Statistische Analyseverfahren
Abschnitt 3: Diskriminanzanalyse

Dr. Andreas Wünsche

TU Bergakademie Freiberg
Institut für Stochastik

Oktober/November 2019

3 Diskriminanzanalyse

3.1 Einführung

- Zielstellung einer Diskriminanzanalyse ist es, einen Merkmalsträger (ein Objekt) mit Hilfe der beobachteten Messwerte zu einer von mehreren Klassen (Gruppen, Populationen, ...) zuzuordnen, wobei in der Regel keine eindeutige deterministische Zuordnung mittels einfacher Entscheidungsregeln möglich ist.
- Zuerst soll der Fall behandelt werden, dass die unterschiedlichen Klassen durch jeweils **bekannte Wahrscheinlichkeitsverteilungen** der Merkmalszufallsvektoren charakterisiert werden.
- Danach wird auf den praktisch relevanteren Fall eingegangen, dass die unterschiedlichen **Wahrscheinlichkeitsverteilungen der Merkmalszufallsvektoren nicht vollständig bekannt** sind.
- Außerdem soll die **lineare Diskriminanzanalyse** von Fisher vorgestellt werden, bei der keine speziellen Verteilungen der Merkmalszufallsvektoren genutzt werden.

Beispiele

- Kredit-Scoring: Beurteilung der Kreditwürdigkeit für (z.B.) Neukunden, wobei ein Kunde anhand von bestimmten erhobenen Daten, wie z.B. Familienstand, Alter, Vermögen, Status als Arbeitsnehmer, Beschäftigungsdauer, etc. in die Klasse der kreditwürdigen oder kreditunwürdigen Kunden eingestuft werden soll.
- Im Zusammenhang mit dem R-Beispieldatensatz „Iris“ kann man das Problem betrachten, eine Schwertlilienpflanze anhand der gemessenen Größen Länge des Kelchblattes, Breite des Kelchblattes, Länge des Blütenblattes und Breite des Blütenblattes zu einer der drei Blumenarten „Iris setosa“ (Borsten-Schwertlilie), „Iris versicolor“ (Verschiedenfarbige Schwertlilie) und „Iris virginica“ (Virginische Schwertlilie) zuzuordnen.

Formales Vorgehen

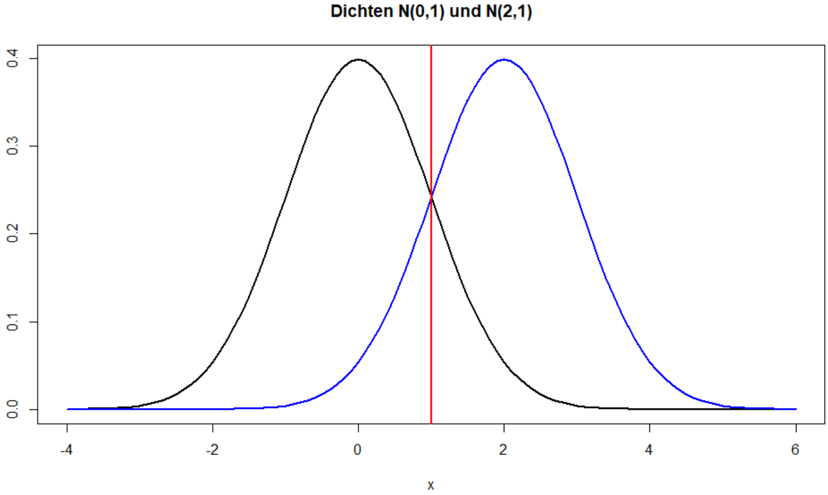
- **Aufgabenstellung:** Klassifikation (Klassierung), Zuordnung eines Merkmalsträgers zu einer von $g \geq 2$ Klassen $\Pi_1, \ldots, \Pi_g$ auf der Grundlage von „Messwerten" $\underline{\mathbf{x}}$ als Realisierungen von zufälligen $p-$dimensionalen Merkmalsvektoren $\underline{\mathbf{X}}_j$.
- **Ziel:** Die Zuordnung soll zu möglichst wenigen Fehlklassifikationen führen.
- **Formales Vorgehen:** Der Merkmalsraum (oft $\mathbb{R}^d$) wird in g disjunkte Regionen $R_1, \ldots, R_g$ eingeteilt (zerlegt).
- **Diskriminanzregel:** Der Merkmalsträger mit Merkmalsvektor $\underline{\mathbf{x}}$ („der Merkmalsvektor $\underline{\mathbf{x}}$") wird der Population Π_i $(i \in \{1, \ldots, g\})$ **genau dann** zugeordnet, wenn $\underline{\mathbf{x}} \in R_i$ gilt.
- Häufige **Annahme**: Die Verteilung des Merkmalsvektors für jede Klasse besitze eine Dichtefunktion f_j , $j = 1, \ldots, g$, bzgl. eines Maßes auf dem Merkmalsraum (oft die übliche Verteilungsdichte).

Beispiel für die Maximum-Likelihood-Diskriminanzregel

- 1 Merkmal, 2 Klassen, Normalverteilungen.
- **Geg.** $p = 1$, $g = 2$, $\Pi_1 = \mathrm{N}(\mu_1, \sigma^2)$, $\Pi_2 = \mathrm{N}(\mu_2, \sigma^2)$, $\mu_1 < \mu_2$; μ_1, μ_2, σ^2 bekannt.
- **Maximum-Likelihood-Diskriminanzregel:**
  Ordne dem Merkmalsträger mit Merkmalswert $x \in \mathbb{R}$ diejenige Klasse zu, deren Dichtefunktion im Punkt x maximal wird, bei Gleichheit kann man beliebig (messbar) zuordnen.
- Da hier für alle $x < \dfrac{\mu_1 + \mu_2}{2}$ gilt $f_1(x) > f_2(x)$, lautet die (oder besser: eine) Diskriminanzregel

$$
\begin{aligned}
x < \frac{\mu_1 + \mu_2}{2} &\longmapsto \Pi_1, \ \text{Zuordnung zu Klasse } 1; \\
x \geq \frac{\mu_1 + \mu_2}{2} &\longmapsto \Pi_2, \ \text{Zuordnung zu Klasse } 2.
\end{aligned}
$$

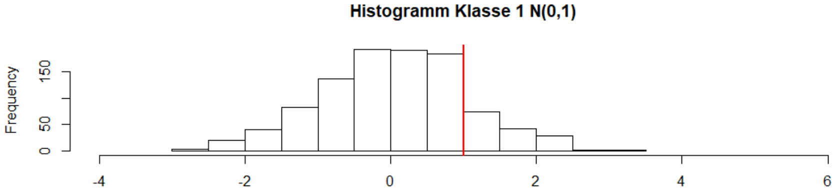
Grafik Dichtefunktionen mit $\mu_1 = 0, \mu_2 = 2, \sigma^2 = 1$

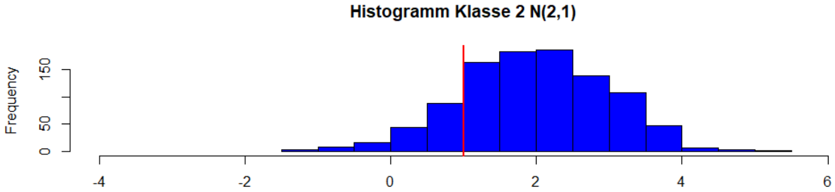


Simulationsstudie Maximum-Likelihood-Diskriminanzregel

- In einer Simulationsstudie wurden jeweils 1000 Realisierungen des Merkmals der Klasse 1 mit der Verteilung $\mathrm{N}(0,1)$ und der Klasse 2 mit der Verteilung $\mathrm{N}(2,1)$ erzeugt.
- Mit der Maximum-Likelihood-Diskriminanzregel werden Realisierungwerte < 1 der Klasse 1 zugeordnet, die anderen (d.h. ≥ 1) der Klasse 2. Die theoretische Wahrscheinlichkeit für eine Fehlklassifikation (eine Realisierung aus der Klasse 1 wird der Klasse 2 zugeordnet bzw. eine Realisierung aus der Klasse 2 wird der Klasse 1 zugeordnet) beträgt jeweils $1 - \Phi(1) \approx 0.159$.
- In der erzeugten Stichprobe werden 148 Realisierungen aus der Klasse 1 der Klasse 2 zugeordnet und 161 Realisierungen aus der Klasse 2 der Klasse 1 zugeordnet, also fehlerhaft klassifiziert.
- Bei einer Trenngrenze von 1.5 statt 1 zum Beispiel würden 74 Realisierungen aus Klasse 1 und 324 Realisierungen aus Klasse 2 fehlerhaft klassifiziert.

