7.1 Allgemeine Lineare Diskriminanzanalyse

Kapitel

7 Diskriminanzanalyse

7.1 Allgemeine Lineare Diskriminanzanalyse

Allgemeine Lineare Diskriminanzanalyse

Zunächst stellen wir die **Kanonische Lineare Diskriminanzanalyse** (**LDA**) für den Zwei-Klassen-Fall als Spezialfall des Bayes-Ansatzes dar.

7.1 Allgemeine Lineare Diskriminanzanalyse

Bei der linearen Diskriminanzanalyse werden die folgenden Annahmen zugrunde gelegt:

(L1) Die Verteilungen innerhalb der Klassen sind Normalverteilungen, wobei unterschiedliche Erwartungswerte μ_i vorausgesetzt werden, aber die Kovarianzmatrizen Σ als identisch für alle Klassen angenommen werden.

(L2) Die Fehlklassifikationskosten sind gleich.

(L3) Die a-priori-Wahrscheinlichkeiten können unterschiedlich sein.

Daraus ergibt sich die folgende **Bayes-Regel**:

"Wähle Klasse 1, wenn $\frac{f_1(x)}{f_2(x)} > \frac{\pi_2}{\pi_1}$."

7.1 Allgemeine Lineare Diskriminanzanalyse

Definition (Einschub):
Eine p-dimensionale reelle Zufallvariable X ist normalverteilt mit Erwartungswertvektor μ und (positiv definiter) Kovarianzmatrix Σ, wenn sie eine Dichtefunktion der Form

$$f_X(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \tag{20}$$

besitzt. Hier bezeichnet $|\Sigma|$ die Determinante der Kovarianzmatrix.

7.1 Allgemeine Lineare Diskriminanzanalyse

Das ist unter Verwendung der Normalverteilungsdichten äquivalent zu:

$$\frac{\exp(-0.5(x-\mu_1)'\Sigma^{-1}(x-\mu_1))}{\exp(-0.5(x-\mu_2)'\Sigma^{-1}(x-\mu_2))} > \frac{\pi_2}{\pi_1} \tag{21a}$$

$$-0.5(x-\mu_1)'\Sigma^{-1}(x-\mu_1) + 0.5(x-\mu_2)'\Sigma^{-1}(x-\mu_2) > \ln\left(\frac{\pi_2}{\pi_1}\right) \tag{21b}$$

$$x'\Sigma^{-1}(\mu_2-\mu_1) < \ln\left(\frac{\pi_1}{\pi_2}\right) + 0.5\mu_2'\Sigma^{-1}\mu_2 - 0.5\mu_1'\Sigma^{-1}\mu_1 \tag{21c}$$

Wenn man die Regel (21c) etwas umschreibt, lässt sie sich leicht verallgemeinern:

$$(\Sigma^{-1}\mu_2)'x - 0.5\mu_2'\Sigma^{-1}\mu_2 + \ln(\pi_2) < (\Sigma^{-1}\mu_1)'x - 0.5\mu_1'\Sigma^{-1}\mu_1 + \ln(\pi_1). \tag{22}$$

7.1 Allgemeine Lineare Diskriminanzanalyse

Allgemein lässt sich nämlich zeigen, dass ganz allgemein die Klasse k nach der Bayes-Regel unter den Annahmen (L1), (L2) und (L3) immer dann gewählt wird, wenn die Funktion

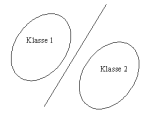
$$h_i(x) := (\Sigma^{-1}\mu_i)'x - 0.5\mu_i'\Sigma^{-1}\mu_i + \ln(\pi_i) \qquad (23)$$

maximal ist für Klasse k.

Durch diese Regel wird der p-dimensionale **Raum** insofern **partitioniert**, dass jedem p-Vektor genau eine Klasse zugeordnet wird, es sei denn, er liegt auf einer Grenze zwischen zwei Klassen.

7.1 Allgemeine Lineare Diskriminanzanalyse

- Ein idealisiertes Beispiel zeigt die folgende Abbildung. Dort sind beide Klassen vollständig durch eine Gerade voneinander zu trennen.



- In der Realität werden sich die Klassen aber im Allg. überlappen, so dass eine solch ideale Trennung unmöglich ist.
- Im Allgemeinen sind die Parameter der Normalverteilungen nicht bekannt. In diesem Fall werden stattdessen ihre **empirischen Momente** verwendet, d.h. Mittelwerte und empirische Kovarianzmatrix.

7.1 Allgemeine Lineare Diskriminanzanalyse

Die Grenzen zwischen jeweils zwei Klassen ergeben sich aus der Gleichsetzung der Funktionen $h_i(x)$ dieser Klassen:

$$\begin{aligned} & (\Sigma^{-1}\mu_i)'x - 0.5\mu_i'\Sigma^{-1}\mu_i + \ln(\pi_i) \\ = \; & (\Sigma^{-1}\mu_j)'x - 0.5\mu_j'\Sigma^{-1}\mu_j + \ln(\pi_j), \end{aligned}$$

d.h. $(\mu_j - \mu_i)'\Sigma^{-1}x = const$,
die Grenzen sind also **Hyperebenen** im $\mathbb{R}^p$.

Hyperebenen sind **Geraden** in zwei Dimensionen, **Ebenen** in drei Dimensionen usw. Im Fall von zwei Klassen in zwei Dimensionen wird zum Beispiel diejenige Gerade gesucht, die beide Klassen "am besten" voneinander trennt.

7.2 Fisher'sche LDA – Einführung

Kapitel

7 Diskriminanzanalyse
7.2 Fisher'sche LDA

Bisher wurden zur Bestimmung der Bayes-Regel Normalverteilungen angenommen. Diese Annahme hat sich aber als überflüssig erwiesen, um die obige Klassifikationsregel abzuleiten.

7.2 Fisher'sche LDA – Einführung

- Die Fisher'sche lineare Diskriminanzanalyse geht lediglich davon aus, dass für p Variable N Beobachtungen gegeben sind und jede Beobachtung sich mindestens einer von G Klassen zuordnen lässt.
- Mit Hilfe dieser N Beobachtungen wird dann eine Klassifikationsregel aufgestellt, mit deren Hilfe zukünftige Beobachtungen klassifiziert, d.h. genau einer der G Klassen zugeordnet werden können.
- Die geometrische Idee der **Fisher'schen Diskriminanzanalyse** ist es, den Beobachtungsraum durch eine Reihe von **Hyperebenen** zu partitionieren und den so entstandenen Teilen des Raumes jeweils genau eine Klasse zuzuordnen.
- Die prognostizierte Klasse einer neuen Beobachtung richtet sich dann danach, in welchen Teil des Raumes die neue Beobachtung fällt.

7.2.1 Fisher'sche LDA – 2 Klassen

Betrachten wir zunächst das **Zwei-Klassen Problem** und die Frage, wie wir die Hyperebene finden, die beide Klassen optimal trennt.

Im Fall von p Variablen (Dimensionen) sind dafür die Koeffizienten $a_1, \ldots, a_p$ zu bestimmen, so dass die Werte der Linearkombination

$$g(x) = a'x \tag{24}$$

die sog. **Diskriminanzkomponente**, für die Elemente der beiden Klassen so unterschiedlich wie möglich sind.

