.

Klassifikations- und Diskriminanzanalyse[1]

U. Mortensen
.
FB Psychologie und Sportwissenschaften, Institut III
Westfälische Wilhelms-Universität Münster

Letzte Korrektur: 13. 03. 2022/26. 09. 2025

[1] KlassifDiskriminanzAllgemein.tex

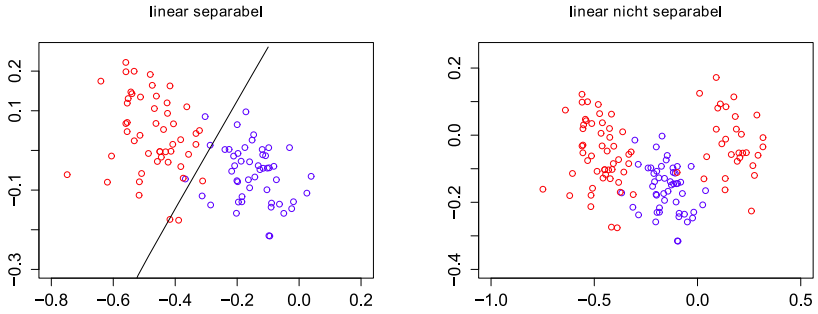
Inhaltsverzeichnis

1 Fishers Lineare Diskriminanzanalyse

Die Klassifikation von Personen, Objekten oder allgemein von Mustern anhand der Ausprägung einer Reihe von Variablen ist eine Aufgabe, die intuitiv auf der Basis von "Erfahrung" oft nur suboptimal gelöst wird: man denke an medizinische oder psychologische Diagnosen, die Zuordnung von Tonscherben zu einer von mehreren möglichen Kulturen, etc. Die Anwendung statistischer Verfahren kann hier hilfreich sein.

Es gebe g Klassen oder Kategorien $\Omega_1, \ldots, \Omega_g$, $g \geq 2$ und p Merkmale, von deren Ausprägung x_j, $j = 1, \ldots, p$ die Klassifikation abhängt. Es gibt zwei mögliche

Abbildung 1: Linear trennbare und linear nicht trennbare Konfigurationen



Ansätze, diese Frage zu diskutieren:

1. Man eine Skala Y oder mehrere Skalen $Y_1, \ldots, Y_r$ bestimmen, in Bezug auf die sich die Klassen maximal unterscheiden, wobei

   $$Y = u_1 x_1 + u_2 x_2 + \cdots + u_p x_p$$

   gelten soll. Die "Gewichte" $u_1, u_2, \ldots, u_p$ werden aus den Daten geschätzt. Diese Methode wurde von Fisher (1936) eingeführt.

2. Man nimmt eine bestimmte bedingte Wahrscheinlichkeitsdichte

   $$f(x_1, \ldots, x_p | \Omega_k)$$

   für die beobachteten Merkmale an, $k = 1, 2, \ldots, g$. Dies ist die Dichte der $x_1, \ldots, x_p$ unter der Bedingung, dass das Objekt oder die Person aus Ω_k ist. Dann lassen sich Entscheidungsregeln aufstellen, durch deren Anwendung sich die Wahrscheinlichkeit einer Fehlentscheidung, oder die mit einer Fehlentscheidung verbundenen Kosten minimisieren lassen.

Es soll zunächst der erste Ansatz beschrieben werden, allgemeinere Ansätze werden im Anschluß daran vorgestellt.

1.1 Fishers Ansatz

Es werden p Variablen $X_1, \ldots, X_p$ bei insgesamt m "Fällen" (Personen, Objekten, ...) gemessen, wobei jeder Fall genau einer von K Gruppen oder Klassen angehört. Gegeben sei eine Stichprobe von insgesamt m Fällen, mit n_1 Fällen der Kategorie C_1, n_2 Fällen in der Kategorie C_2, etc., bis n_k Fälle in der Kategorie C_K, und $m = \sum_k n_k$. Die Messungen für die j-te Variable X_j werden zu einem m-dimensionalen Vektor $\mathbf{x}_j$ zusammengefasst, und diese Vektoren werden wiederum zu einer (m, p)-Matrix X integriert. Die Struktur von X und der weiter unten eingeführten Vektoren $\mathbf{y}$ und $\mathbf{u}$ werden in (1.1) gezeigt:

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_1 1} \\ \hline y_{12} \\ y_{22} \\ \vdots \\ y_{n_2 2} \\ \hline \vdots \\ \hline y_{1K} \\ y_{2K} \\ \vdots \\ y_{n_K K} \end{pmatrix}, \quad X = \left(\begin{array}{cccc} X_{111} & X_{112} & \cdots & X_{11p} \\ X_{211} & X_{212} & \cdots & X_{21p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_1 11} & X_{n_1 12} & \cdots & X_{n_1 1p} \\ \hline X_{121} & X_{122} & \cdots & X_{12p} \\ X_{221} & X_{222} & \cdots & X_{22p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_2 21} & X_{n_2 22} & \cdots & X_{n_2 2p} \\ \hline \vdots & \vdots & \cdots & \vdots \\ \hline X_{1K1} & X_{1K2} & \cdots & X_{1Kp} \\ X_{2K1} & X_{2K2} & \cdots & X_{2Kp} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_K K1} & X_{n_K K2} & \cdots & X_{n_K Kp} \end{array}\right), \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix} \tag{1.1}$$

Die Elemente von X haben hier drei Indices: der erste zeigt den Fall in einer Gruppe oder Kategorie an, der zweite die Gruppe oder Kategorie, und der dritte die jeweilige Variable. Das Ziel der Analyse ist, einen Teilraum des von den $\mathbf{x}_j$ aufgespannten Vektorraums zu bestimmen, in dem die Fälle der verschiedenen Kategorien maximal getrennt erscheinen. Im einfachsten Fall ist dieser Teilraum 1-dimensional, d.h. eine Gerade, auf den die Konfiguration der Fälle projiziert wird und dessen Orientierung so gewählt wird, dass die Mittelwerte für die einzelnen Kategorien maximal separiert werden. Die Gerade wird durch den Vektor

$$\mathbf{y} = X\mathbf{u} \tag{1.2}$$

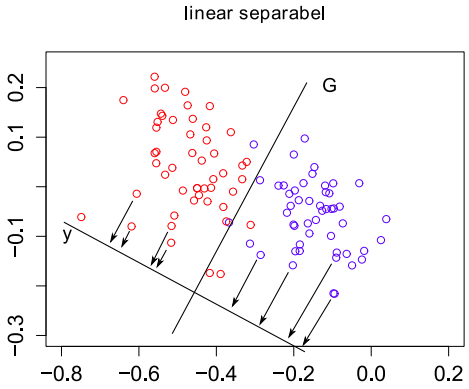
definiert. Falls der Teilraum mehr als nur eine Dimension hat, etwa r, so gibt es r Vektoren $\mathbf{y}_1, \ldots, \mathbf{y}_r$ und ebenso viele Vektoren $\mathbf{u}_1, \ldots, \mathbf{u}_r$, die zu Matrizen $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_r]$ und $U = [bu_1, \ldots, \mathbf{u}_r]$ zusammengefasst werden können; (1.2) nimmt dann die Form

$$Y = XU \tag{1.3}$$

an.

Die Aufgabe ist nun, die Vektoren $\mathbf{u}_\ell$, $\ell = 1, \ldots, r$ aus den Daten zu schätzen.

Abbildung 2: Klassifikation nach Fisher (1936) (I): Es gibt $K = 2$ Klassen Ω_1 (blau), Ω_2 (rot), eine mögliche Trennlinie G, eine zu G orthogonale Projektionsgerade Y, wobei Y durch einen Vektor $\mathbf{y}$ bestimmt wird.



Aus (1.1) geht hervor, dass für jede Klasse ein Mittelwert $\bar{y}_k$ und eine Varianz s_k^2 berechnet werden können. s_k^2 ist eine Schätzung der Varianz innerhalb einer Klasse. Wie bei der Varianzanalyse kann dann die durchschnittliche Varianz "innerhalb" der Klassen bestimmt werden. Ebenso kann aus den $\bar{y}_k$ eine Varianz "zwischen" den Klassen berechnet werden; es genügt, die zugehörigen Quadratsummen zu bilden. Die Quadratsummen ergeben sich wie bei der Varianzanalyse aus Zerlegung der Gesamtvarianz bzw. der zur Gesamtvarianz korrespondierenden Gesamtquadratumme der Komponenten von $\mathbf{y}$. Es ergibt sich die bekannte Zerlegung