Histogramme Simulationsstudie





Beispiel für die Bayessche Diskriminanzregel

- **Geg.** $p = 1$, $g = 2$, $\Pi_1 = \mathrm{N}(\mu_1, \sigma^2)$, $\Pi_2 = \mathrm{N}(\mu_2, \sigma^2)$, $\mu_1 < \mu_2$; μ_1, μ_2, σ^2 bekannt, π_1, π_2 a-priori-Wahrscheinlichkeiten für die 1. bzw. 2. Klasse.
- **Bayessche Diskriminanzregel:** Ordne dem Merkmalsträger mit Merkmalswert $x \in \mathbb{R}$ diejenige Klasse zu, deren a-posteriori-Wahrscheinlichkeit $\pi(j|x)$ maximal wird, bei Gleichheit kann man beliebig (messbar) zuordnen.
- Es gelten

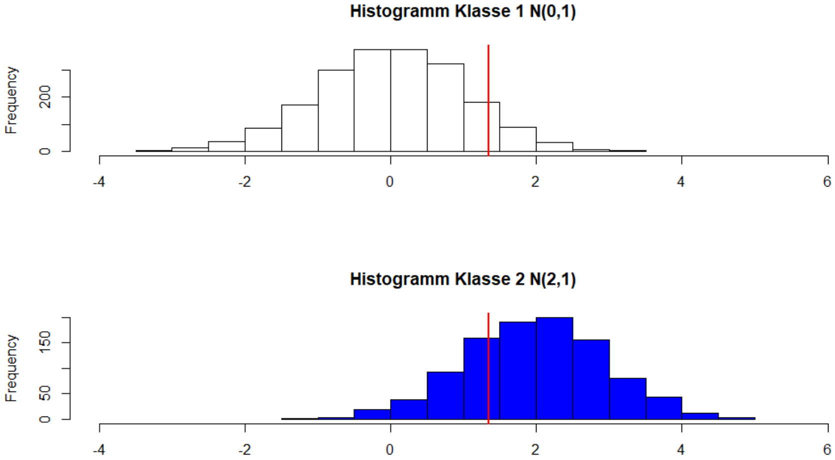
$$\pi(j|x) = \frac{f_j(x)\pi_j}{f_1(x)\pi_1 + f_2(x)\pi_2} \propto f_j(x)\pi_j\,, \quad j = 1, 2\,;$$

$$\pi(1|x) > \pi(2|x) \quad \Leftrightarrow \quad x < \frac{1}{2}(\mu_1 + \mu_2) + \frac{\sigma^2}{\mu_2 - \mu_1} \ln\left(\frac{\pi_1}{\pi_2}\right).$$

Simulationsstudie BAYESsche Diskriminanzregel

- In einer Simulationsstudie wurden 2000 Realisierungen des Merkmals der Klasse 1 mit der Verteilung $\mathrm{N}(0,1)$ und 1000 Realisierungen der Klasse 2 mit der Verteilung $\mathrm{N}(2,1)$ erzeugt. Dies entspricht der Situation mit a-priori-Wahrscheinlichkeiten $\pi_1 = \frac{2}{3}\,,\ \pi_2 = \frac{1}{3}\,.$
- Mit der BAYESschen Diskriminanzregel werden Realisierungwerte $< 1 + \frac{1}{2}\ln(2) \approx 1.347$ der Klasse 1 zugeordnet, die anderen der Klasse 2.
- In der erzeugten Stichprobe werden 177 Realisierungen aus der Klasse 1 der Klasse 2 zugeordnet und 255 Realisierungen aus der Klasse 2 der Klasse 1 zugeordnet (also fehlerhaft klassifiziert), dies entspricht einer relativen Häufigkeit von 0.144.
- Bei einer Trenngrenze von 1 statt 1.347 zum Beispiel würden 314 Realisierungen aus Klasse 1 und 154 Realisierungen aus Klasse 2 fehlerhaft klassifiziert, dies entspricht einer relativen Häufigkeit von 0.156.

Histogramme Simulationsstudie



3.2 Diskrimination bei bekannten Verteilungen

- Allgemeine **Maximum-Likelihood-Diskriminanzregel** (ML-Diskriminanzregel) bei bekannten Verteilungen:

$$\underline{\mathbf{x}} \in R_i \,, \quad \text{falls} \quad L(i;\underline{\mathbf{x}}) = \max_{j=1,\ldots,g} L(j;\underline{\mathbf{x}}) \qquad \text{mit}$$

$$L(j;\underline{\mathbf{x}}) := f_j(\underline{\mathbf{x}})\,, \quad j = 1,\ldots,g\,, \quad \text{(Likelihood-Funktion)}$$

  mit speziellen Vereinbarungen im Fall mehrfacher Maxima, so dass eine messbare Zuordnung realisiert wird.

- **Beispiel** $\Pi_1 = \begin{pmatrix} 0 & 1 \\ 0.5 & 0.5 \end{pmatrix}, \quad \Pi_2 = \begin{pmatrix} 0 & 1 \\ 0.25 & 0.75 \end{pmatrix};$

$$x = 0: \ L(1;0) = 0.5 > 0.25 = L(2;0)\,, \ \text{Zuordnung zu } \Pi_1\,;$$

$$x = 1: \ L(1;1) = 0.5 < 0.75 = L(2;1)\,, \ \text{Zuordnung zu } \Pi_2\,.$$

ML-Diskrimination für Normalverteilungen mit identischen regulären Kovarianzmatrizen

- **Satz 3.2.1** **Geg.** $\Pi_j = \mathrm{N}_p(\underline{\mu}_j, \underline{\underline{\boldsymbol{\Sigma}}})$, $\underline{\underline{\boldsymbol{\Sigma}}}$ regulär, $j = 1, \ldots, g$.
  Die ML-Diskriminanzregel ist dann gegeben durch: $\underline{\mathbf{x}} \in R_i \Leftrightarrow$
  $$(\underline{\mathbf{x}} - \underline{\mu}_i)^\mathsf{T} \underline{\underline{\boldsymbol{\Sigma}}}^{-1} (\underline{\mathbf{x}} - \underline{\mu}_i) = \min_{j=1,\ldots,g} (\underline{\mathbf{x}} - \underline{\mu}_j)^\mathsf{T} \underline{\underline{\boldsymbol{\Sigma}}}^{-1} (\underline{\mathbf{x}} - \underline{\mu}_j)$$
  und eine geeignete Vereinbarung im Fall mehrfacher Minima.
- **Def. 3.2.2** Die Zahl $d_M(\underline{\mathbf{x}}, \underline{\mu}_j) := \left((\underline{\mathbf{x}} - \underline{\mu}_j)^\mathsf{T} \underline{\underline{\boldsymbol{\Sigma}}}^{-1} (\underline{\mathbf{x}} - \underline{\mu}_j) \right)^{1/2}$ ist der MAHALANOBIS-Abstand von $\underline{\mathbf{x}}$ zum Erwartungswertvektor $\underline{\mu}_j$ (oft wird auch keine Quadratwurzel gezogen).
- **Bem.** Nimmt der MAHALANOBIS-Abstand zwischen dem Merkmalsvektor $\underline{\mathbf{x}}$ und dem Erwartungswertvektor $\underline{\mu}_i$ der $i-$ten Klasse das Minimum unter allen MAHALANOBIS-Abständen zwischen $\underline{\mathbf{x}}$ und $\underline{\mu}_j$ $(j = 1, \ldots, g)$ an, so wird der Merkmalsträger mit dem Merkmalsvektor $\underline{\mathbf{x}}$ der $i-$ten Klasse zugeordnet.