7.2.1 Fisher'sche LDA – 2 Klassen

- Das Problem dabei ist, dass man einzelnen Punkten zwar einfach **Abstände** zuordnen kann, nicht aber **Stichproben**, um die es sich bei den Klassen handelt.
- Deshalb sollen die Stichproben der einzelnen Klassen durch ihre **Mittelwerte** charakterisiert werden.
- Es wird gefordert, dass die (Projektionen der) Mittelwerte auf der Diskriminanzkomponente den größtmöglichen Abstand voneinander haben sollen.
- Dabei ist zu berücksichtigen, dass die x_j im Allg. auf verschiedenen Skalen gemessen werden. Deshalb wird das Abstandsmaß bzgl. der Varianzen und Kovarianzen standardisiert.
- Bei der linearen Diskriminanzanalyse wird (wiederum) unterstellt, dass alle Klassen **dieselbe Kovarianzstruktur** aufweisen (aber hier eben nicht unbedingt normalverteilt).

7.2.1 Fisher'sche LDA – 2 Klassen

Wenn S eine Schätzung dieser **Kovarianzmatrix** innerhalb der Klassen ist, dann ist $a'Sa$ die geschätzte Varianz innerhalb der Klassen in Richtung des Vektors a. Das führt zu dem zu maximierenden Abstand:

$$\frac{|a'(\bar{x}_2 - \bar{x}_1)|}{a'Sa} \tag{25}$$

wobei $\bar{x}_i$ der Mittelwert der Beobachtungen in Klasse i ist.

Ein äquivalentes Kriterium ist:

$$D(a) = \frac{(a'(\bar{x}_2 - \bar{x}_1))^2}{a'Sa}. \tag{26}$$

7.2.1 Fisher'sche LDA – 2 Klassen

- Man kann zeigen, dass $D(a)$ maximal wird in der Richtung $a_{opt} = S^{-1}(\bar{x}_2 - \bar{x}_1)$, denn a_{opt} ist Lösung von $\frac{\partial}{\partial a} D(a) = 0$.
- Damit sind die **Koeffizienten der Diskriminanzkomponente** bestimmt.
- Die **Klassifikationsregel** lautet dann:
  Eine neue Beobachtung x wird derjenigen Klasse zugerechnet, deren projizierter Mittelpunkt $a'_{opt}\bar{x}_i$ am nächsten bei der Projektion $a'_{opt}x$ liegt.
- **Klassifikationsregel**:

$$(a'_{opt}(x - \bar{x}_i))^2 = \min_i !$$

7.2.1 Fisher'sche LDA – 2 Klassen

Die **trennende Gerade** wird also senkrecht zu der Diskriminanzkomponente gewählt, wobei sie diese an der Stelle

$$\frac{g_{opt}(\bar{x}_2) + g_{opt}(\bar{x}_1)}{2}, g_{opt}(x) := a'_{opt}x \qquad (27)$$

durchstößt, d.h. in der Mitte zwischen den Projektionen der beiden Mittelwerte.

Tatsächlich entspricht die Fisher'sche Regel der Bayes-Regel bei Normalverteilungen (nur im Fall von 2 Klassen!), denn falls die a-priori Wahrscheinlichkeiten in Formel (21c) gleich sind, ergibt sich daraus die Regel:
"Wähle Klasse 1, wenn $a'_{opt}x < 0.5(a'_{opt}\mu_1 + a'_{opt}\mu_2)$ ist mit $a_{opt} = \Sigma^{-1}(\mu_2 - \mu_1)$."

7.2.2 Fisher'sche LDA: Mehrere Klassen

Bisher haben wir nur eine Trennrichtung gefunden. Tatsächlich ist bei zwei Klassen auch nur eine trennende Hyperebene sinnvoll, wie wir noch sehen werden. Anders sieht das Problem bei mehr als zwei Klassen aus.

Dazu wird häufig die folgende **Verallgemeinerung der Fisher'schen Methode** verwendet.

7.2.2 Fisher'sche LDA: Mehrere Klassen

Bei zwei Klassen haben wir zur Bestimmung der Diskriminanzkomponente den Quotienten $D(a)$ maximiert:

$$D(a) = \frac{(a'(\bar{x}_1 - \bar{x}_2))^2}{a'Sa}.$$

Der Zähler ist aber gleich $a'(\bar{x}_1 - \bar{x}_2)(\bar{x}_1 - \bar{x}_2)'a$
und dieser Ausdruck ist proportional zu

$$a'((\bar{x}_1 - \bar{x})(\bar{x}_1 - \bar{x})' + (\bar{x}_2 - \bar{x})(\bar{x}_2 - \bar{x})')a,$$

wobei $\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N}$ das Gesamtmittel aller Beobachtungen ist.

7.2.2 Fisher'sche LDA: Mehrere Klassen

Der mittlere Term in diesem Ausdruck erlaubt auf einfache Weise eine Verallgemeinerung auf mehr als zwei Klassen. Dazu wird die **Zwischen-den-Klassen-Kovarianzmatrix** B definiert:

$$B := \frac{1}{G} \sum_{i=1}^{G} (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})' \tag{28}$$

7.2.2 Fisher'sche LDA: Mehrere Klassen

Damit wird dann die maximale Trennung zwischen den Klassen durch die Maximierung des Ausdrucks

$$\frac{a'Ba}{a'Sa} \tag{29}$$

erreicht. Dabei ist S die gemittelte sog. **gepoolte Kovarianzmatrix** innerhalb der Klassen:

$$S := \frac{1}{N-G} \sum_{i=1}^{G} \sum_{j=1}^{N_i} (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_i)', \tag{30}$$

wobei jeweils N_i Beobachtungen x_{ij} aus der Klasse A_i angenommen werden, $N = N_1 + \ldots + N_G$.

7.2.2 Fisher'sche LDA: Mehrere Klassen

- Wir maximieren also die Trennung zwischen den Klassenmitteln, standardisiert mit dem Zusammenhang der Variablen innerhalb der Klassen.
- Die gesuchte Richtung a_{opt} ergibt sich als standardisierter **Eigenvektor** zum größten Eigenwert der Matrix $S^{-1}B$.
  Beweis: Mardia et al., S. 318, 479.
- Diese Richtung heißt erste **Diskriminanzkomponente**.
- Der Ansatz hat den Vorteil, dass sich weitere Diskriminanzkomponenten durch die Eigenvektoren der nächst kleineren Eigenwerte ergeben.
- Diese Eigenvektoren resultieren aus der Maximierung von (29) unter der Nebenbedingung, dass nur Richtungen senkrecht zu den bereits gefundenen gesucht werden.

7.2.2 Fisher'sche LDA: Mehrere Klassen

- Die relevante Anzahl Diskriminanzkomponenten ist dadurch bestimmt, dass mit weiteren Komponenten keine bessere Klassifikation möglich ist.
- Maximal lassen sich G-1 Diskriminanzkomponenten bestimmen.
  Beweis: $S^{-1}B$ hat maximal $\min(p, G-1)$ Eigenwerte ungleich Null; denn B hat maximal Rang $G-1$, weil $\sum_{i=1}^{G} N_i(\bar{x}_i - \bar{x}) = 0$, und S hat maximal Rang p.

