.

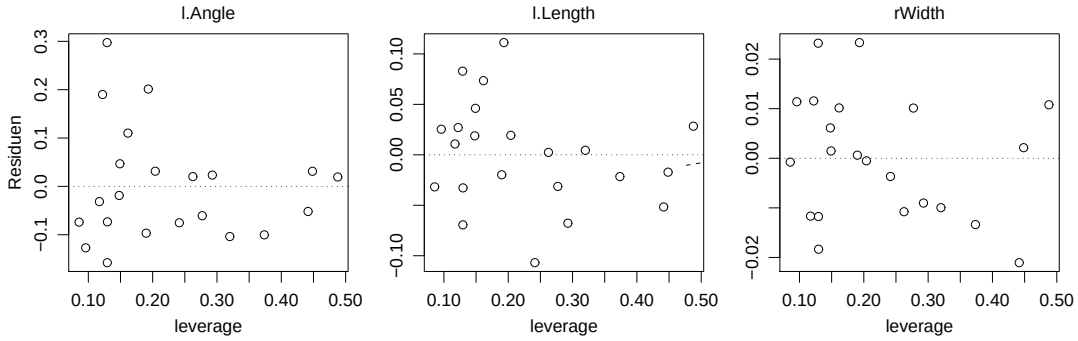


Abbildung 6.2.g (ii): Streudiagramm der Residuen gegen Hebelarmwerte für das Beispiel der Fossilien

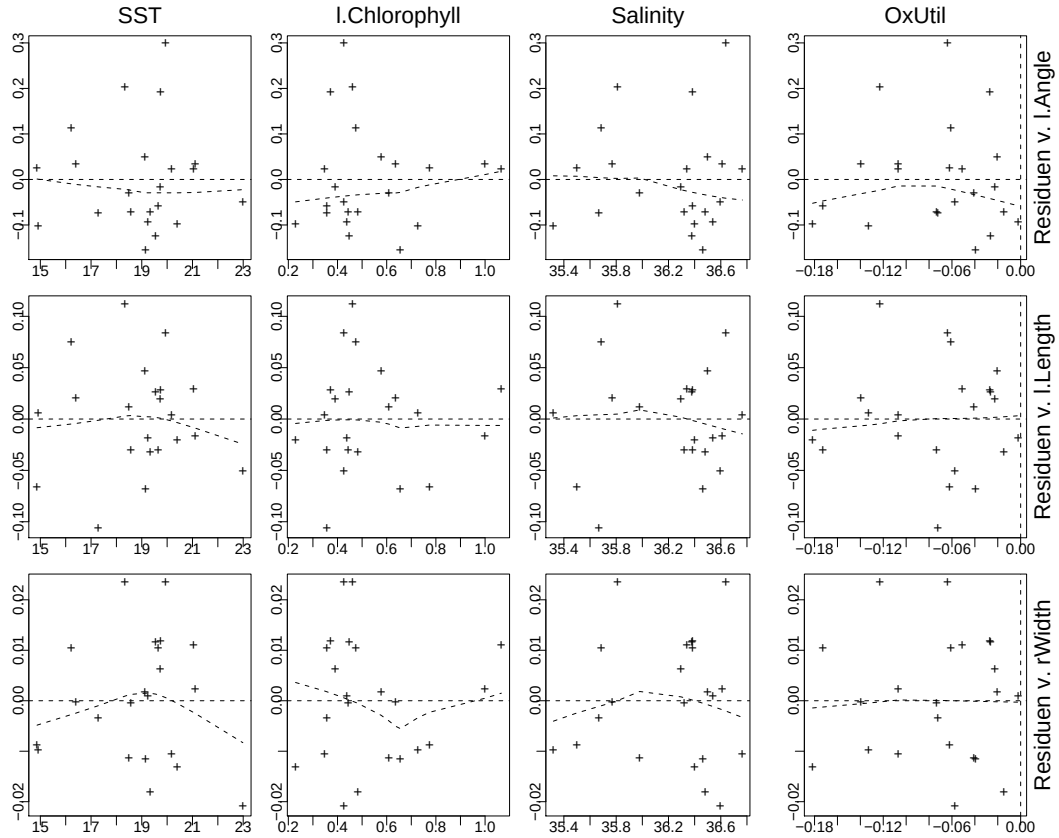


Abbildung 6.2.g (iii): Streudiagramme der Residuen gegen die Eingangsgrössen für das Beispiel der Fossilien