Allgemeiner Fall von 2 Klassen: Diskriminanzfunktion

- **Bem.**
  Im Fall mit $g = 2$ gilt: $\underline{\mathbf{x}} \in R_1$, falls $L(1; \underline{\mathbf{x}}) > L(2; \underline{\mathbf{x}})$, d.h. falls
  $$\ln \frac{L(1; \underline{\mathbf{x}})}{L(2; \underline{\mathbf{x}})} = \ln L(1; \underline{\mathbf{x}}) - \ln L(2; \underline{\mathbf{x}}) > 0\,.$$
- **Def. 3.2.3**
  Die Funktion
  $$h(\underline{\mathbf{x}}) := \ln \frac{L(1; \underline{\mathbf{x}})}{L(2; \underline{\mathbf{x}})} = \ln L(1; \underline{\mathbf{x}}) - \ln L(2; \underline{\mathbf{x}})\,, \qquad \underline{\mathbf{x}} \in \mathbb{R}^p\,,$$
  wird Diskriminanzfunktion genannt.
- **Bem.**
  Die Diskriminanzfunktion teilt $\mathbb{R}^p$ in zwei Teilmengen. Mit der ML-Diskriminanzregel liegt $\underline{\mathbf{x}}$ in R_1, wenn $h(\underline{\mathbf{x}}) > 0$ gilt.

Zwei Normalverteilungen mit gleichen Kovarianzmatrizen

- **Satz 3.2.4**
  Geg. $\Pi_j = \mathrm{N}_p(\underline{\mu}_j, \underline{\underline{\mathbf{\Sigma}}})$, $\underline{\underline{\mathbf{\Sigma}}}$ regulär, $j = 1, 2$.
  Dann gelten

$$\underline{\mathbf{x}} \in R_1 \;\Leftrightarrow\; h(\underline{\mathbf{x}}) = (\underline{\mu}_1 - \underline{\mu}_2)^{\mathsf{T}} \underline{\underline{\mathbf{\Sigma}}}^{-1} \left(\underline{\mathbf{x}} - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2) \right) > 0 \,.$$

$$\underline{\mathbf{x}} \in R_2 \;\Leftrightarrow\; h(\underline{\mathbf{x}}) \leq 0 \,.$$

  Folglich ist die Diskriminanzfunktion $h(\underline{\mathbf{x}})$ (affin) linear, ihre Nullstellenmenge $\{\underline{\mathbf{x}} \in \mathbb{R}^p \,:\, h(\underline{\mathbf{x}}) = 0\}$ beschreibt eine Hyperebene (die Trennebene) im Raum $\mathbb{R}^p$, die durch den Schwerpunkt (den Mittelpunkt) der beiden Erwartungswertvektoren $(\underline{\mu}_1 + \underline{\mu}_2)/2$ geht.
- **Bem.**
  Wird eine affin lineare Diskriminanzfunktion zur Klassifizierung genutzt, spricht man auch von „linearer Diskriminanzanalyse (LDA)“.

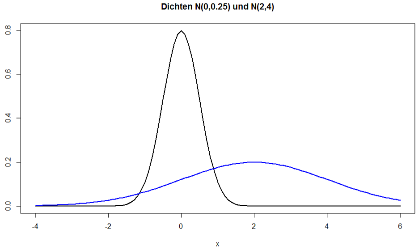
Zwei Normalverteilungen mit regulären Kovarianzmatrizen

Bem.

Besitzen die Normalverteilungen unterschiedliche Kovarianzmatrizen, führt die Maximum-Likelihood-Diskriminanzregel zu einer quadratischen Diskriminanzfunktion. Das entsprechende Verfahren wird auch „quadratische Diskriminanzanalyse (QDA)“ genannt.

Im univariaten Fall ($p = 1$) gilt dann

$$x \in R_1 \;\Leftrightarrow\; x^2\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right) - 2x\left(\frac{\mu_1}{\sigma_1^2} - \frac{\mu_2}{\sigma_2^2}\right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2}\right) < 2\ln\left(\frac{\sigma_2}{\sigma_1}\right).$$



Beispiel 3.2.5

Geg. $\Pi_j = \mathrm{N}_2(\underline{\mu}_j, \underline{\underline{\boldsymbol{\Sigma}}})\,;\;\; j = 1, 2\,;$

$$\underline{\underline{\boldsymbol{\Sigma}}} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}\,; \qquad \underline{\mu}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}\,; \qquad \underline{\mu}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\,.$$

Merkmalsträger mit Merkmalsvektor $\underline{\mathbf{x}} = \left(\frac{1}{2}, 1\right)^{\mathsf{T}}$.

BAYESsche Diskriminanzregel

- Wenn man bereits „Vorurteile“ bzw. Vorinformationen über die Zugehörigkeiten zu den g Klassen hat, kann man diese mit in die Diskriminanzregel aufnehmen.
- **Geg.** zusätzlich a-priori-Wahrscheinlichkeiten $\pi_1 \geq 0, \ldots, \pi_g \geq 0$ mit $\sum_{j=1}^{g} \pi_j = 1$ für die Klassen $\Pi_1, \ldots, \Pi_g$.

$\Rightarrow$ a-posteriori-Wahrscheinlichkeiten

$$\pi(i|\underline{\mathbf{x}}) = \frac{L(i;\underline{\mathbf{x}})\pi_i}{\sum_{j=1}^{g} L(j;\underline{\mathbf{x}})\pi_j} \propto L(i;\underline{\mathbf{x}})\pi_i\,, \quad i = 1, \ldots, g\,.$$

- BAYES**sche Diskriminanzregel:** Ordne den Merkmalsvektor $\underline{\mathbf{x}}$ zur Klasse i zu, d.h. $\underline{\mathbf{x}} \in R_i$, falls

$$\pi_i L(i;\underline{\mathbf{x}}) = \max_{j=1,\ldots,g} \pi_j L(j;\underline{\mathbf{x}})$$

(mit speziellen Vereinbarungen im Fall mehrfacher Maxima).

Bemerkungen zur BAYESschen Diskriminanzregel

- Die Maximum-Likelihood-Diskriminanzregel erhält man im Fall $\pi_1 = \ldots = \pi_g = \frac{1}{g}$.
- Im Spezialfall $g = 2$ führt das auf eine Verschiebung des kritischen Wertes der Diskriminanzfunktion: $\underline{\mathbf{x}} \in R_1 \;\Leftrightarrow\; h(\underline{\mathbf{x}}) > \ln(\pi_2/\pi_1)$.
- Insbesondere lautet die BAYESsche Diskriminanzregel bei 2 Klassen von p-dimensional normalverteilten Merkmalsvektoren mit gleichen Kovarianzmatrizen $\underline{\underline{\boldsymbol{\Sigma}}} = \underline{\underline{\boldsymbol{\Sigma}}}_1 = \underline{\underline{\boldsymbol{\Sigma}}}_2$ (vgl. Satz 3.2.4)

$$\underline{\mathbf{x}} \in R_1 \;\Leftrightarrow\; h(\underline{\mathbf{x}}) = (\underline{\mu}_1 - \underline{\mu}_2)^{\mathsf{T}} \underline{\underline{\boldsymbol{\Sigma}}}^{-1} \left(\underline{\mathbf{x}} - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2) \right) > \ln\left(\frac{\pi_2}{\pi_1}\right) ;$$

$$\underline{\mathbf{x}} \in R_2 \;\Leftrightarrow\; h(\underline{\mathbf{x}}) \le \ln\left(\frac{\pi_2}{\pi_1}\right) ;$$

bzw. mit der Diskriminanzfunktion $h_1(\underline{\mathbf{x}}) := h(\underline{\mathbf{x}}) - \ln\left(\frac{\pi_2}{\pi_1}\right)$:

Zuordnung zu Π_1 falls $h_1(\underline{\mathbf{x}}) > 0$, sonst zu Π_2 .