$$QSges = QS_{zw} + QS_{inn}, \tag{1.4}$$

mit QS_{ges} die Quadratsumme für die Gesamtvarianz, QS_{zw} die Quadratsumme "zwischen" den Klassen (korrespondierend zur Variaz der $\bar{y}_k$), und QS_{inn} die Quadratsumme der Varianz "innerhalb" der Klassen. $\mathbf{y}$ wird durch die Maximierung des Quotienten

$$\lambda = \lambda(u_1, \cdots, u_p) = \frac{QS_{zw}(u_1, \cdots, u_p)}{QS_{inn}(u_1, \cdots, u_p)} \tag{1.5}$$

als Funktion des Vektors $\mathbf{u} = (u_1, \ldots, u_p)'$ bestimmt. Im einfachsten Fall hat man den Fall von nur zwei Klassen (Fisher, 1936):

$$\lambda = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_1^2 + s_2^2} \tag{1.6}$$

wobei s_1^2 und s_2^2 die Varianzen in den zwei Klassen (Gruppen) G_1 und G_2 sind. Hier wird der allgemeine Fall entwickelt werden. Dazu hat man die folgenden Definitionen:

Definition 1.1 *Der Vektor* $\boldsymbol{y} = X\boldsymbol{u}$ *heißt* lineare Diskriminanzfunktion *oder* kanonische Variable*. Die in (1.5) definierte Größe* λ *heißt* Diskriminanzkriterium*.*

Der Begriff der Diskriminanzfunktion wurde zuerst von Fisher (1936) eingeführt. Die alternative Bezeichnung *kanonische Variable* ergibt sich aus einem Zusammenhang mit der Kanonischen Korrelation, auf den später (vergl. Abschnitt 1.8) noch eingegangen wird. Man erhält $\mathbf{y}$, indem man den Vektor $\mathbf{u}$ bestimmt. Dies ist dann die *Diskriminanzanalyse*. Der Ausdruck 'lineare Diskriminanzanalyse' ergibt sich aus dem Ansatz, dass für $\mathbf{y}$ der lineare Ansatz $\mathbf{y} = X\mathbf{u}$ verwendet wird.

1.2 Die Varianzzerlegung

Gemäß (1.1) hat man

$$\begin{aligned} y_{ik} &= u_1 x_{ik1} + u_2 x_{ik2} + \cdots + u_p x_{ikp}, \quad i = 1, \ldots, n_k && (1.7)\\ \bar{y}_k &= u_1 \bar{x}_{k1} + u_2 \bar{x}_{k2} + \cdots + u_p \bar{x}_{kp} && (1.8)\\ \bar{y} &= u_1 \bar{x}_1 + u_2 \bar{x}_2 + \cdots + u_p \bar{x}_p && (1.9) \end{aligned}$$

wobei $\bar{y}_k$ der Mittelwert für die k-te Gruppe und $\bar{y}$ der Gesamtmittelwert ist.

Es sei QS_{ges} die Quadratsumme, die berechnet werden muß, wenn man die Gesamtvarianz aller y_{ik}-Werte berechnen möchte. Es zeigt sich, daß man QS_{ges} in Teilsummen zerlegen kann, die der Varianz zwischen den Gruppen und der gemittelten Varianz innerhalb der Gruppen entspricht. Es gilt insbesondere der

Satz 1.1 *Es sei* $N = n_1 + \cdots + n_K$ *und*

$$\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{ik}, \quad \bar{y} = \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{n_k} y_{ik},$$

$$QS_{ges} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ik} - \bar{y})^2, \quad QS_{inn} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ik} - \bar{y}_k)^2, \quad QS_{zw} = \sum_{k=1}^{K} n_k (\bar{y}_k - \bar{y})^2. \tag{1.10}$$

Dann gilt

$$QS_{ges} = QS_{zw} + QS_{inn} \tag{1.11}$$

Beweis: Quadratsummenzerlegung wie bei der Regressions- bzw. Varianzanalyse.

□

Natürlich kann es sein, daß nur eine Skala oder Dimension Y nicht hinreicht, um zu einer optimalen Zuordnung zu kommen; die im Folgenden zu beschreibende Analyse liefert alle Skalen, die *für die gegebenen Messungen* $X_1, \cdots, X_p$ die jeweils optimale Entscheidung erlauben.