7.2.2 Fisher'sche LDA: Mehrere Klassen

Analog zum 2-Klassen-Fall wählen wir als

Fisher'sche Klassifikationsregel im Mehrklassenfall:

Wähle die **prognostizierte Klasse** k so, dass

$$\sum_{l=1}^{r} \frac{(a_l'(x - \bar{x}_g))^2}{a_l' S a_l} = \min_g$$

bei r Diskriminanzkomponenten, wobei die $a_l, l = 1, \ldots, r$, die ersten r Eigenvektoren von $S^{-1}B$ sind.

7.2.2 Fisher'sche LDA: Mehrere Klassen

- Es wird also diejenige Klasse gewählt, deren Mittelwert nach Projektion auf r Diskriminanzkomponenten im Mittel die kleinste **Distanz** von dem zu klassifizierenden Objektvektor x hat, wobei jede Komponente normiert wird mit ihrer Varianz.
- Dabei sollte r so festgelegt werden, dass die Eigenwerte der nicht berücksichtigten Eigenvektoren klein sind relativ zu der Summe der berücksichtigten Eigenwerte.
- Eine wiederum idealisierte Darstellung der Trennung von drei Klassen in zwei Dimensionen findet sich in der folgenden Abbildung, die dem folgenden **Beispiel** nachempfunden ist.

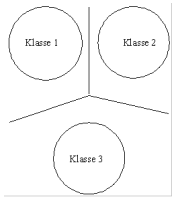
7.2.3 Fisher'sche LDA: Beispiel A

Die Klassenmitten liegen bei $\mu_1 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\mu_3 = \begin{pmatrix} 0 \\ -2 \end{pmatrix}$,

die gemeinsame Kovarianzmatrix hat die Form $\Sigma = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.25 \end{pmatrix}$

und die a-priori-Wahrscheinlichkeiten sind gleich:

$\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$.



7.2.3 Fisher'sche LDA: Beispiel A

Damit gilt für die **Trenngeraden der Bayes-Regel**:

$(\Sigma^{-1}\mu_i)'x - 0.5\mu_i'\Sigma^{-1}\mu_i + \ln(\pi_i) = (\Sigma^{-1}\mu_j)'x - 0.5\mu_j'\Sigma^{-1}\mu_j + \ln(\pi_j)$:

- $-2x_1 + 4x_2 - 3 = 2x_1 + 4x_2 - 3$,
  d.h. $x_1 = 0$ ist die Trenngerade der Klassen 1, 2
- $-2x_1 + 4x_2 - 3 = -8x_2 - 8$,
  d.h. $x_2 = \frac{1}{6}x_1 - \frac{5}{12}$ ist die Trenngerade der Klassen 1, 3
- $2x_1 + 4x_2 - 3 = -8x_2 - 8$,
  d.h. $x_2 = -\frac{1}{6}x_1 - \frac{5}{12}$ ist die Trenngerade der Klassen 2, 3.
- Die Geraden treffen sich im Punkt $\left(0, \frac{5}{12}\right)$.

7.2.3 Fisher'sche LDA: Beispiel A

Für die Diskriminanzkomponenten der **Fisher'schen LDA** gilt, wenn man die theoretischen Werte der Klassenmitten und der Kovarianzmatrix verwendet:
Gesucht sind die Eigenvektoren der Matrix (inkonsistent notiert):

$$\begin{aligned}\Sigma^{-1}B &= \begin{pmatrix}0.5 & 0\\ 0 & 0.25\end{pmatrix}^{-1}\frac{1}{G}\sum_{i=1}^{G}(\bar{x}_i-\bar{x})(\bar{x}_i-\bar{x})'\\ &= \begin{pmatrix}2 & 0\\ 0 & 4\end{pmatrix}\frac{1}{3}\left[\begin{pmatrix}-1\\ 1\end{pmatrix}(-1\ 1)+\begin{pmatrix}1\\ 1\end{pmatrix}(1\ 1)+\begin{pmatrix}0\\ -2\end{pmatrix}(0\ -2)\right]\\ &= \begin{pmatrix}2 & 0\\ 0 & 4\end{pmatrix}\frac{1}{3}\begin{pmatrix}2 & 0\\ 0 & 6\end{pmatrix}=\begin{pmatrix}\frac{4}{3} & 0\\ 0 & 8\end{pmatrix},\end{aligned}$$

denn für die Gesamtmitte gilt: $\bar{\mu}=\begin{pmatrix}0\\ 0\end{pmatrix}$.

7.2.3 Fisher'sche LDA: Beispiel A

- Eigenwerte von $\begin{pmatrix}\frac{4}{3} & 0\\ 0 & 8\end{pmatrix}$: 8 und $\frac{4}{3}$
- zugehörige Eigenvektoren: $a_1=\begin{pmatrix}0\\ 1\end{pmatrix}$ und $a_2=\begin{pmatrix}1\\ 0\end{pmatrix}$
- Normierung: $a_1'\Sigma a_1=0.25$ und $a_2'\Sigma a_2=0.5$.

Betrachten wir jetzt die Trennung zwischen Klassen 1 und 3 **nach Fisher**, dann gilt:

$$\sum_{l=1}^{2}\frac{(a_l'(x-\mu_1))^2}{a_l'\Sigma a_l} = \sum_{l=1}^{2}\frac{(a_l'(x-\mu_3))^2}{a_l'\Sigma a_l}$$

7.2.3 Fisher'sche LDA: Beispiel A

Einsetzen liefert:

$$\begin{aligned}&\frac{1}{0.25}\left((0\quad 1)\begin{pmatrix}x_1+1\\ x_2-1\end{pmatrix}\right)^2+\frac{1}{0.5}\left((1\quad 0)\begin{pmatrix}x_1+1\\ x_2-1\end{pmatrix}\right)^2\\ =\ &\frac{1}{0.25}\left((0\quad 1)\begin{pmatrix}x_1\\ x_2+2\end{pmatrix}\right)^2+\frac{1}{0.5}\left((1\quad 0)\begin{pmatrix}x_1\\ x_2+2\end{pmatrix}\right)^2\end{aligned}$$

Auflösen ergibt:

$$\begin{aligned}4(x_2-1)^2+2(x_1+1)^2 &= 4(x_2+2)^2+2x_1^2,\\ 4x_2^2-8x_2+4+2x_1^2+4x_1+2 &= 4x_2^2+16x_2+16+2x_1^2\\ -8x_2+4x_1+6 &= 16x_2+16,\\ x_2 &= \frac{1}{6}x_1-\frac{5}{12}\end{aligned}$$

Da für die anderen Trennungen Analoges gilt, stimmen die **Trenngeraden nach Bayes und Fisher** überein.

7.2.4 Fisher'sche LDA: Dimensionsreduktion

Eine ganz andere Art, die **Diskriminanzkomponenten** zu verwenden ist, sie als **neue Koordinaten** zu nutzen. Dabei werden die Beobachtungsvektoren zur Veranschaulichung der Klassifikation auf die relevanten Diskriminanzkomponenten projiziert.

7.2.4 Fisher'sche LDA: Dimensionsreduktion

- Im Allg. kommt man ohne großen Informationsverlust mit deutlich weniger als $G-1$ Diskriminanzkomponenten aus. Deshalb zählt die Diskriminanzanalyse zu den Methoden der **Dimensionsreduktion**.
- Die **Lineare Diskriminanzanalyse** wird also auch zur Reduktion der Dimension p verwendet mit dem Ziel einer einfacheren Darstellung der Klassifikationsregeln ohne großen Informationsverlust.
- Besonders anschaulich ist die graphische Darstellung der Klassifikationsregeln im Fall von 2 Komponenten. Die bisherigen Abbildungen von Partitionen können als grafische Darstellungen von **Projektionen auf 2 Diskriminanzkomponenten** aufgefasst werden.