Fortsetzung Beispiel 3.2.5

Zusätzlich **geg.**

$$\pi_1 = \frac{2}{3}\,, \qquad \pi_2 = \frac{1}{3}\,.$$

3.3 Diskrimination, wenn Verteilungen bis auf Parameter bekannt sind

- Sind die Verteilungen der zufälligen Merkmalsvektoren für die einzelnen Klassen nicht bekannt, benötigt man eine Lernstichprobe (Trainingsstichprobe) von Merkmalsträgern mit beobachteten Merkmalsvektoren, für die die Zugehörigkeit zu einer Klasse bekannt sein muss. Dann kann man für weitere Merkmalsträger die Diskrimination z.B. mit Hilfe geschätzter Parameter durchführen.
- **Vor. 3.3.1**
  Es liegt eine Datenmatrix $\underline{\underline{\mathbf{x}}}$ vor, mit jeweils n_j Realisierungen der Population Π_j , $j = 1, \ldots, g$, d.h.

$$\underline{\underline{\mathbf{x}}} = (\underline{\underline{\mathbf{x}}}_1^{\mathsf{T}}, \ldots, \underline{\underline{\mathbf{x}}}_g^{\mathsf{T}})^{\mathsf{T}}, \quad \underline{\underline{\mathbf{x}}}_j \in \mathbb{R}^{n_j \times p}, \quad \underline{\underline{\mathbf{x}}}_j = (\underline{\mathbf{x}}_{j1}, \ldots, \underline{\mathbf{x}}_{jn_j})^{\mathsf{T}}.$$

  Mit diesen Daten soll eine passende Diskriminanzregel gelernt werden.

Prinzip der Stichproben-ML-Diskriminanzregel

Prinzip 3.3.2

(i) Die unbekannten Parameter von Π_j werden mit $\underline{\underline{\mathbf{x}}}_j$ geschätzt.

(ii) Danach benutzt man die ML-Diskriminanzregel zur Zuordnung mit geschätzten statt theoretischen Parametern.

Univariates Beispiel Stichproben-ML-Diskriminanzregel

3.3.3 Bsp.

Geg. $p = 1$, $g = 2$, $\Pi_1 = \mathrm{N}(\mu_1, \sigma^2)$, $\Pi_2 = \mathrm{N}(\mu_2, \sigma^2)$, $\mu_1 < \mu_2$;
μ_1, μ_2, σ^2 unbekannt;
Lernstichproben $\underline{\mathbf{x}}_1 = (x_{11}, \ldots, x_{1n_1})^\mathsf{T}$, $\underline{\mathbf{x}}_2 = (x_{21}, \ldots, x_{2n_2})^\mathsf{T}$;
Schätzwerte $\mu_j \approx \hat{\mu}_j = \overline{x}_j = (x_{j1} + \ldots + x_{jn_j})/n_j$, $j = 1, 2$.

Diskriminanzregel (Fall $\overline{x}_1 < \overline{x}_2$): $x \in R_1$ falls $x < \frac{1}{2}(\overline{x}_1 + \overline{x}_2)$.

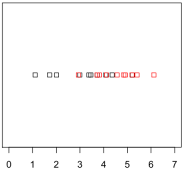
Zahlenbeispiel:

$\underline{\mathbf{x}}_1 = (4.09, 1.11, 3.73, 5.21, 2.99, 4.36, 3.46, 2.01, 1.72, 3.38)^\mathsf{T}$,
$\underline{\mathbf{x}}_2 = (4.57, 5.41, 3.82, 4.12, 5.20, 4.91, 6.12, 3.72, 2.93, 4.85)^\mathsf{T}$,
$\overline{x}_1 = 3.21$, $\overline{x}_2 = 4.57$,

$$\frac{1}{2}(\overline{x}_1 + \overline{x}_2) = 3.89\,.$$

$x \in R_1$ falls $x < 3.89$,

sonst $x \in R_2$.



Punktdiagramm Daten: $\underline{\mathbf{x}}_1$ schwarz, $\underline{\mathbf{x}}_2$ rot

Schätzung der Parameter der Klassenverteilungen

- **Vor. 3.3.4**
  Neben Vor. 3.3.1 sind die Daten in den g Klassen p-dimensional normalverteilt mit den Erwartungswertvektoren $\underline{\mu}_1, \ldots, \underline{\mu}_g$ und übereinstimmenden Kovarianzmatrizen $\underline{\underline{\boldsymbol{\Sigma}}} = \underline{\underline{\boldsymbol{\Sigma}}}_1 = \ldots = \underline{\underline{\boldsymbol{\Sigma}}}_g$, d.h. $\underline{\mathbf{X}}_j \sim \mathrm{N}_p(\underline{\mu}_j, \underline{\underline{\boldsymbol{\Sigma}}})$, $j = 1, \ldots, g$; $\underline{\mu}_j$ und $\underline{\underline{\boldsymbol{\Sigma}}}$ sind unbekannt.
- **Beh. 3.3.5**
  Schätzwerte aus erwartungstreuen Schätzfunktionen aus den Datenmatrizen $\underline{\underline{\mathbf{x}}}_j$ sind unter Vor. 3.3.4
  - für die Erwartungswertvektoren
  $$\hat{\underline{\mu}}_j = \overline{\underline{\mathbf{x}}}_j = \frac{1}{n_j} \sum_{k=1}^{n_j} \underline{\mathbf{x}}_{jk}, \quad j = 1, \ldots, g;$$
  - für die Kovarianzmatrizen
  $$\hat{\underline{\underline{\boldsymbol{\Sigma}}}}_j = \underline{\underline{\mathbf{s}}}_j = \frac{1}{n_j - 1} \sum_{k=1}^{n_j} (\underline{\mathbf{x}}_{jk} - \overline{\underline{\mathbf{x}}}_j)(\underline{\mathbf{x}}_{jk} - \overline{\underline{\mathbf{x}}}_j)^{\mathsf{T}}, \quad j = 1, \ldots, g.$$

Schätzung der gemeinsamen Kovarianzmatrix

Beh. 3.3.6

Eine geeignete Schätzmatrix (mit einer erwartungstreuen Schätzung) der gemeinsamen Kovarianzmatrix $\underline{\underline{\boldsymbol{\Sigma}}}$ ist unter Vor. 3.3.4 und mit $n = n_1 + \ldots + n_g$ gegeben durch

$$\begin{aligned}\underline{\underline{\hat{\boldsymbol{\Sigma}}}} = \underline{\underline{\tilde{\mathbf{s}}}} &= \frac{1}{n-g} \sum_{j=1}^{g} (n_j - 1) \underline{\underline{\mathbf{s}}}_j \\ &= \frac{1}{n-g} \sum_{j=1}^{g} \sum_{k=1}^{n_j} (\underline{\mathbf{x}}_{jk} - \underline{\overline{\mathbf{x}}}_j)(\underline{\mathbf{x}}_{jk} - \underline{\overline{\mathbf{x}}}_j)^{\mathsf{T}} \\ &=: \frac{1}{n-g} \underline{\underline{\mathbf{w}}} \,.\end{aligned}$$

Beispiel 3.3.7 (Simulation für Beispiel 3.2.5)