h In Ergänzung zu den Residuen-Analysen für jede Zielgrösse lohnt es sich, eine **Streudiagramm-Matrix der Residuen**-Matrix (Abbildung 6.2.h (i)) zu betrachten. Es fallen im Beispiel (mindestens) zwei extreme Punkte mit grossen Residuen für alle

Variablen auf. Wären mehr Beobachtungen vorhanden, so könnte man die Rechnungen ohne diese beiden Punkte wiederholen.

Aus den Residuen und ihrer geschätzten Kovarianzmatrix erhält man in der früher besprochenen Art (3.2.o) die **Mahalanobis-Beträge**, die man mit Hilfe eines Quantil-Quantil-Diagramms mit der entsprechenden Verteilung, der „Wurzel-Chiquadrat-Verteilung“, vergleichen kann (Abbildung 6.2.h (ii)). Damit überprüft man einen Aspekt der Annahme der multivariaten Normalverteilung der Fehler $\underline{E}_i$.

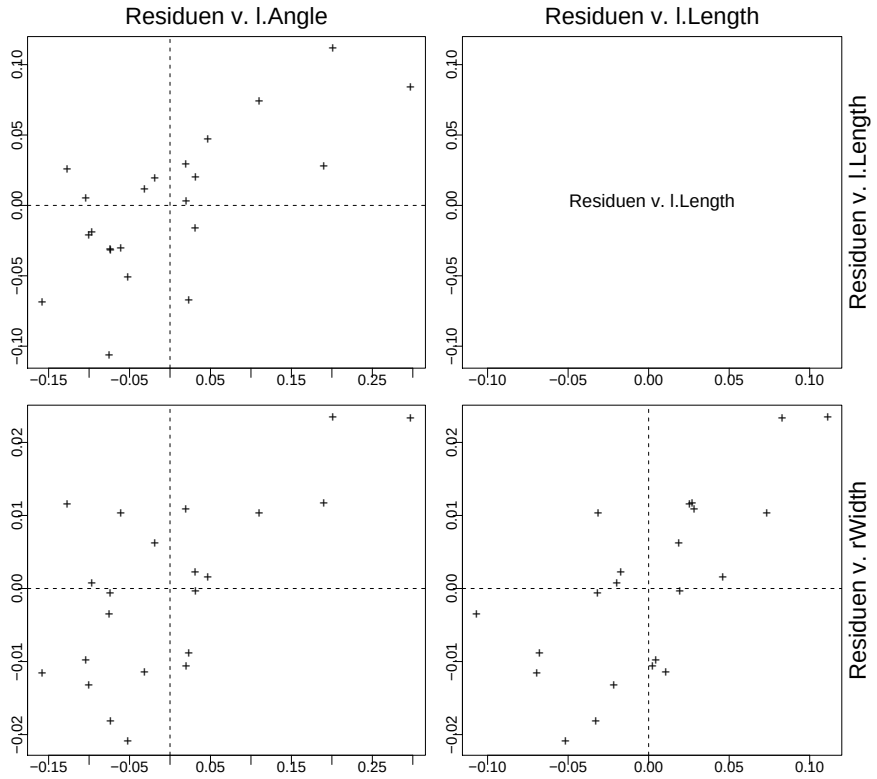


Abbildung 6.2.h (i): Streudiagramm-Matrix der Residuen für das Beispiel der Fossilien

i* **Vorhersage.** Es soll für einen gegebenen Satz $\underline{x}_0$ von Werten der Regressoren eine neue Beobachtung $\underline{Y}_0$ gemacht werden. Was können wir im Voraus über die Verteilung von $\underline{Y}_0$ sagen?