Um die Abhängigkeit von den u_j, $1 \leq j \leq p$ explizit zu machen, muß (1.5) umgeschrieben werden. Durch Einsetzen ergibt sich

$$QS_{inn} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (u_1(X_{k1i} - \bar{x}_{k1}) + \cdots + u_p(X_{kpi} - \bar{x}_{kp}))^2 \quad (1.12)$$

$$QS_{zw} = \sum_{k=1}^{K} n_k (u_1(\bar{x}_{k1} - \bar{x}_1) + \cdots + u_p(\bar{x}_{kp} - \bar{x}_p))^2 \quad (1.13)$$

$$QS_{ges} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (u_1(X_{ik1} - \bar{x}) + \cdots + u_p(X_{ikp} - \bar{x}))^2 \quad (1.14)$$

Man kann nun die rechten Seiten von (1.12) und (1.13) in den Ausdruck (1.5) für $\lambda(u_1, \cdots, u_p)$ einsetzen, bezüglich der u_j maximieren (d.h. nach den u_j differenzieren, die Ableitungen gleich Null setzen und nach den $\hat{u}_j$, für die diese Gleichungen gelten, auflösen). Aber diese Maximierung wird (i) einfacher, und (ii) ergibt sich eine bessere Vergleichbarkeit mit anderen Methoden, wenn (1.5) und damit die Ausdrücke für QS_{zw} und QS_{inn} in Matrixform angeschrieben werden.

1.3 Die Matrixschreibweise für Quadrate von Summen

Quadrate von Summen können leicht vektoriell dargestellt werden. Es sei $\mathbf{u} = (u_1, \ldots, u_p)'$ und $\mathbf{z} = (z_1, \ldots, z_p)'$ ein beliebiger Vektor. Es gilt

$$(\mathbf{u}'\mathbf{z})^2 = (u_1 z_1 + u_2 z_2 + \cdots + u_p z_p)^2 = u_1^2 z_1^2 + \cdots + u_p^2 z_p^2 + \sum_{i \neq j} u_i u_j z_i z_j. \quad (1.15)$$

Für die Produkte $z_i z_j$ kann man das äußere Produkt

$$\mathbf{z}\mathbf{z}' = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{pmatrix} (z_1, z_2, \ldots, z_p) = \begin{pmatrix} z_1^2 & z_1 z_2 & \cdots & z_1 z_p \\ z_2 z_1 & z_2^2 & \cdots & z_2 z_p \\ \vdots & \vdots & \ddots & \vdots \\ z_p z_1 & z_p z_2 & \cdots & z_p^2 \end{pmatrix}$$

betrachten, und man rechnet leicht nach, dass dann

$$\mathbf{u}' \begin{pmatrix} z_1^2 & z_1 z_2 & \cdots & z_1 z_p \\ z_2 z_1 & z_2^2 & \cdots & z_2 z_p \\ \vdots & \vdots & \ddots & \vdots \\ z_p z_1 & z_p z_2 & \cdots & z_p^2 \end{pmatrix} \mathbf{u} = \mathbf{u}'\mathbf{z}\mathbf{z}'\mathbf{u} = u_1^2 z_1^2 + \cdots + u_p^2 z_p^2 + \sum_{i \neq j} u_i u_j z_i z_j$$

ist, d.h. man hat

$$(\mathbf{u}'\mathbf{z})^2 = \mathbf{u}'\mathbf{z}\mathbf{z}'\mathbf{u}. \tag{1.16}$$

Es sei $x_{kij} = X_{kij} - \bar{x}$ und $\mathbf{x}_{ki} = (x_{ki1}, \ldots, x_{kip})'$. Der Ausdruck für QS_{ges} läßt sich dann in der Form

$$T = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{j=1}^{p} u_j \mathbf{x}_{ki} \mathbf{x}'_{ki} u_j \tag{1.17}$$

schreiben. .

Es sei nun $x_{kji} = X_{kji} - \bar{x}_{kj}$, und $\mathbf{x}_{ki} = (x_{k1i}, \ldots, x_{kpi})'$. Dann ist (vergl. (1.12))

$$(u_1 x_{k1i} + u_2 x_{k2i} + \cdots + u_p x_{kpi})^2 = \mathbf{u}'\mathbf{x}_{ki}\mathbf{x}'_{ki}\mathbf{u}. \tag{1.18}$$

und

$$\begin{aligned} QS_{inn} &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbf{u}'\mathbf{x}_{ki}\mathbf{x}'_{ki}\mathbf{u} \\ &= \mathbf{u}' \left(\sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbf{x}_{ki}\mathbf{x}'_{ki} \right) \mathbf{u} \end{aligned} \tag{1.19}$$