Lernstichprobe Klasse 1

k	x_{1k1}	x_{1k2}
1	1.83	-0.41
2	0.72	-0.57
3	0.64	0.92
4	1.09	2.66
5	3.25	2.49
6	1.83	1.08
7	2.31	-0.04
8	3.50	2.32
9	2.17	3.36
10	0.57	1.79
$\overline{\underline{\mathbf{x}}}_1 =$	1.791	1.360
$\underline{\mathbf{s}}_1^2 =$	1.102	1.907
$r_1 =$	0.336	

Lernstichprobe Klasse 2

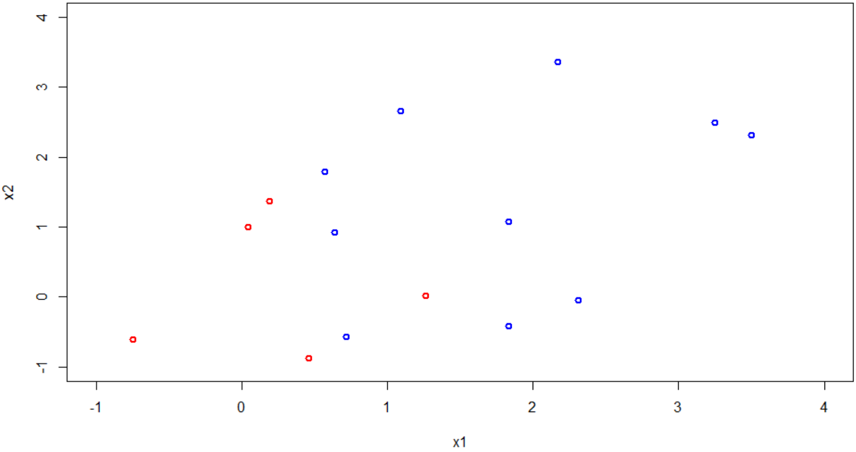
k	x_{2k1}	x_{2k2}
1	-0.75	-0.60
2	1.26	0.02
3	0.04	1.00
4	0.19	1.37
5	0.46	-0.88
$\overline{\underline{\mathbf{x}}}_2 =$	0.240	0.182
$\underline{\mathbf{s}}_2^2 =$	0.528	0.962
$r_2 =$	0.053	

$$\Rightarrow \quad \underline{\underline{\tilde{\mathbf{s}}}} = \begin{pmatrix} 0.925 & 0.348 \\ 0.348 & 1.616 \end{pmatrix}$$

$$\underline{\underline{\tilde{\mathbf{s}}}}^{-1} = \begin{pmatrix} 1.176 & -0.254 \\ -0.254 & 0.673 \end{pmatrix}$$

Streudiagramm Beispiel 3.3.7

blau: Merkmalsvektoren Klasse 1 , rot: Merkmalsvektoren Klasse 2.



Stichproben-ML-Diskriminanzregel

Beh. 3.3.8
Unter obigen Voraussetzungen 3.3.4 gilt

- die folgende Stichproben-ML-Diskriminanzregel: $\underline{\mathbf{x}} \in R_i \;\Leftrightarrow$

$$\underline{\mathbf{x}} \in R_i \;\Leftrightarrow\; (\underline{\mathbf{x}} - \underline{\hat{\mu}}_i)^{\mathsf{T}} \underline{\underline{\hat{\boldsymbol{\Sigma}}}}^{-1} (\underline{\mathbf{x}} - \underline{\hat{\mu}}_i) = \min_{j=1,\dots,g} (\underline{\mathbf{x}} - \underline{\hat{\mu}}_j)^{\mathsf{T}} \underline{\underline{\hat{\boldsymbol{\Sigma}}}}^{-1} (\underline{\mathbf{x}} - \underline{\hat{\mu}}_j)$$

  und eine geeignete Vereinbarung im Fall mehrfacher Minima.

- Im Spezialfall von 2 Klassen erhält man mit der aus der Lernstichprobe geschätzten („gelernten") Diskriminanzfunktion

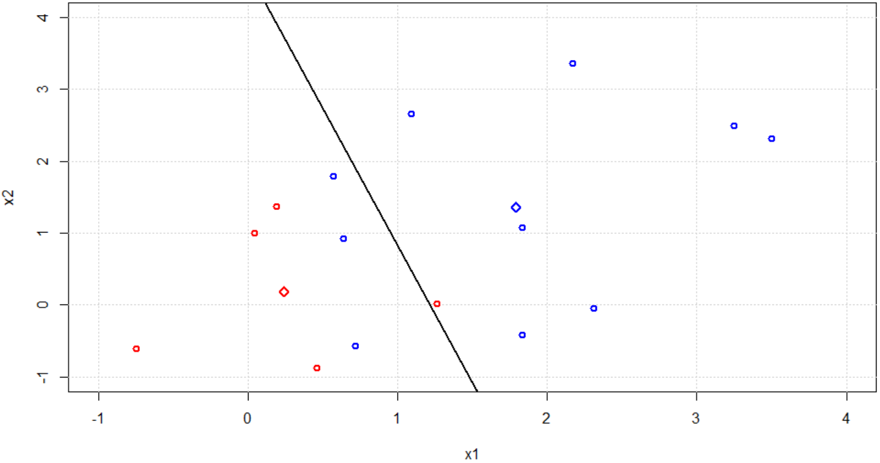
$$\hat{h}(\underline{\mathbf{x}}) = (\underline{\overline{\mathbf{x}}}_1 - \underline{\overline{\mathbf{x}}}_2)^{\mathsf{T}} \underline{\underline{\tilde{\mathbf{s}}}}^{-1} \left(\underline{\mathbf{x}} - \frac{1}{2} (\underline{\overline{\mathbf{x}}}_1 + \underline{\overline{\mathbf{x}}}_2) \right)$$

  die (oder besser: eine) Zuordnungsregel

$$\underline{\mathbf{x}} \in R_1 \quad \Leftrightarrow \quad \hat{h}(\underline{\mathbf{x}}) > 0\,, \qquad \underline{\mathbf{x}} \in R_2 \quad \Leftrightarrow \quad \hat{h}(\underline{\mathbf{x}}) \leq 0\,.$$

Streudiagramm Beispiel 3.3.7 mit Trenngerade

blau: Merkmalsvektoren Klasse 1 , rot: Merkmalsvektoren Klasse 2 ;
Rauten: Mittelwertvektoren.



BAYESsche Stichprobendiskriminanzregel (2 Klassen)

- **Beh. 3.3.9**
  Unter obigen Voraussetzungen 3.3.4 für 2 Klassen gilt mit a-priori-Wahrscheinlichkeiten π_1, π_2 mit der angepassten geschätzten Diskriminanzfunktion

$$\hat{h}_1(\underline{\mathbf{x}}) = \hat{h}(\underline{\mathbf{x}}) - \ln\left(\frac{\pi_2}{\pi_1}\right)$$

  die BAYESsche Stichprobendiskriminanzregel

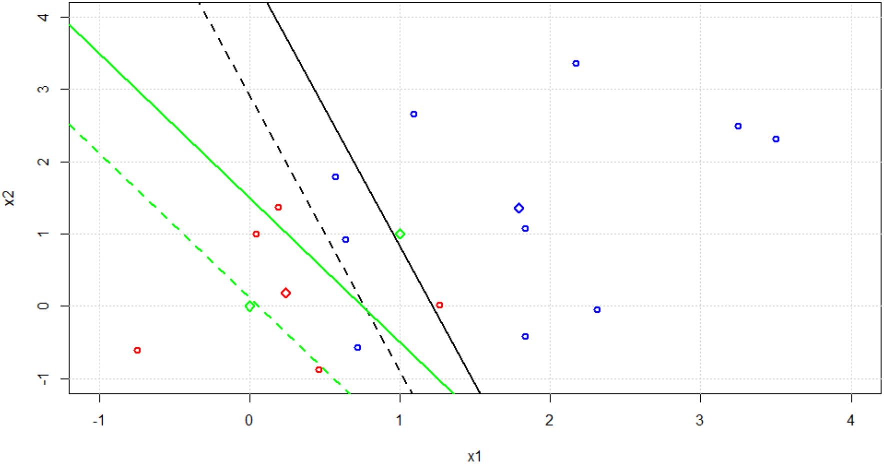
$$\underline{\mathbf{x}} \in R_1 \quad \Leftrightarrow \quad \hat{h}_1(\underline{\mathbf{x}}) > 0\,, \qquad \underline{\mathbf{x}} \in R_2 \quad \Leftrightarrow \quad \hat{h}_1(\underline{\mathbf{x}}) \le 0\,.$$