Bei bekannten Parametern ist das Problem trivial: Die gesamte Verteilung der neuen Beobachtung ist durch das Regressionsmodell 6.1.b gegeben. Die beste Vorhersage ist der Erwartungswert von $\underline{Y}_0$, $\mathcal{E}\langle \underline{Y}_0^T \rangle = \underline{x}_0^T \boldsymbol{\beta}$ (transponiert geschrieben).

In der Realität muss der Zusammenhang von $\underline{x}$ und $\underline{Y}$ aus „Trainingsdaten“ $\boldsymbol{X}$, $\boldsymbol{Y}$ geschätzt werden. Das führt zur Schätzung $\widehat{\boldsymbol{\beta}}$, die wir an Stelle von $\boldsymbol{\beta}$ einsetzen. Die beste Vorhersage wird also $\underline{\widehat{Y}}_0^T = \underline{x}_0^T \widehat{\boldsymbol{\beta}}$. Die Verteilung der Vorhersage lässt sich wie im eindimensionalen Fall aus der Verteilung der geschätzten Koeffizienten herleiten.

j* **Vorhersagebereich.** Der Vorhersagebereich soll die *Beobachtung* mit vorgegebener Wahrscheinlichkeit enthalten (während ein Vertrauensbereich einen *Parameter* mit einer solchen Wahrscheinlichkeit enthält). Analog zum Vorhersage-Intervall für eine einige Zielgrösse gibt die Summe der Kovarianzmatrizen für die geschätzte beste Vorhersage und für den Zufallsfehler der neuen Beobachtung, $\mathrm{var}\langle \underline{\widehat{Y}} \rangle + \widehat{\boldsymbol{\Sigma}}$, die Grösse und Form des gesuchten Bereiches an.

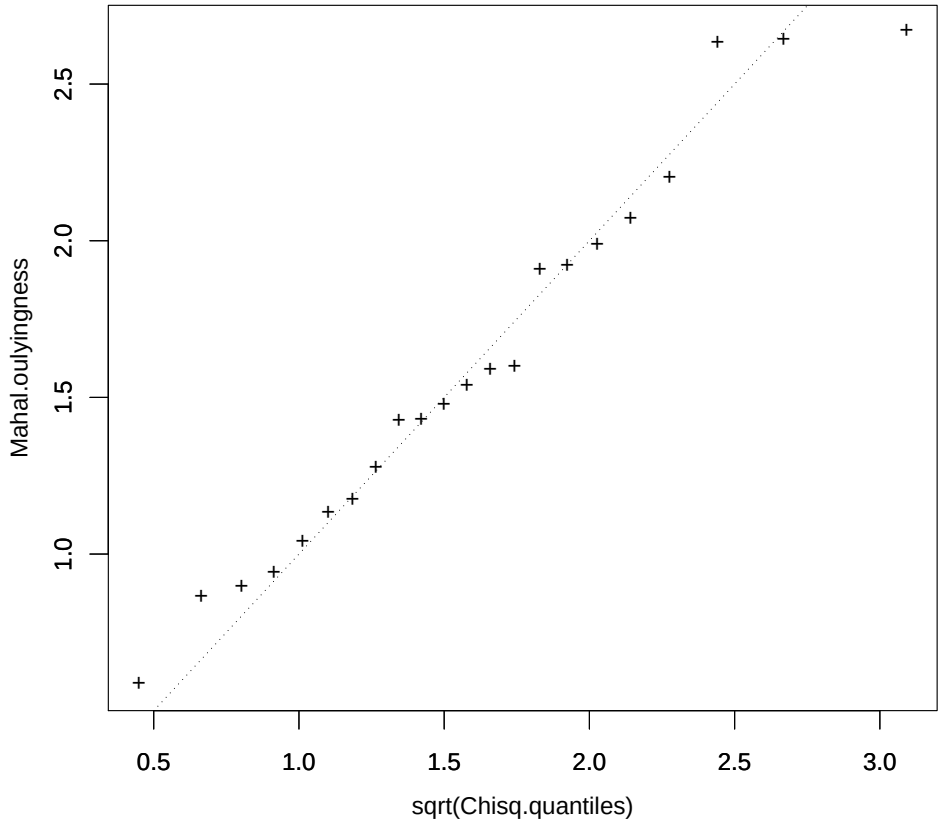


Abbildung 6.2.h (ii): Q-Q-Diagramm der Längen der multivariat standardisierten Residuen für das Beispiel der Fossilien