7.2.4 Fisher'sche LDA: Dimensionsreduktion

Definition

Fisher's Lineare Diskriminanzanalyse (LDA)

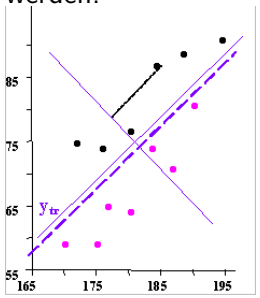
Die **LDA** findet für einen Datensatz ein orthogonales Koordinatensystem, in dessen 1. Koordinatenrichtung (**1. Diskriminanzkomponente** d_1) die Mittelwertvektoren der Klassen bestmöglich getrennt sind, in dessen 2. Koordinatenrichtung (d_2) bestmöglich für alle orthogonalen Richtungen, usw.
Jedes d_i entspricht einer Linearkombination der ursprünglichen p Merkmale:

$$d_i = c_{1i} \cdot x_1 + \ldots + c_{pi} \cdot x_p, \quad i = 1, \ldots, r.$$

Ziel ist oft eine Dimensionsreduktion mit $r < \min(p, G-1)$.

7.2.5 Fisher'sche LDA: Beispiel B

Beispiel: 2D-Diskriminanzanalyse mit 2 Klassen
13 Männer wurden (von einem "Supervisor") auf der Basis ihrer Größe und ihres Gewichts den beiden Klassen "**dick**" und "**dünn**" zugeordnet. Diese Zuordnung kann fehlerfrei durch eine Klassifikationsregel mit Hilfe der Diskriminanzanalyse reproduziert werden.



7.2.5 Fisher'sche LDA: Beispiel B

	Merkmal		
Träger	x=Größe	y=Gewicht	Klasse
1	170	60	dünn
2	172	76	dick
3	175	60	dünn
4	176	75	dick
5	177	66	dünn
6	180	65	dünn
7	180	78	dick
8	183	75	dünn
9	185	87	dick
10	187	72	dünn
11	188	90	dick
12	190	92	dünn
13	194	92	dick

7.2.5 Fisher'sche LDA: Beispiel B

Hier gilt für die 1. Diskriminanzkomponente:

$$a_{opt} = 11 \begin{bmatrix} S_{x_1}^2 + S_{x_2}^2 & S_{x_1y_1} + S_{x_2y_2} \\ S_{x_1y_1} + S_{x_2y_2} & S_{y_1}^2 + S_{y_2}^2 \end{bmatrix}^{-1} \begin{pmatrix} \bar{x}_2 - \bar{x}_1 \\ \bar{y}_2 - \bar{y}_1 \end{pmatrix},$$

wobei $N - G = 11$,

$S_{x_1}^2 := \sum_{j=1}^{N_1} (x_{1j} - \bar{x}_1)^2, \quad S_{x_1y_1} := \sum_{j=1}^{N_1} (x_{1j} - \bar{x}_1)(y_{1j} - \bar{y}_1),$

N_1 = Anzahl der Beobachtungen in Klasse 1 und Analoges für Klasse 2.

7.2.5 Fisher'sche LDA: Beispiel B

Es gilt aber: $\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$.

Also:

$$\begin{pmatrix} a_{opt1} \\ a_{opt2} \end{pmatrix} = \begin{bmatrix} S_{y_1}^2 + S_{y_2}^2 & -(S_{x_1y_1} + S_{x_2y_2}) \\ -(S_{x_1y_1} + S_{x_2y_2}) & S_{x_1}^2 + S_{x_2}^2 \end{bmatrix} \begin{pmatrix} \bar{x}_2 - \bar{x}_1 \\ \bar{y}_2 - \bar{y}_1 \end{pmatrix} \cdot \frac{11}{(S_{x_1}^2 + S_{x_2}^2)(S_{y_1}^2 + S_{y_2}^2) - (S_{x_1y_1} + S_{x_2y_2})^2}$$

Die dazugehörige **Trennachse**, senkrecht zu der ersten **Diskriminanzachse**, wird bestimmt durch die Tatsache, dass sämtliche Orthogonalprojektionen auf die erste Diskriminanzachse von Punkten auf der Trennachse mit dem Mittelpunkt zwischen den Projektionen der Klassenmittelwertsvektoren zusammenfallen:

7.2.5 Fisher'sche LDA: Beispiel B

$$\begin{aligned} & a_{opt1}x + a_{opt2}y_{tr} &&= 0.5(a_{opt1}\bar{x}_1 + a_{opt2}\bar{y}_1 + a_{opt1}\bar{x}_2 + a_{opt2}\bar{y}_2) \\ \Leftrightarrow \quad & a_{opt1}x + a_{opt2}y_{tr} &&=: c_{mittel} \\ & y_{tr} &&= \frac{c_{mittel}}{a_{opt2}} - \frac{a_{opt1}}{a_{opt2}}x \end{aligned}$$

- Daraus ergibt sich eine Trenngerade mit einer Steigung von 0.948.
- Tatsächlich gibt es eine ganze Schar von Trenngeraden mit dieser Steigung, die die beiden Klassen perfekt trennen.
- Ein Beispiel ist die gestrichelte Gerade in der einleitenden Abbildung: $y = 0.948x - 100$:
  das Gewicht sollte "< (0.948 Größe in cm - 100) in kg" sein, damit ein Mann als "dünn" bezeichnet wird.

Übung: Welche Klasse ist Klasse 1, "dick" oder "dünn"?

7.2.6 LDA: Anwendungsbeispiel C

Anwendungsbeispiel
Zur Illustration der Vorgehensweise bei einer Klassifikationsaufgabe wird im Folgenden die Phasen westdeutscher Konjunkturzyklen zwischen 1967 und 1993 betrachtet, die mit Hilfe von vier ökonomischen Variablen unterschieden werden sollen.

Y = Phasen von Konjunkturzyklen (PHASEN)

1. Aufschwungphase
2. Obere Wendepunktphase
3. Abschwungphase
4. Untere Wendepunktphase

7.2.6 LDA: Anwendungsbeispiel C

X_1 = Abhängig Erwerbstätige (EWAJW)
X_2 = Lohnstückkosten (LSTKJW)
X_3 = Preisindex des Bruttosozialprodukts (PBSPJW)
X_4 = Kurzfristiger Zinssatz (Dreimonatsgeld) (ZINSK)

- Die Ausgangsdaten sind für alle Quartale von 1967 bis 1993 in die vier Phasen unterteilt. Die Variablen X_1 bis X_3 sind (um Trend- und saisonale Einflüsse auszuschalten) in Wachstumsraten gegenüber dem entsprechenden Vorjahresquartal transformiert.

7.2.6 LDA: Anwendungsbeispiel C

- Mit Hilfe der Diskriminanzanalyse soll geprüft werden, welchen Beitrag die X-Variablen zur Unterscheidung der einzelnen Phasen der Konjunkturzyklen leisten. Dabei werden die a-priori-Wahrscheinlichkeiten der Klassen (= Phasen) aus den Klassengrößen berechnet.
- In einer folgenden Tabelle sind die Klassencharakteristika bei der Diskriminanzanalyse dargestellt.
- Man erkennt, dass sich die Mittelwerte der Variablen in den einzelnen Phasen der Konjunkturzyklen deutlich voneinander unterscheiden — eine wesentliche Voraussetzung für eine sinnvolle Anwendung der Diskriminanzanalyse.