- **Bem.**
  Möglicherweise werden die a-priori-Wahrscheinlichkeiten π_1, π_2 auch aus der Stichprobe geschätzt (und in der geschätzten Diskriminanzfunktion statt π_1, π_2 genutzt):

$$\hat{\pi}_1 = \frac{n_1}{n_1 + n_2}\,, \qquad \hat{\pi}_2 = \frac{n_2}{n_1 + n_2}\,.$$

Streudiagramm Bsp. 3.3.7 mit BAYESscher Trenngerade

blau: Merkmalsvektoren Klasse 1 , rot: Merkmalsvektoren Klasse 2 ;
gestrichelt: Trenngerade BAYESsche Diskriminanzregel;
grün: Erwartungswertvektoren, Trenngeraden theoretische Verteilungen.



3.4 Einige relevante statistische Tests

3.4.1 Test auf Normalverteilung

- Verschiedene Apassungstests können für multivariate Daten verallgemeinert werden. Dies trifft auch auf die Fragestellung der Überprüfung einer vorliegenden Normalverteilung zu.
- Eine Variante einer Verallgemeinerung für multivariate Daten des Shapiro-Wilk-Testes auf Normalverteilung kann mit dem R-Paket `mvnormtest` (Befehl `mshapiro.test()`) realisiert werden.

Testergebnisse für Beispiel 3.3.7

- 1. Lernstichprobe (vgl. R-Skript für Bezeichnungen)

```
> mshapiro.test(t(bsp3_3_7_lk1))
    Shapiro-Wilk normality test
data: Z
W = 0.9108, p-value = 0.2865
```

- 2. Lernstichprobe

```
> mshapiro.test(t(bsp3_3_7_lk2))
    Shapiro-Wilk normality test
data: Z
W = 0.95874, p-value = 0.7992
```

- Zusammengefasste Lernstichprobe

```
> mshapiro.test(t(bsp3_3_7_lg))
    Shapiro-Wilk normality test
data: Z
W = 0.96272, p-value = 0.7395
```

3.4.2 Test auf Kovarianzhomogenität

- Voraussetzung zur Anwendung der MANOVA (siehe 3.4.3) und der Nutzung der oben angegebenen Diskriminanzfunktionen ist die Gleichheit (Homogenität) der Kovarianzmatrizen. Ein möglicher Test dazu für normalverteilte Daten ist der M-Test von Box.
- **Voraussetzung** Die Daten sind p-dimensional normalverteilt.
- **Hypothesen**
  - Die Kovarianzmatrizen sind in allen g Klassen gleich,
    $$\mathrm{H}_0 \;:\; \underline{\underline{\boldsymbol{\Sigma}}}_1 = \ldots = \underline{\underline{\boldsymbol{\Sigma}}}_g = \underline{\underline{\boldsymbol{\Sigma}}} \,.$$
  - Es gibt unterschiedliche Kovarianzmatrizen in den g Klassen,
    $$\mathrm{H}_\mathrm{A} \;:\; \text{es gibt } \; j \neq l \;:\; \underline{\underline{\boldsymbol{\Sigma}}}_j \neq \underline{\underline{\boldsymbol{\Sigma}}}_l \,.$$

- **Schätzer für Teilstichproben** $\hat{\underline{\mu}}_j := \overline{\underline{\mathbf{X}}}_j = \dfrac{1}{n_j} \sum_{k=1}^{n_j} \underline{\mathbf{X}}_{jk}$;

für Kovarianzmatrix $\underline{\underline{\mathbf{S}}}_j := \dfrac{1}{n_j - 1} \sum_{k=1}^{n_j} (\underline{\mathbf{X}}_{jk} - \overline{\underline{\mathbf{X}}}_j)(\underline{\mathbf{X}}_{jk} - \overline{\underline{\mathbf{X}}}_j)^\mathsf{T}$.

Fortsetzung M-Test von BOX

- **Gepoolte Kovarianzmatrix**

$$\underline{\underline{\tilde{\mathbf{S}}}} := \frac{1}{n-g}\sum_{k=1}^{g}(n_k-1)\underline{\underline{\mathbf{S}}}_k =: \frac{1}{n-g}\underline{\underline{\mathbf{W}}},$$

$\underline{\underline{\mathbf{W}}}$ ist die Inner-Gruppen-Streumatrix, auch Inner-Gruppen-SPP-Matrix (**w**ithin groups **s**um of **s**quares and **p**roducts).

- **Testgröße**

$$\begin{aligned} T &= (1-c)\left[(n-g)\ln\det\underline{\underline{\tilde{\mathbf{S}}}} - \sum_{k=1}^{g}(n_k-1)\ln\det\underline{\underline{\mathbf{S}}}_k\right] \\ &= (1-c)\sum_{k=1}^{g}(n_k-1)\ln\det\left(\underline{\underline{\mathbf{S}}}_k^{-1}\underline{\underline{\tilde{\mathbf{S}}}}\right) \end{aligned}$$

$$\text{mit} \quad c = \frac{2p^2+3p-1}{6(p+1)(g-1)}\left(\sum_{k=1}^{g}\frac{1}{n_k-1} - \frac{1}{n-g}\right).$$

Fortsetzung M-Test von Box

- **Asymptotische Verteilung**

$$T \overset{\text{asymt.}}{\sim} \chi^2_{p(p+1)(g-1)/2}$$

Die Approximation ist gut, falls $n_k > 20\,,\; g \leq 5\,, p \leq 5\,.$

- **Kritischer Bereich**

$$K = \left\{ t \in \mathbb{R}\,:\; t > \chi^2_{p(p+1)(g-1)/2\,;\,1-\alpha} \right\}\,.$$

Testergebnisse für Beispiel 3.3.7

(vgl. R-Skript für Berechnungen)

- $p(p+1)(g-1)/2 = 3$;
- $\chi^2_{p(p+1)(g-1)/2\,;\,1-\alpha} \approx 7.82$ für $\alpha = 0.05$;
- $c \approx 0.205$;
- $\underline{\underline{\mathbf{s}}}_1 \approx \begin{pmatrix} 1.102 & 0.486 \\ 0.486 & 1.907 \end{pmatrix}$, $\underline{\underline{\mathbf{s}}}_2 \approx \begin{pmatrix} 0.528 & 0.038 \\ 0.038 & 0.962 \end{pmatrix}$,

  $\underline{\underline{\tilde{\mathbf{s}}}} \approx \begin{pmatrix} 0.925 & 0.348 \\ 0.348 & 1.616 \end{pmatrix}$;
- $\det \underline{\underline{\tilde{\mathbf{S}}}} \approx 1.374$, $\quad \det \underline{\underline{\mathbf{S}}}_1 \approx 1.865$, $\quad \det \underline{\underline{\mathbf{S}}}_2 \approx 0.506$;
- Realisierungswert der Testgröße $t \approx 0.990$

$\Rightarrow$ $t < \chi^2_{p(p+1)(g-1)/2\,;\,1-\alpha}$, $t \notin K$, H_0 wird nicht abgelehnt, es gibt keine signifikanten Unterschiede zwischen den Kovarianzmatrizen.