6.S S-Funktionen

a Zur Durchführung von multivariaten Varianzanalysen und Regressionen dienen die Funktionen `lm` und `manova`

```
> t.r  <- lm( cbind(Sepal.Length,Sepal.Width,Petal.Length,
Petal.Width)~Species, data=iris)
```

erzeugt, da `lm` hier mit mehreren Zielgrössen aufgerufen wird, ein Objekt der Klasse `mlm`, für die

```
> summary(t.r)
```

die Resultate für alle Zielgrössen nacheinander auflistet.

b Für multivariate Tests braucht ersetzt man `lm` durch `manova` im vorhergehenden Aufruf. Die Funktion `summary(t.r, test="Wilks")` führt dann den Test durch. Wenn das Modell mehrere Terme (Faktoren, Ausgangsgrössen) umfasst, werden entsprechend viele Tests durchgeführt – Vorsicht! Es sind „Type I“ Tests, die für ein schrittweise aufgebautes Modell jeweils prüfen, ob der nächste Term eine signifikante Verbesserung des Modells bringt.

c Da R zurzeit in dieser Beziehung lückenhaft ist, stellt der Autor einige Funktionen zur Verfügung. Man erhält sie über die Website `r-forge.r-project.org`.

d **Funktion `regr`.** Die Funktion `regr`, die viele Regressionmodelle anpassen kann, erlaubt auch multivariate Regression und liefert die Information, die hier empfohlen wird, wenn man das Ergebnis druckt oder `plot` übergibt.

e Die Funktion `plot.regr`, die mit `plot(t.r)` für ein `regr`-Ergebnis aktiviert wird, liefert eine umfassende Residuen-Analyse. Wenn einige Ziel- und Ausgangsgrössen im Modell sind, produziert das viele Seiten grafischen Output! In Kürze, was diese Funktion zeigt:

- Streudiagramme der Residuen gegen die angepassten Werte für alle Zielgrössen. Diese Streudiagramme dienen dazu, die generelle Form der Regressionsfunktionen zu prüfen und insbesondere Hinweise auf allfällige Transfomationen der Zielgrössen zu geben.
- Streudiagramme der Absolutwerte der Residuen gegen die angepassten Werte. Man kann gegebenenfalls Abweichungen von der Voraussetzung der gleichen Varianzen für alle Beobachtungen entdecken.
- Normalverteilungs-Diagramme.
- Streudiagramme der Residuen gegen die „Hebelarm"-Werte (leverages). Sie zeigen einflussreiche Beobachtungen an.
- Streudiagramm-Matrix der Residuen für die verschiedenen Zielgrössen.
- Streudiagramm-Matrix der Residuen gegen die Ausgangsgrössen im Modell. Sie können Hinweise auf Abweichungen von Linearitätsannahmen und Verbesserungsmöglichkeiten durch Transformation von Ausgangsgrössen geben.

f Die Funktion `regr` ruft weitere Funktionen auf, die auch einzeln benützt werden können für Resultate von `lm` und `anova` (?).

g **Funktion `drop1.mlm`.** Zunächst gibt es da eine Funktion `drop1.mlm`, die man aufrufen kann mit

```
> drop1(t.r)
```

Sie liefert die „Type III" Tests, prüft also, ob die einzelnen Terme des Modells weggelassen werden können, ohne dass sich die Anpassung signifikant verschlechtert.

h **Funktion `summary.mreg`.** Eine Zusammenfassung der Koeffizienten in Form einer Tabelle, die der Matrix $\boldsymbol{\beta}$ entspricht und somit alle Zielgrössen umfasst, erhält man durch

```
> summary.mreg(t.r)
```

Die Funktion liefert zudem eine analoge Tabelle für die Standardfehler und die P-Werte, die angeben, ob ein einzelner Koeffizient signifikant von 0 verschieden ist (ob also die entsprechende Ausgangsgrösse für eine bestimmte Zielgrösse aus dem Modell weggelassen werden kann – eine univariate Betrachtungsweise).