Es werde

$$W = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbf{x}_{ki}\mathbf{x}'_{ki} \tag{1.20}$$

gesetzt; W ist die Matrix der zusammengefassten ("pooled") Varianz-Kovarianz-Quadratsummen "innerhalb", und statt (1.19) kann man

$$QS_{inn} = \mathbf{u}' W \mathbf{u} \tag{1.21}$$

schreiben. Für QS_{zw} erhält man analog

$$QS_{zw} = \sum_{k=1}^{K} n_k \mathbf{u}'(\bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}})'\mathbf{u} = \mathbf{u}' \left(\sum_{k=1}^{K} n_k (\bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}})' \right) \mathbf{u}. \tag{1.22}$$

Setzt man

$$B = \sum_{k=1}^{K} n_k (\bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}})(\bar{\mathbf{x}}_{k\cdot} - \bar{\mathbf{x}})' \tag{1.23}$$

so kann man

$$QS_{zw} = \mathbf{u}' B \mathbf{u} \tag{1.24}$$

schreiben. B ist die Varianz-Kovarianz-Matrix der Mittelwerte.

Wie man leicht nachrechnet, gilt

$$T = B + W. \tag{1.25}$$

Anmerkung: QS_{inn} und QS_{zw} sind Quadratsummen der y-Werte; wie (1.11) (Seite 7) zeigt und wie von der Varianzanalyse bekannt ist, treten bei der Zerlegung von QS_{ges} keine Kreuzprodukte, also keine Kovarianzen auf. Substituiert man aber für die y-Werte die Ausdrücke $u_1\mathbf{x}_1 + \cdots + u_p\mathbf{x}_p$ (vergl. (1.12), (1.13)), so treten die Kovarianzen zwischen den Komponenten der $\mathbf{x}_{jk}$ auf, d.h. die Quadratsummen QS_{inn} und QS_{zw} sind durch die Varianz-Kovarianz-Matrizen W und B definiert.

Mit (1.21) und (1.24) erhält man für $\lambda = QS_{inn}/QS_{zw}$ den Ausdruck

$$\lambda(\mathbf{u}) = \frac{QS_{zw}}{QS_{inn}} = \frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'W\mathbf{u}} \tag{1.26}$$

Die Aufgabe ist nun, λ als Funktion von $\mathbf{u}$ zu maximieren.

1.4 Bestimmung der Lösung für u und λ

Der Vektor $\mathbf{u}$ der Gewichte soll so gewählt werden, dass $\lambda = \mathbf{u}'B\mathbf{u}/\mathbf{u}'W\mathbf{u}$ maximal wird. Zur Vorbereitung werde der Begriff des Rayleigh-Quotienten eingeführt: Es sei A eine symmetrische Matrix; dann heißt

$$\lambda = \frac{\mathbf{x}'A\mathbf{x}}{\mathbf{x}'\mathbf{x}} \tag{1.27}$$

ein *Rayleigh-Quotient*. Im Anhang, Abschnitt 5.1 wird der Satz von Courant-Fischer bewiesen, demzufolge λ maximal wird, wenn für $\mathbf{x}$ der erste Eigenvektor $\mathbf{p}_1$ von A eingesetzt wird; dann ist $\lambda = \lambda_1$ und λ_1 ist der zu $\mathbf{P}_1$ korrespondierende Eigenwert; für $\mathbf{x} = \mathbf{p}_n$ der Eigenvektor, der zum kleinsten Eigenwert λ_n korrespondiert, wird $\lambda = \lambda_n$ minimal.

Es gilt nun der

Satz 1.2 *Der gesuchte Vektor* $\boldsymbol{u}$ *von Gewichten ist ein Eigenvektor der Matrix* $W^{-1}B$*, und das Diskriminanzkriterium* λ *ist der zugehörige Eigenwert dieser Matrix, d.h. es gilt*

$$W^{-1}B\boldsymbol{u} = \lambda\boldsymbol{u} \tag{1.28}$$

Beweis: Der Satz kann bewiesen werden, indem man die partiellen Ableitungen $\partial\lambda/\partial u_j$ bildet und gleich Null setzt; die Lösungen des entstehenden Gleichungssystems liefern die $\hat{u}_j$, für die λ maximal wird. Hier wird ein auf dem Satz von Courant-Fischer beruhender Beweis vorgeführt. Er hat den Vorteil, auf eine Aussage über die Beziehung zwischen Diskriminanzfunktionen zu führen, s. Abschnitt 5. Gesucht ist

$$\max_{\mathbf{u}\neq 0} \frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'W\mathbf{u}}.$$