7.2.6 LDA: Anwendungsbeispiel C

```
R> library("klaR") # lädt automatisch auch Paket "MASS" (für lda())
R> data("B3")
R> B3 <- B3[46:153,]  # 1967 - 1993
R> Xvars <- c("EWAJW", "LSTKJW", "PBSPJW", "ZINSK")
R> by(B3[,Xvars], B3$PHASEN, function(x)
R+    cbind(Mittelw. = mean(x), Standardab. = sd(x)))

B3$PHASEN: 1                    B3$PHASEN: 3
       Mittelw. Standardab.            Mittelw. Standardab.
EWAJW     1.056       0.889     EWAJW     0.718       1.495
LSTKJW    2.337       1.721     LSTKJW    7.167       3.592
PBSPJW    2.882       1.058     PBSPJW    5.177       1.578
ZINSK     4.844       1.260     ZINSK     9.962       2.015
------------------------------  ------------------------------
B3$PHASEN: 2                    B3$PHASEN: 4
       Mittelw. Standardab.            Mittelw. Standardab.
EWAJW     2.569       0.639     EWAJW    -1.450       1.612
LSTKJW    3.853       2.276     LSTKJW    5.104       4.454
PBSPJW    4.239       1.011     PBSPJW    4.916       2.335
ZINSK     8.264       1.139     ZINSK     6.275       1.972
```


7.2.6 LDA: Anwendungsbeispiel C

- In einer weiteren Tabelle finden sich die a-priori-Wahrscheinlichkeiten der Phasen:

```
R> cbind(Anzahl = table(B3$PHASEN),
R+       `a-priori Wk.` = table(B3$PHASEN) / nrow(B3))

  Anzahl a-priori Wk.
1     48        0.444
2     14        0.130
3     25        0.231
4     21        0.194
```

Tab. 4 : A-priori-Wahrscheinlichkeiten der Klassen

7.2.6 LDA: Anwendungsbeispiel C

- Danach besteht also eine Wahrscheinlichkeit von 44% dafür, dass ein Fall der ersten Gruppe (Aufschwung) angehört, während für die restlichen Gruppen die Wahrscheinlichkeiten deutlich kleiner sind.
- Die Größe der Absolutwerte der normierten Koeffizienten der drei Diskriminanzkomponenten (, die aus der Standardisierung aller eingesetzten Variablen auf die Standardabweichung 1 resultieren) in der nächsten Tabelle deutet an, dass
  - Abhängig Erwerbstätige (EWAJW) und
  - Kurzfristzinsen (ZINSK)

  die wichtigeren Variablen in diesem Beispiel sind.

7.2.6 LDA: Anwendungsbeispiel C

```
R> ldaobj <- lda(PHASEN ~ EWAJW + LSTKJW + PBSPJW + ZINSK, data = B3)
R> ldaobj$scaling

          LD1    LD2    LD3
EWAJW  -0.836 -0.545  0.135
LSTKJW  0.136 -0.045 -0.649
PBSPJW  0.241  0.338  1.299
ZINSK   0.429 -0.441  0.008
```

Tab. 5 : Normierte Koeffizienten der Diskriminanzkomponenten

- Die aus den 3 Diskriminanzkomponenten resultierende Klassifikationsregel führt für den Lerndatensatz 1967-1993 zu dem in der nun folgenden Tabelle enthaltenen Ergebnis:

7.2.6 LDA: Anwendungsbeispiel C

```
R> cbind(errormatrix(B3$PHASEN, predict(ldaobj)$class),
+>       errormatrix(B3$PHASEN, predict(ldaobj)$class, relative = TRUE))
       predicted           |        predicted
true     1  2  3  4 -SUM-   | true        1     2     3     4 -SUM-
  1     45  3  0  0     3   |   1     0.938 0.062 0.000 0.000 0.062
  2      3 10  1  0     4   |   2     0.214 0.714 0.071 0.000 0.286
  3      0  2 20  3     5   |   3     0.000 0.080 0.800 0.120 0.200
  4      1  0  2 18     3   |   4     0.048 0.000 0.095 0.857 0.143
  -SUM-  4  5  3  3    15   |   -SUM- 0.267 0.333 0.200 0.200 0.139

R> CV <- lda(PHASEN ~ EWAJW + LSTKJW + PBSPJW + ZINSK,
+>           data = B3, CV = TRUE)$class
R> cbind(errormatrix(B3$PHASEN, CV),
+>       errormatrix(B3$PHASEN, CV, relative = TRUE))
       predicted           |        predicted
true     1  2  3  4 -SUM-   | true        1     2     3     4 -SUM-
  1     44  3  0  1     4   |   1     0.917 0.063 0.000 0.021 0.083
  2      3 10  1  0     4   |   2     0.214 0.714 0.071 0.000 0.286
  3      0  2 20  3     5   |   3     0.000 0.080 0.800 0.120 0.200
  4      2  0  4 15     6   |   4     0.095 0.000 0.191 0.714 0.286
  -SUM-  5  5  5  4    19   |   -SUM- 0.263 0.263 0.263 0.211 0.176
```


7.2.6 LDA: Anwendungsbeispiel C

- Insgesamt wurden 86.1%, d.h. 93 von 108 Quartalen korrekt zugeordnet, mit erheblichen Unterschieden in den einzelnen Phasen der insgesamt vier Konjunkturzyklen.
- Da die offensichtliche Fehlerrate den wahren Fehler im Allg. unterschätzt, enthält der untere Teil der Tabelle den (kreuzvalidierten) **leave-one-out-Fehler**, der insgesamt etwa 4% höher liegt.
- Eine Alternativmethode ist hier, wenn jeweils einer der vier Konjunkturzyklen bei der Schätzung der Diskriminanzkomponenten unberücksichtigt bleibt (Trainingsdatensatz) und die Phasen dieses Zyklus (Testdatensatz) dann prognostiziert werden:
  Man muss davon ausgehen, dass man Daten aus bisher gar nicht beobachteten Konjunkturzyklen vorhersagen muss!
- Software: In R gibt es z.B. die Funktion `lda()` im Paket *MASS*.

7.2.7 Einschub: Leave-one-out Methode

Fehlerratenschätzung

Leave-one-out Methode

- Eine Klassifikationsmethode wird bei einer Größe N des Lerndatensatzes jeweils auf $(N-1)$ Beobachtungen angewendet und für die N-te Beobachtung getestet.
- Dieses Vorgehen wird N mal, d.h. für jede Beobachtung, wiederholt, wobei jedes Mal eine neue Klassifikationsregel bestimmt wird.
- Als Fehlerrate wird die Anzahl Fehler bei den individuellen Testfällen geteilt durch N verwendet.

7.3 Diskriminanzanalyse: Variablenselektion

Kapitel

7 Diskriminanzanalyse
7.3 Variablenselektion in der LDA

(aus: Mardia et al., 1979, 322-324)

- Im letzten Abschnitt haben wir ein Beispiel für die Beurteilung der Wichtigkeit von Variablen für die Trennung der Klassen gesehen.
- Dort wurden Gewichte der standardisierten Diskriminanzkomponenten zur Beurteilung herangezogen: je größer, desto besser.