3.4.3 Test auf Gleichheit der Erwartungswertvektoren

- **Bem.**
  Dies ist eine Aufgabenstellug der multivariaten Varianzanalyse (MANOVA). Analog zum univariaten Fall wird die Streuung innerhalb der Klassen verglichen mit der Streuung zwischen den Klassen.
- **Voraussetzungen**
  - Die Daten sind p-dimensional normalverteilt.
  - Die Kovarianzmatrizen in den g Klassen stimmen überein, d.h.
    $$\underline{\underline{\boldsymbol{\Sigma}}}_1 = \underline{\underline{\boldsymbol{\Sigma}}}_2 = \ldots = \underline{\underline{\boldsymbol{\Sigma}}}_g = \underline{\underline{\boldsymbol{\Sigma}}}\,.$$
- **Hypothesen**
  - Die Erwartungswertvektoren sind in allen g Klassen gleich (in diesem Fall ist die Diskriminanzanalyse nicht sinvoll),
    $$\mathrm{H}_0 \;:\; \underline{\mu}_1 = \ldots = \underline{\mu}_g\,.$$
  - Es gibt unterschiedliche Erwartungswertvektoren in den g Klassen (in diesem Fall ist die Diskriminanzanalyse sinvoll),
    $$\mathrm{H}_\mathrm{A} \;:\; \text{es gibt } \; j \neq l \;:\; \underline{\mu}_j \neq \underline{\mu}_l\,.$$

Forts. Test auf Gleichheit der Erwartungswertvektoren

- **Streuungsmatrix innerhalb der Klassen**

$$\underline{\underline{\mathbf{W}}} := \sum_{j=1}^{g}(n_j - 1)\underline{\underline{\mathbf{S}}}_j = \sum_{j=1}^{g}\sum_{k=1}^{n_j}(\underline{\mathbf{X}}_{jk} - \overline{\underline{\mathbf{X}}}_j)(\underline{\mathbf{X}}_{jk} - \overline{\underline{\mathbf{X}}}_j)^{\mathsf{T}};$$

- **Streuungsmatrix zwischen den Klassen**

$$\underline{\underline{\mathbf{B}}} := \sum_{j=1}^{g} n_j(\overline{\underline{\mathbf{X}}}_j - \overline{\underline{\mathbf{X}}})(\overline{\underline{\mathbf{X}}}_j - \overline{\underline{\mathbf{X}}})^{\mathsf{T}} \quad \text{mit} \quad \overline{\underline{\mathbf{X}}} = \frac{1}{n}\sum_{j=1}^{g}\sum_{k=1}^{n_j}\underline{\mathbf{X}}_{jk};$$

- **totale Streuungsmatrix**

$$\underline{\underline{\mathbf{S}}}_{total} := \sum_{j=1}^{g}\sum_{k=1}^{n_j}(\underline{\mathbf{X}}_{jk} - \overline{\underline{\mathbf{X}}})(\underline{\mathbf{X}}_{jk} - \overline{\underline{\mathbf{X}}})^{\mathsf{T}};$$

- **Streuungszerlegung**

$$\underline{\underline{\mathbf{S}}}_{total} = \underline{\underline{\mathbf{W}}} + \underline{\underline{\mathbf{B}}};$$

Forts. Test auf Gleichheit der Erwartungswertvektoren

- **Testgrößen** basieren z.B. auf der Spur der Matrix $\underline{\underline{\mathbf{W}}}^{-1}\,\underline{\underline{\mathbf{B}}}$ oder auf

$$\Lambda = \frac{\det \underline{\underline{\mathbf{W}}}}{\det\left(\underline{\underline{\mathbf{B}}} + \underline{\underline{\mathbf{W}}}\right)} = \det\left(\underline{\underline{\mathbb{I}}} + \underline{\underline{\mathbf{W}}}^{-1}\underline{\underline{\mathbf{B}}}\right)^{-1} .$$

- **Testgröße im Zweigruppenfall ($g = 2$)**

$$T = \frac{(n-p-1)n_1 n_2}{pn(n-2)}\left(\underline{\overline{\mathbf{X}}}_1 - \underline{\overline{\mathbf{X}}}_2\right)^{\mathsf{T}}\underline{\underline{\tilde{\mathbf{S}}}}^{-1}\left(\underline{\overline{\mathbf{X}}}_1 - \underline{\overline{\mathbf{X}}}_2\right).$$

- **Kritischer Bereich**

$$K = \left\{t \in \mathbb{R} \,:\, t > F_{p;n-p-1;1-\alpha}\right\} .$$

Testergebnisse für Beispiel 3.3.7

(vgl. R-Skript für Berechnungen)

- $F_{p;n-p-1;1-\alpha} \approx 3.89$ für $\alpha = 0.05$.
- Realisierungswert der Testgröße $t \approx 4.364$.

$\Rightarrow$ $t \in K$, H_0 wird abgelehnt, es gibt signifikante Unterschiede zwischen den Erwartungswertvektoren.

3.5 Wahrscheinlichkeit für Fehlklassifikation

- Bei bekannten Verteilungen kann man Wahrscheinlichkeiten für eine Fehlklassifikation (theoretisch) berechnen.
- Ist

$$p_{ij} := \int_{R_i} L(j; \underline{\mathbf{x}}) \, \mathrm{d}\underline{\mathbf{x}} = \int_{R_i} f_j(\underline{\mathbf{x}}) \, \mathrm{d}\underline{\mathbf{x}}$$

  die Wahrscheinlichkeit dafür, dass ein Merkmalsträger aus der j-ten Klasse der i-ten Klasse zugeordnet wird, dann ist p_{ij} für $i \neq j$ eine Fehlklassifikationswahrscheinlichkeit .
- Die Fehlklassifikationswahrscheinlichkeiten sollten möglichst klein sein.
- Sind die Verteilungen der Populationen $\Pi_1, \ldots, \Pi_g$ nicht bekannt, können verschiedene Ansätze zur Schätzung dieser Wahrscheinlichkeiten genutzt werden.

3.5.1 Nutzung geschätzter Parameter

- **Bsp.** **Geg.** $g = 2\,,\ \Pi_i = \mathrm{N}_p(\underline{\mu}_i, \underline{\underline{\boldsymbol{\Sigma}}})\,,\ i = 1, 2\,.$
- Die Diskriminanzfunktion (bei bekannten Parametern) lautet dann

$$h(\underline{\mathbf{x}}) = (\underline{\mu}_1 - \underline{\mu}_2)^{\mathsf{T}} \underline{\underline{\boldsymbol{\Sigma}}}^{-1} \left(\underline{\mathbf{x}} - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2) \right) , \quad \underline{\mathbf{x}} \in \mathbb{R}^p\,.$$

- Für einen Zufallsvektor $\underline{\mathbf{X}}$ aus Π_1 gilt $h(\underline{\mathbf{X}}) \sim \mathrm{N}_1\left(\frac{1}{2}\Delta^2, \Delta^2\right)$,

ist $\underline{\mathbf{X}}$ aus Π_2 gilt $h(\underline{\mathbf{X}}) \sim \mathrm{N}_1\left(-\frac{1}{2}\Delta^2, \Delta^2\right)$, mit

$$\Delta^2 := (\underline{\mu}_1 - \underline{\mu}_2)^{\mathsf{T}} \underline{\underline{\boldsymbol{\Sigma}}}^{-1} (\underline{\mu}_1 - \underline{\mu}_2)$$

(Quadrat des Mahalanobis-Abstandes zwischen $\underline{\mu}_1$ und $\underline{\mu}_2$).