Literaturverzeichnis

Anderberg, M. R. (1973). *Cluster Analysis for Applications*, Academic Press, N. Y.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, Wiley, N. Y.

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, Springer-Verlag, N. Y.

Becker, R. A., Chambers, J. M. and Wilks, A. R. (1988). *The S Language; A Programming Environment for Data Analysis and Graphics*, Wadsworth & Brooks/Cole, Pacific Grove.

Bilodeau, M. and Brenner, D. (1999). *Theory of Multivariate Statistics*, Springer Texts in Statistics, Springer-Verlag, New York.

Bock, H. H. (1974). *Automatische Klassifikation*, Vandenhoeck & Rupprecht, Göttingen.

Bollmann, J., Henderiks, J. and Brabec, B. (2002). Global calibration of gephyrocapsa coccolith abundance in holocene sediments for paleotemperature assessment, *Paleoceanography* **17**(3): 1035.

Bortz, J. (1977). *Lehrbuch der Statistik für Sozialwissenschaftler*, Springer Lehrbücher, Springer, Berlin.

Brown, P. J. (1993). *Measurement, Regression, and Calibration*, Clarendon Press, Oxford, U.K.

Chambers, J. M. (1998). *Programming with Data; A Guide to the S Language*, Springer-Verlag, New York.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S*, Wadsworth & Brooks/Cole.

Chatfield, C. and Collins, A. J. (1980). *Introduction to Multivariate Analysis*, Science Paperbacks, Chapman and Hall, London.

Cleveland, W. S. (1993). *Visualizing Data*, Hobart Press, Summit, New Jersey. *2 Ex.*

Cleveland, W. S. (1994). *The Elements of Graphing Data*, 2nd edn, Hobart Press, Summit, New Jersey.

Cooley, W. W. and Lohnes, P. R. (1971). *Multivariate Data Analysis*, Wiley, New York.

Deichsel, G. and Trampisch, H. J. (1985). *Clusteranalyse und Diskriminanzanalyse*, VEB Gustav Fischer Verlag (Stuttgart).

Everitt, B. (1980). *Cluster Analysis, Second Edition*, Halsted Press, Wiley.

Everitt, B. S. (1978). *Graphical Techniques for Multivariate Data*, Heinemann Educational Books.

Fahrmeir, L., Hamerle, A. and Tutz, G. (eds) (1996). *Multivariate statistische Verfahren*, 2nd edn, de Gruyter, Berlin.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Ann. Eugenics* **7**: 179–184.

Flury, B. (1997). *A first course in multivariate statistics*, Springer texts in statistics, Springer-Verlag, NY.

Flury, B. und Riedwyl, H. (1983). *Angewandte multivariate Statistik*, Gustav Fischer, Stuttgart.

Friedman, Hastie and Tibshirani (2000). Additive logistic regression: a statistical view of boosting, *Annals of Statistics* **28**: 377–386.

Fuller, W. A. (1987). *Measurement Error Models*, Wiley, N. Y.

Gnanadesikan, R. (1997). *Methods for Statistical Data Analysis of Multivariate Observations*, Series in Probability and Statistics, 2nd edn, Wiley, NY.

Gordon, A. D. (1981). *Classification. Methods for the Exploratory Analysis of Multivariate Data*, Chapman & Hall, London.

Green, P. E. and Caroll, J. D. (1976). *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, New York.

Harman, H. H. (1960, 1967). *Modern Factor Analysis*, 2nd edn, University of Chicago Press.

Harris, R. J. (1975). *A Primer of Multivariate Statistics*, Academic Press, New York.

Hartigan, J. A. (1975). *Clustering algorithms*, Wiley.

Hastie, T. and Tibshirani, R. (1994). Discriminant analysis by gaussian mixtures, *Journal of the Royal Statistical Society B* **?**: ?