7.3 Diskriminanzanalyse: Variablenselektion

- Streng genommen gilt diese Methode aber nur, wenn die **Orginalvariablen unkorreliert** sind. Das ist meistens nicht der Fall. Dann muss zur Beurteilung der Relevanz der Originalvariablen die Kovarianzmatrix S der Variablen innerhalb der Klassen mit herangezogen werden.
- Aus der Vielfalt der Methoden zur Bestimmung relevanter Variablen stellen wir nur eine wichtige vor, bei der eine F-Statistik unter der **Null-Hypothese** bestimmt wird, dass eine oder mehrere Variablen keinen Einfluss auf den Mahalanobis-Distanz zwischen **zwei Klassen** hat.
- Die geschätzte Diskriminanzkomponente hat die Form

$$a_{opt} = S^{-1}(\bar{x}_2 - \bar{x}_1).$$

7.3 Diskriminanzanalyse: Variablenselektion

- Die **Mahalanobis-Distanz** zwischen den Populationen der beiden Klassen wird geschätzt durch:

$$D_p^2 = (\bar{x}_2 - \bar{x}_1)' S^{-1} (\bar{x}_2 - \bar{x}_1).$$

- Wenn $(p-k)$ Variablen keinen Einfluss auf die Trennung der Klassen haben, dann gilt (evtl. nach Umordnung der Komponenten): $\Sigma^{-1}(\mu_2 - \mu_1) = (\alpha_1, \alpha_2)'$ mit $\alpha_2 = 0$ aus $\mathbb{R}^{p-k}$.

7.3 Diskriminanzanalyse: Variablenselektion

- Unter der **Nullhypothese** $H_0 : \alpha_2 = 0$ gilt dann:
  - die $(p - k)$ Variablen haben keinen Einfluss auf die Mahanalobis-Distanz und
  - die **Teststatistik**

$$T_M = \frac{N - G - p + 1}{p - k} \frac{c^2(D_p^2 - D_k^2)}{N - G + c^2 D_k^2}$$

  mit $c^2 = \frac{N_1 N_2}{N}$ ist $F_{p-k,N-G-p+1}$-verteilt, wobei D_k^2 die Mahalanobis-Distanz ist auf der Basis der ersten k Variablen.
- Die Nullhypothese wird abgelehnt, falls gilt:

$$T_M > F_{p-k,N-G-p+1}(1 - \alpha).$$

7.3 Diskriminanzanalyse: Variablenselektion

- Dieser Test kann schrittweise verwendet werden, indem die Gruppe der relevanten Variablen (, die zu α_1 gehört) sukzessive erweitert wird mit jeweils derjenigen einzelnen Variablen, die die Trennung der Klassen möglichst stark verbessert, die etwa die Mahalonobis-Distanz D_k am stärksten erhöht.
- Dabei wird die Gruppe der relevanten Variablen solange vergrößert, bis für die andere Gruppe kein signifikanter Einfluss mehr nachgewiesen werden kann (**gierige Vorwärtsselektion**).

7.3 Diskriminanzanalyse: Variablenselektion

- Diese **schrittweise Diskriminanzanalyse** kann so erweitert werden, dass, falls eine irrelevante Gruppe gefunden wurde, in der Gruppe der relevanten Variablen testweise versucht wird, die unwichtigste Variable zu eliminieren, indem diejenige Variable identifiziert wird, die D_k am wenigsten verkleinert. Wenn danach die andere Gruppe immer noch irrelevant ist, wird die Elimination realisiert und erneut eine Vorwärtsselektion versucht (**Vorwärts-Rückwärts-Selektion**), die u. U. eine andere Variable als die eliminierte identifiziert.
- Es gibt auch die reine **Rückwärtsselektion**.

7.4 Quadratische Diskriminanzanalyse

Kapitel

7 Diskriminanzanalyse
7.4 Quadratische Diskriminanzanalyse

Wird die Annahme (L1) durch die folgende Annahme (Q1) ersetzt, spricht man von **quadratischer Diskriminanzanalyse**:

(Q1) Für die Verteilungen innerhalb der Klassen werden Normalverteilungen mit unterschiedlichen Erwartungswerten und (unterschiedlichen!!!) Kovarianzmatrizen Σ_i angenommen.

7.4 Quadratische Diskriminanzanalyse

- Da bei einer quadratischen Diskriminanzanalyse wesentlich mehr Parameter bestimmt werden müssen als bei einer linearen, kann man aus dem allgemeineren Modell in der Regel nur bei entsprechend großen Stichproben Kapital schlagen.
- Hier lässt sich (analog zu (23)) zeigen, dass Klasse k nach der Bayes-Regel unter den Annahmen (Q1), (L2) und (L3) immer dann gewählt wird, wenn die Funktion

$$h_i^Q(x) := -0.5(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) + \ln(\pi_i) - 0.5\ln(\det(\Sigma_i)) \tag{31}$$

maximal ist für Klasse k.
- Zur expliziten Berechnung werden die entsprechenden empirischen Momente für μ_i und Σ_i und die relative Häufigkeit für π_i eingesetzt (sog. **plug-in Schätzer**).

7.4 Quadratische Diskriminanzanalyse

- Als **Klassenwahrscheinlichkeit** ergibt sich entsprechend:

$$\begin{aligned} P(A_i|x) &= \exp[-0.5(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i) \\ &\quad + \ln(\pi_i) - 0.5\ln(\det(\Sigma_i))] \end{aligned}$$

für gegebenes x, abgesehen vom normalisierenden Faktor.
- Bei der quadratischen Diskriminanzanalyse sind die **Trennflächen zwischen zwei Klassen** Linien gleicher Höhe von Konfidenzellipsoiden unterschiedlicher Größe und Ausrichtung.
- Das führt im Gegensatz zur Linearen Diskriminanzanalyse (mit linearen Trennungen) zu nichtlinearen Trennungen, wie das folgende Beispiel zeigt.
- Software: In R gibt es z.B. die Funktion `qda()` im Paket *MASS*.

7.4 Beispiel: Quadratische vs. Lineare DA

Beispiel (Westdeutscher Konjunkturzyklus):

Wir betrachten nur 2 ökonomische Erklärungsgrößen (Bezeichnungen geändert):

L	Wachstumsrate der Anzahl Beschäftigten (EWAJW)
LC	Wachstumsrate der Lohnstückkosten (LSTKJW)
LTP	Unterer Wendepunkt (der Konjunktur)
Up	Aufschwung
UTP	Oberer Wendepunkt
Down	Abschwung

7.4 Beispiel: Quadratische vs. Lineare DA

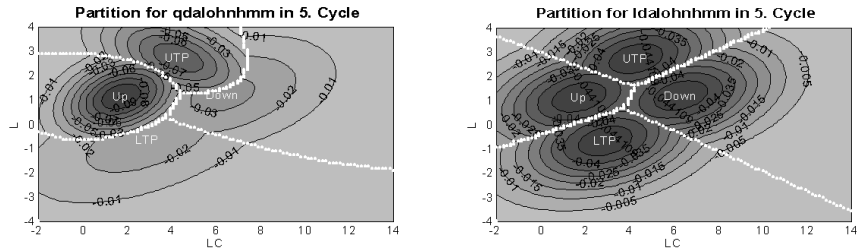


Abb. 1 : Partition bei Quadratischer (links) und linearer (rechts) Diskriminanzanalyse

7.5 Regularisierte Diskriminanzanalyse

Kapitel

7 Diskriminanzanalyse
7.5 Regularisierte Diskriminanzanalyse

- Das Hauptproblem der Quadratischen Diskriminanzanalyse besteht in der großen Anzahl zu schätzender Parameter.
- Das führt insbesondere zu Eigenwerten der Kovarianzmatrix, die Null oder nahezu Null sind. Damit ist diese Matrix nicht mehr invertierbar.
- Als Ausweg werden in der Literatur viele sog. **Regularisierungs- oder Glättungsmethoden** angeboten.