Sowohl B wie auch W sind symmetrische Matrizen. Also existiert für W die Spektralzerlegung $W = P\Lambda P'$. Mit $\Lambda^{1/2}$ werde die Diagonalmatrix bezeichnet, in deren Diagonalzellen die Wurzeln $\sqrt{\lambda_j}$ der Eigenwerte λ_j von W stehen; offenbar gilt $\Lambda^{1/2}\Lambda^{1/2} = \Lambda$, so dass man $W = P\Lambda^{1/2}\Lambda^{1/2}P'$ schreiben kann. Man definiert $W^{1/2} = P\Lambda^{1/2}$ als die *Wurzel* der Matrix W; in der Tat findet man

$$W^{1/2}W^{1/2} = W = P\Lambda^{1/2}\Lambda^{1/2}P' = P\Lambda P'.$$

Weiter sei

$$\mathbf{v} = W^{1/2}\mathbf{u}.$$

Dann ist

$$\lambda = \frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'W^{1/2}W^{1/2}\mathbf{u}} = \frac{\mathbf{v}'W^{-1/2}BW^{-1/2}\mathbf{v}}{\mathbf{v}'\mathbf{v}},$$

denn $\mathbf{u} = W^{-1/2}\mathbf{v}$. Damit ist λ ein Rayleigh-Quotient in Bezug auf die symmetrische Matrix $A = W^{-1/2}BW^{-1/2}$, und λ nimmt nach dem Satz von Courant-Fischer (s. Anhang, Abschnitt 5.1, Seite 94) den maximalen Wert λ_1 an, der sich als größter Eigenwert von A ergibt, wenn $\mathbf{v} = \mathbf{v}_1$ der zugehörige Eigenvektor von A ist. Es muß demnach

$$W^{-1/2}BW^{-1/2}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$$

gelten. Multipliziert man von links mit $W^{-1/2}$, so ergibt sich

$$W^{-1}BW^{-1/2}\mathbf{v}_1 = \lambda_1 W^{-1/2}\mathbf{v}_1.$$

Es war aber $W^{-1/2}\mathbf{v}_1 = \mathbf{u}_1$, so dass

$$W^{-1}B\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

folgt, d.h. $\mathbf{u}_1$ ist ein Eigenvektor von $W^{-1}B$ und λ_1 ist der zugehörige Eigenvektor.

$\mathbf{P}_1$ muß nicht der einzige Eigenvektor von $M = W^{-1}B$ sein. Man bildet die Matrix $\tilde{M} = M - \lambda_1\mathbf{P}_1\mathbf{P}_1'$ und wendet den Satz 1.2 auf $\tilde{M}$ an, so dass sich $\mathbf{P}_2$ mit dem zugehörigen Eigenwert λ_2 ergibt, etc. □

Fasst man alle Eigenvektoren $\mathbf{u}$ zu einer Matrix U und alle Eigenwerte zu einer Diagonalmatrix Λ zusammen, so kann (1.28) in der Form

$$W^{-1}BU = U\Lambda \tag{1.29}$$

schreiben.

Anmerkung: Satz 1.2 bezieht sich zunächst nur auf einen Gewichtsvektor $\mathbf{u}$. Andererseits ist es möglich, dass die Matrix $W^{-1}B$ mehr als einen von Null verschiedenen Eigenwert λ hat, so dass mehr als ein Vektor $\mathbf{u}$ existiert. Es ist also möglich, dass $r > 1$ Eigenwerte $\lambda_j \neq 0$ und damit r Eigenvektoren $\mathbf{u}_j$, $j = 1, \ldots, r$ existieren. Dementsprechend ist U eine $(p \times r)$-Matrix, wobei der

Wert von r durch Inspektion der Ergebnisse festgesetzt wird – nach Maßgabe des durch die einzelnen Eigenwerte erklärten Anteils an QS_{zw}. □

Da die Matrix $W^{-1}B$ nicht symmetrisch ist, sind die $\mathbf{u}_j$ zwar linear unabhängig, aber nicht notwendig orthogonal. Deswegen ist der folgende Satz bemerkenswert:

Satz 1.3 *$W^{-1}B$ habe mehr als einen von Null verschiedenen Eigenwert. Die zu diesen Eigenwerten korrespondierenden (Rechts-)Eigenvektoren $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_r$ definieren Diskriminanzfunktionen $\boldsymbol{y}_j = X\boldsymbol{u}_j$, $j = 1, \ldots, r$ und es gilt*

$$\boldsymbol{y}_j'\boldsymbol{y}_k = 0, \quad j \neq k, \tag{1.30}$$

d.h. die Diskriminanzfunktionen (Kanonische Variablen) sind orthogonal.