Fortsetzung Beispiel

- Hieraus folgt

$$p_{12} = \mathbb{P}(h(\underline{\mathbf{X}}) > 0|\Pi_2) = \Phi\left(-\frac{\Delta}{2}\right) = p_{21} = \mathbb{P}(h(\underline{\mathbf{X}}) < 0|\Pi_1)\,.$$

- Mit geschätzten Parametern erhält man mit

$$\hat{\Delta}^2 = (\overline{\underline{\mathbf{x}}}_1 - \overline{\underline{\mathbf{x}}}_2)^{\mathsf{T}} \underline{\underline{\tilde{\mathbf{s}}}}^{-1} (\overline{\underline{\mathbf{x}}}_1 - \overline{\underline{\mathbf{x}}}_2) \quad \text{und}$$

$$\underline{\underline{\tilde{\mathbf{s}}}} = \frac{1}{n_1 + n_2 - 2}\left((n_1 - 1)\underline{\underline{\mathbf{s}}}_1 + (n_2 - 1)\underline{\underline{\mathbf{s}}}_2\right) :$$

$$\hat{p}_{12} = \hat{p}_{21} = \Phi\left(-\frac{\hat{\Delta}}{2}\right)\,.$$

- Für Beispiel 3.2.5 bzw. 3.3.7

$$p_{12} = \Phi\left(-\frac{\sqrt{1.5}}{2}\right) = 0.2701\,, \qquad \hat{p}_{12} = \Phi\left(-\frac{1.684}{2}\right) = 0.1998\,.$$

3.5.2 Resubstitutionsmethode

- **Prinzip** Man wendet die aus einer Lernstichprobe konstruierte Diskriminanzregel auf die Lernstichprobe selber an und bestimmt die relative Häufigkeit von Fehlklassifikationen.
- Sei n_{ij} die Anzahl der n_j Individuen von Π_j, deren Merkmalsvektor $\underline{\mathbf{x}}$ in R_i liegt, die also zu Π_i zugeordnet werden müssen.
- Dann ist eine Schätzung der indivuellen Fehlerraten gegeben durch

$$\hat{p}_{ij} = \frac{n_{ij}}{n_j}\,, \quad i \neq j\,.$$

- Für Beispiel 3.3.7 (ML-Diskriminanzregel, siehe Streudiagramm Folie 31)

$$n_{12} = 1\,, \quad n_{21} = 3\,, \qquad \hat{p}_{12} = \frac{1}{5} = 0.2\,, \quad \hat{p}_{21} = \frac{3}{10} = 0.3\,.$$

3.5.3 Cross-Validation-Prinzip („jack-knifing“)

- **Ausgangspunkt** Die in 3.5.1 und 3.5.2 vorgestellten Methoden sind oft zu optimistisch, da diesselben Daten, die die Diskriminanzregel definieren, zu deren Bewertung herangezogen werden.
- Cross-Validation (Kreuzvalidierung) ist eine Methode zur Bewertung von statistischen Verfahren, bei der Teile des (bekannten) Datenmaterials nicht zur Konstruktion des Verfahrens, sondern zu seiner Kontrolle genutzt werden.
- Wird jede einzelne Beobachtung einmal zur Kontrolle genutzt, spricht man auch von einer Leave-One-Out-Methode oder Leave-One-Out-Kreuzvalidierung.

Leave-One-Out-Methode

- **Geg.** $\underline{\underline{\mathbf{x}}}^{\mathrm{T}} = (\underline{\underline{\mathbf{x}}}_1^{\mathrm{T}}, \ldots, \underline{\underline{\mathbf{x}}}_j^{\mathrm{T}}, \ldots, \underline{\underline{\mathbf{x}}}_g^{\mathrm{T}})$.
- **Vorgehen**
  (i) Streiche $r-$tes Datum $\underline{\mathbf{x}}_{jr}$ aus $\underline{\underline{\mathbf{x}}}_j$.
  (ii) Bestimme die Diskriminanzregel mit den verbleibenden $n-1$ Daten, das Ergebnis sind die Regionen $R_1^{(jr)}, \ldots, R_g^{(jr)}$.
  (iii) Wende die Diskriminanzregel auf das gestrichene Datum $\underline{\mathbf{x}}_{jr}$ an; mache dies n_j mal, d.h. für $r = 1, \ldots, n_j$.
  (iv) Sei n_{ij}^* die Anzahl der Fälle, bei denen $\underline{\mathbf{x}}_{jr}$ aus Π_j in $R_i^{(jr)}$ liegt, d.h. Π_i zugeordnet werden würde.
  (v) Die gesuchten Schätzungen sind

$$\hat{p}_{ij} = \frac{n_{ij}^*}{n_j} \, .$$

3.6 Fishers lineare Diskriminanzfunktion

- **Bem.** Bei dieser Methode wird der Verteilungstyp der Populationen $\Pi_1, \ldots, \Pi_g$ **nicht** vorausgesetzt.
- **Prinzip** Man finde eine lineare Funktion der Datenmatrix, d.h. einen Vektor $\underline{\mathbf{a}} \in \mathbb{R}^p$ mit

$$\underline{\mathbf{z}} := \begin{pmatrix} \underline{\underline{\mathbf{x}}}_1 \, \underline{\mathbf{a}} \\ \vdots \\ \underline{\underline{\mathbf{x}}}_g \, \underline{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \underline{\mathbf{z}}_1 \\ \vdots \\ \underline{\mathbf{z}}_g \end{pmatrix} \qquad (\underline{\mathbf{z}}_j \in \mathbb{R}^{n_j}),$$

so dass in $\underline{\mathbf{z}}$

$$\frac{\text{Streuung zwischen den Gruppen}}{\text{Streuung innerhalb der Gruppen}} \stackrel{!}{=} \max .$$

Dann ist die Variation zwischen den Gruppen so groß wie möglich und die Variation innerhalb der Gruppen so klein wie möglich und damit sind die Gruppen bestmöglichst „getrennt“.

FISHERs lineare Diskriminanzfunktion

- Man kann zeigen, dass diese Aufgabenstellung äquivalent ist zu

$$\frac{\underline{\mathbf{a}}^{\mathsf{T}} \underline{\underline{\mathbf{b}}}\, \underline{\mathbf{a}}}{\underline{\mathbf{a}}^{\mathsf{T}} \underline{\underline{\mathbf{w}}}\, \underline{\mathbf{a}}} \overset{!}{=} \max_{\underline{\mathbf{a}}} .$$

  ($\underline{\underline{\mathbf{b}}}$ und $\underline{\underline{\mathbf{w}}}$ sind die entsprechenden Ausdrücke für die Variation zwischen (**b**etween) und innerhalb (**w**ithin) der Gruppen, siehe 3.4.3.) Man kann weiterhin zeigen, dass die Lösung dieses Problems folgendermaßen bestimmt werden kann.

- **Def.**
  Sei $\underline{\mathbf{a}}^* \neq \underline{\mathbf{0}}_p$ ein Eigenvektor zum größten Eigenwert von $\underline{\underline{\mathbf{w}}}^{-1} \underline{\underline{\mathbf{b}}}$.
  Die Funktion $\mathbb{R}^p \ni \underline{\mathbf{x}} \mapsto (\underline{\mathbf{a}}^*)^{\mathsf{T}} \underline{\mathbf{x}} \in \mathbb{R}$ heißt FISHERs lineare Diskriminanzfunktion.

Diskriminanzregel

Diskriminanzregel

$\underline{\mathbf{x}}$ wird der Klasse i zugeordnet, d.h. $\underline{\mathbf{x}} \in R_i$, falls für alle $j \neq i$ gilt

$$|(\underline{\mathbf{a}}^*)^\mathsf{T}\underline{\mathbf{x}} - (\underline{\mathbf{a}}^*)^\mathsf{T}\overline{\underline{\mathbf{x}}}_i| < |(\underline{\mathbf{a}}^*)^\mathsf{T}\underline{\mathbf{x}} - (\underline{\mathbf{a}}^*)^\mathsf{T}\overline{\underline{\mathbf{x}}}_j|,$$

d.h. der transformierte Vektor zu $\underline{\mathbf{x}}$ kommt dem transformierten Zentrum der Gruppe i am nächsten.