7.5 Regularisierte Diskriminanzanalyse

- Friedman (1989) beschreibt eine 2-Parameter-Familie von solchen Regularisierungsmethoden. Er "glättet" die Schätzung S_i der Klassen-Kovarianzmatrix Σ_i sowohl
  - "in Richtung" der (Schätzung der) gepoolten Kovarianzmatrix S, also auch
  - "in Richtung" der Einheitsmatrix I,

  d.h. er schlägt vor, folgenden geglätteten **Schätzer der Kovarianzmatrix** zu verwenden.

7.5 Regularisierte Diskriminanzanalyse

- Zunächst seien

$$\begin{aligned}
\hat{\mu}_i &= \bar{X}_i = \frac{1}{W_i} \sum_{c(j)=i} w_j X_j \\
S_i &= \sum_{c(j)=i} w_j (X_j - \bar{X}_i)(X_j - \bar{X}_i)' \\
S_i(\delta) &= (1-\delta) S_i + \delta S; \quad S = \sum_i S_i \\
W_i(\delta) &= (1-\delta) W_i + \delta W; \quad W_i = \sum_{c(j)=i} w_j, \quad W = \sum_i W_i \\
\hat{\Sigma}_i(\delta) &= S_i(\delta)/W_i(\delta),
\end{aligned}$$

  wobei $c(j)$ die Klasse der j-ten Beobachtung ist und w_j Gewichte.

7.5 Regularisierte Diskriminanzanalyse

- Die geschätzte Kovarianz der Klasse i ergibt sich dann durch

$$\hat{\Sigma}_i(\delta, \lambda) := (1-\lambda)\hat{\Sigma}_i(\delta) + \frac{\lambda \cdot tr[\hat{\Sigma}_i(\delta)]}{p} I,$$

  wobei δ und λ möglichst optimal gewählte Konstanten zwischen Null und Eins sind:
  - Je größer δ ist, desto näher ist die Methode an der Linearen Diskriminanzanalyse.
  - Je größer λ ist, desto sicherer ist die Invertierbarkeit der Schätzung der Kovarianzmatrix.
- Für die weitere Diskussion dieses Problems s. z.B. die Diplomarbeiten von Pouwels (2001) und Röver (2003).
- Software: Im R Paket *klaR* gibt es z.B. die Funktion `rda()` (Regularized Discriminant Analysis), eine Implementation der Friedman-Klasse.

7.6 Klassifikation bei Existenz von Subklassen

Kapitel

7 Diskriminanzanalyse
7.6 Mixture Discriminant Analysis

(Hastie, T., Tibshirani, R. und Friedman, J., 2001, S. 399-405)

Eine andere Erweiterung zur Linearen Dikriminanzanalyse (LDA) stellt die **Mixture Discriminant Analysis (MDA)** dar.

Diese ist besonders geeignet für den Fall, dass mehrere Unterklassen der einzelnen Klassen bestehen, die unterschiedliche Charakteristika in den Daten besitzen.

7.6 Prinzip der MDA

An dieser Stelle soll das **Prinzip der MDA** vorgestellt werden (ohne weitere Details, siehe Literatur).

Die Normalverteilungsannahme der Klassen (L1) der (Linearen) Diskriminanzanalyse lässt sich erweitern, indem sich die Verteilung einer Beobachtung X, gegeben Klasse k, als **Mischverteilung** aus mehreren Normalverteilungen beschreiben lässt:

7.6 Prinzip der MDA

$$P(X|G=k) = \sum_{r=1}^{R_k} \pi_{kr} \Phi(X, \mu_{kr}, \Sigma), \text{ mit} \tag{32}$$

- $r = 1, \ldots, R_k$ die Unterklassen der Klasse k,
  - Die Anzahl der Unterklassen muss für die verschiedenen Klassen k nicht gleich groß sein, wird aber als bekannt vorausgesetzt.
- π_{kr} die Anteile der verschiedenen Subklassen an der Mischverteilung (summieren sich zu 1 auf),
- μ_{kr} der Erwartungswert der Subklasse r von Klasse k,
- Σ die Kovarianzen, die wie in der LDA der Einfachheit halber als identisch für alle Subklassen aller Klassen angenommen werden.

7.6 Prinzip der MDA

Wenn die derartigen Verteilungen aller Klassen bekannt sind, lassen sich die **a posteriori Zugehörigkeiten** einer Beobachtung $X = x$ zur Klasse k bestimmen durch:

$$P(G=k|X=x) = \frac{\pi_k \sum_{r=1}^{R_k} \pi_{kr} \Phi(X, \mu_{kr}, \Sigma)}{\sum_{L=1}^{K} \pi_L \sum_{r=1}^{R_L} \pi_{Lr} \Phi(X, \mu_{Lr}, \Sigma)},$$

wobei π_k die a priori Wahrscheinlichkeit für ein Vorliegen der Klasse k bedeutet.

7.6 Prinzip der MDA

Es tritt das Problem auf, dass zwar die Klassen k der Objekte, nicht jedoch die Zugehörigkeiten zu den Subklassen bekannt sind. Diese (und die daraus resultierenden Modelle der Verteilungen der Subklassen) lassen sich mit Hilfe des **EM-Algorithmus** bestimmen.

7.6.1 Der EM-Algorithmus

EM-Algorithmus

Der EM-Algorithmus

Der EM-Algorithmus besteht aus der wiederholten Abfolge von **E-step (expectation)** und **M-step (maximization)** bis hin zur Konvergenz.

7.6.1 Der EM-Algorithmus

E-step:

- Für ein gegebenes Modell von Subklassen-Verteilungsparametern (μ_{kr}, π_{kr}) werden **Zugehörigkeitsgewichte** $W(c_{kr})$ der einzelnen Objekte aus Klasse k zu den einzelnen Subklassen c_{kr} berechnet.
- Das Vorgehen ist identisch mit der Schätzung der a posteriori Wahrscheinlichkeit der Klassen in der Linearen Diskriminanzanalyse, wenn die Parameter der Klassen bekannt sind, es gilt:

$$W(c_{kr}|x_i, k) = \frac{\pi_{kr}\Phi(x_i, \mu_{kr}, \Sigma)}{\sum_{L=1}^{R_k} \pi_{kL}\Phi(x_i, \mu_{kL}, \Sigma)} \quad (33)$$

$W(c_{kr}|x_i, k)$ beschreibt dabei das Gewicht der Beobachtung x_i (die aus Klasse k stammen muss !) welches die Zugehörigkeit zur Subklasse c_{kr} angibt.