Jewell, P. L., Güsewell, S., Berry, N. R., Käuferle, D., Kreuzer, M. and Edwards, P. (2005). Vegetation patterns maintained by cattle grazing on a degraded mountain pasture. *Manuscript*

Johnson, N. L. and Kotz, S. (1972). *Continuous Multivariate Distributions*, A Wiley Publication in Applied Statistics, Wiley, New York.

Johnson, R. A. and Wichern, D. W. (1982, 1988, 1992). *Applied Multivariate Statistical Analysis*, Prentice Hall Series in Statistics, 3rd edn, Prentice Hall Int.,Englewood Cliffs,N.J.,USA.

Karson, M. J. (1982). *Multivariate Statistical Methods*, The Iowa State University Press, Ames.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, N. Y.

Kendall, M. G. (1957, 1961). *A Course in Multivariate Analysis*, Griffin's Statistical Monographs & Courses, No.2, 2nd edn, Charles Griffin, London.

Krzanowski, W. J. (2000). *Principles of Multivariate Analysis; A User's Perspective*, Oxford statistical science series; 3, 2nd edn, Oxford University Press, Oxford, UK.

Lawley, D. N. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method*, Butterworths Mathematical Texts, 2nd edn, Butterworths, London.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, N. Y.

Manly, B. F. J. (1986, 1990). *Multivariate Statistical Methods: A Primer*, Chapman and Hall, London.

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

Maxwell, A. E. (1977). *Multivariate Analysis in Behavioural Research*, Monographs on Applied Probability and Statistics, Chapman and Hall, London.

Morrison, D. F. (1967, 1976). *Multivariate Statistical Methods*, McGraw-Hill Series in Probability and Statistics, 2nd edn, McGraw-Hill Book Co., New York.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, Wiley, N. Y.

Rencher, A. C. (1995). *Methods of Multivariate Analysis*, Wiley, N. Y.

Rencher, A. C. (1998). *Multivariate Statistical Inference and Applications*, Wiley, N. Y.

Renner, R. M. (1993). The resolution of a compositional data set into mixtures of fixed source compositions, *Applied Statistics — Journal of the Royal Statistical Society C* **42**: 615–631.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge UK.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*, number 72 in *Monographs on Statistics and Applied Probability*, Chapman and Hall.

Seber, G. A. F. (1984). *Multivariate Observations*, Wiley, N. Y.

Sokal, R. R. and Sneath, P. H. A. (1963). *Principles of Numerical Taxonomy*, Freeman, San Francisco.

Späth, H. (1977). *Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion*, Oldenbourg; München, Wien.

Späth, H. (1983). *Cluster-Formation und -Analyse: Theorie, FORTRAN-Programme und Beispiele*, Oldenbourg; München, Wien.

Srivastava, M. S. and Carter, E. M. (1983). *An Introduction to Applied Multivariate Statistics*, North Holland.

Stahel, W. A. (2002). *Statistische Datenanalyse: Eine Einführung für Naturwissenschaftler*, 4. Aufl., Vieweg, Wiesbaden.

Steinhausen, D. and Langer, K. (1977). *Clusteranalyse: Einführung in Methoden und Verfahren der automatischen Klassifikation*, de Gruyter, Berlin.

Tatsuoka, M. M. (1971). *Multivariate Analysis: Techniques for Educational and Psychological Research*, Wiley, New York.

Timm, N. H. (2002). *Applied Multivariate Analysis*, Springer-Verlag, N. Y.

Tufte, E. R. (1983). *The Visual Display of Quantitative Information*, Graphics Press, Cheshire.

Tufte, E. R. (1990). *Envisioning Information*, Graphics Press, Cheshire.

Tufte, E. R. (1997). *Visual Explanations; Images and quantities, evidence and narrative*, Graphics Press, Cheshire.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-Plus*, Statistics and Computing, 3rd edn, Springer-Verlag, New York.

Venables, W. N. and Ripley, B. D. (2000). *S Programming*, Statistics and Computing, Springer-Verlag, New York.