7.6.1 Der EM-Algorithmus

- Mit den so berechneten Subklassen-Zugehörigkeiten $W(c_{kr}|x_i, k)$ lassen sich im **M-step** die Modellparameter der Mischverteilung mit Maximum-Likelihood Schätzern neu bestimmen, wobei die Beobachtungen x_i in Klasse k mit $W(c_{kr}|x_i, k)$ gewichtet werden.
- Aus der neuen Schätzung der Verteilungsparameter können die Subklassen-Zugehörigkeiten neu bestimmt werden, usw. ... bis zur Konvergenz des Algorithmus (siehe Hastie et al., 2001, pp. 236-243).

7.6.2 Beispiel: Der Waveform-Datensatz

Das Verfahren soll anhand des Waveform-Datensatzes (Hastie et al., 2001, S. 402-404) vorgeführt werden: dieser führt zu einem Klassifikationsproblem mit simulierten Daten in 3 Klassen und 21 Variablen.

Gegeben seien 3 Wellenformen über 21 **Variablen**:

$$h_1(j) = \max(6 - |j - 11|, 0), j = 1, 2, .., 21,$$
$$h_2(j) = h_1(j - 4)$$
$$h_3(j) = h_1(j + 4)$$

Die Wellen h_2 und h_3 sind dabei die nach links bzw. rechts verschobene Welle h_1.

7.6.2 Beispiel: Der Waveform-Datensatz

- Es ergibt sich folgende Gleichung für die Verteilung der Objekte der verschiedenen Klassen:

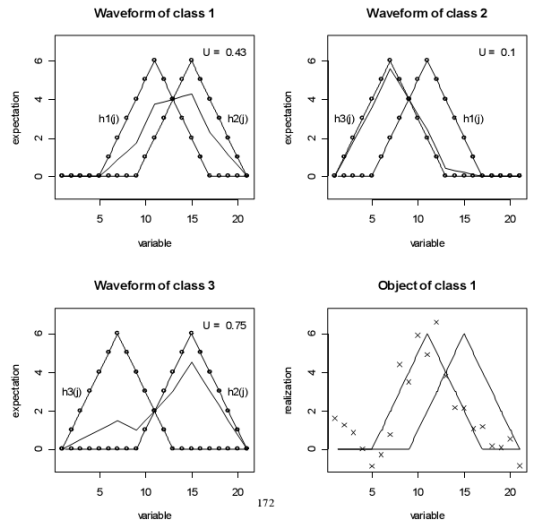
$$\text{Klasse 1: } X_j = U \cdot h_1(j) + (1 - U)h_2(j) + \epsilon_j$$
$$\text{Klasse 2: } X_j = U \cdot h_1(j) + (1 - U)h_3(j) + \epsilon_j$$
$$\text{Klasse 3: } X_j = U \cdot h_2(j) + (1 - U)h_3(j) + \epsilon_j$$

  - Dabei ist ϵ_j der zwischen den Variablen j unabhängig standard normalverteilte Zufallsfehler und $U \sim Unif(0, 1)$ ist die selbe gleichverteilte Zufallszahl über alle Variablen j für eine Beobachtung.
- Die folgende Abbildung zeigt die beiden beteiligten Wellenformen sowie exemplarisch den Erwartungswert eines Objektes in Abhängigkeit von der Realisierung der gleichverteilten Zufallszahl U über alle 21 Variablen. Unten rechts ist eine mögliche Realisierung eines Objekts aus Klasse 1 dargestellt.

7.6.2 Beispiel: Der Waveform-Datensatz

- Durch diese Wellenformen werden die Erwartungswerte der Klassen in den 21 Variablen beschrieben.
- Für jede der 3 Klassen werden dabei zwei andere Wellenformen zur Berechnung des Erwartungswertes verwendet.
- Für die Realisierung eines Objekts einer Klasse ergeben sich dessen Erwartungswerte in den 21 Variablen als gleichverteilte Zufallszahl zwischen den beiden Wellenformen.
- Um ihren Erwartungswert herum sind die Objekte unabhängig normalverteilt mit einer Varianz von 1.

7.6.2 Beispiel: Der Waveform-Datensatz



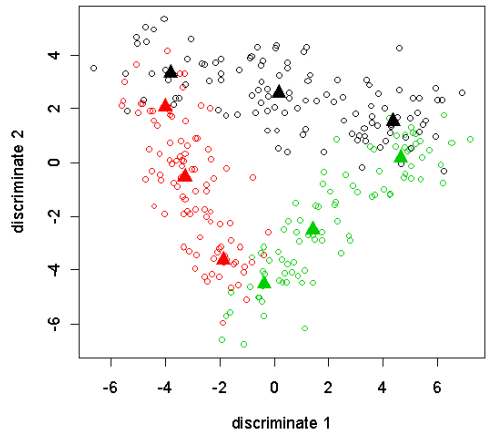
7.6.2 Beispiel: Der Waveform-Datensatz

Dieser Datensatz wurde von Breiman et al. (1984) eingeführt, gilt als schwieriges Klassifikationsproblem und besitzt eine Bayes-Fehlklassifikationsrate von 14.9%.

Simuliert wurden im Trainingsdatensatz 300 Beobachtungen (mit gleicher a priori Wahrscheinlichkeit aller 3 Klassen) und im Testdatensatz 500 Beobachtungen.

Vergleich von MDA (mit 3 Subklassen je Klasse) mit LDA und QDA ergibt eine Verbesserung des Klassifikationsergebnisses, obwohl QDA anhand der offensichtlichen Fehlerrate auf den Trainingsdaten sehr gute Ergebnisse vermuten ließe:

7.6.2 Beispiel: Der Waveform-Datensatz



Projektion auf 2 Diskriminanzachsen, Subklassenmittel als fette Dreiecke

7.6.2 Beispiel: Der Waveform-Datensatz

	Trainingsfehler	Testfehler
LDA	0.121	0.191
QDA	**0.039**	0.205
MDA (je 3 Subklassen)	0.087	**0.169**

Tab. 6 : Vergleich von Diskrimimanzanalyseverfahren auf dem Waveform Datensatz.

Die folgende Grafik zeigt insbesondere, dass die drei Klassen auf den Seiten eines Dreiecks liegen. Das war zu erwarten, weil die drei definierenden Funktionen h_i drei Punkte im 21-dimensionalen Raum bilden und die Klassen zufällige Konvexkombinationen von jeweils zwei dieser Funktionen sind.

7.7 Literatur: Diskriminanzanalyse

- Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984): Classification and Regression Trees, Wadsworth.
- Friedman, J. (1989): Regularized Discriminant Analysis; Journal of the American Statistical Association 84, 165-175.
- Hastie, T. Tibshirani, R., Friedman, J (2001). The Elements of Statistical learning - Data Mining, Inference and Prediction, Springer, NY.
- Heilemann, U., Weihs, C. (2004): Diskriminanzanalyse; in: Taschenbuch der Statistik, Voß, W. (ed.), 2. Auflage, Fachbuchverlag Leipzig, 583-608.
- Mardia, K.V., Kent, J.T., Bibby, J.M. (1979): Multiviarate Analysis; Academic Press, London
- Pouwels, B. (2001): Diskriminanzanalyse bei fast-singulären Kovarianzmatrizen; Diplomarbeit am Fachbereich Statistik, Universität Dortmund.
- Röver, C. (2003): Musikinstrumentenerkennung mit Hilfe der Hough-Transformation; Diplomarbeit am Fachbereich Statistik, Universität Dortmund.