Beweis: Setzt man $X = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p]$, d.h. X sei eine Matrix, deren Spaltenvektoren die $\mathbf{x}_j$ sind, so ist

$$\mathbf{y}_j = X\mathbf{u}_j. \tag{1.31}$$

X sei spaltenzentriert, d.h. die Spalten von X seien durch $\mathbf{x}_j = \mathbf{X}_j - \bar{\mathbf{x}}_j$ definiert, wobei die Komponenten von $\bar{\mathbf{x}}_j$ gleich dem Mittelwert der j-ten Prädiktorwerte sind:

$$\bar{\mathbf{x}}_j = (\bar{x}_j, \bar{x}_j, \ldots, \bar{x}_j)'.$$

Es ist dann

$$\mathbf{y}_j'\mathbf{y}_k = \mathbf{u}_j'X'X\mathbf{u}_k = \mathbf{u}_j'\Sigma\mathbf{u}_k,$$

wobei $\Sigma = X'X$ proportional zur Matrix der Kovarianzen zwischen den $\mathbf{x}_j$ ist (der Proportionalitätsfaktor ist $1/m$). Nun wird Σ aber durch die Matrix W geschätzt (vergl. Gleichung (1.21), Seite 9), $\mathbf{u}_j$ sind die Eigenvektoren von $W^{-1}B$ und $\mathbf{u}_j = W^{-1/2}\mathbf{v}_j$, $\mathbf{v}_j$ ein Eigenvektor der symmetrischen Matrix $W^{-1/2}BW^{-1/2}$, d.h. $\mathbf{v}_j'\mathbf{v}_j = 1$ und $\mathbf{v}_j'\mathbf{v}_k = 0$ für $j \neq k$. Schreibt man also W für Σ, so erhält man

$$\mathbf{y}_j'\mathbf{y}_k = \mathbf{v}_j'W^{-1/2}WW^{-1/2}\mathbf{v}_k = \mathbf{v}_j'\mathbf{v}_k = 0,$$

und das war zu zeigen. □

Falls es also mehrere Eigenwerte ungleich Null gibt, existieren dazu korrespondierende Eigenvektoren, die in einer Matrix U zusammengefasst werden können; fasst man die $\mathbf{y}$ dann zu einer Matrix Y zusammen, so erhält man

$$Y = XU, \quad U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1r} \\ u_{21} & u_{22} & \cdots & u_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pr} \end{pmatrix} \tag{1.32}$$

für die orthogonale Matrix Y der kanonischen Variablen.

1.5 Klassifikation von Beobachtungen

Es werde zunächst die klassifikation des i-ten Falles betrachtet. Die Messwerte sind durch die i-te Zeile $\tilde{\mathbf{x}}_i$ von X gegeben, und die Koordinaten dieses Falles auf den den neuen Achsen sind $\tilde{\mathbf{y}}_i$: man hat gemäßt (1.32)

$$\tilde{\mathbf{y}}_i' = \tilde{\mathbf{x}}_i' U. \tag{1.33}$$

Hier ist berücksichtigt worden, dass Vektoren stets als Spaltenvektoren angeschrieben werden, so dass $\tilde{\mathbf{y}}_i'$ der als Zeilenvektor geschriebene Vektor $\tilde{\mathbf{y}}_i$ ist, – analog für $\tilde{\mathbf{x}}_i$. Geht man wieder zur Schreibweise von Spaltenvektoren über, so erhält man

$$\tilde{\mathbf{y}}_i = U'\tilde{\mathbf{x}}_i = \begin{pmatrix} u_{11}, & u_{21}, & \cdots & u_{ps} \\ u_{12}, & u_{22}, & \cdots & u_{p1} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1r}, & u_{2r}, & \cdots & u_{pr} \end{pmatrix} \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{is} \end{pmatrix}, \tag{1.34}$$

Die Komponenten $y_{i1}, \ldots, y_{ir}$ von $\mathbf{y}$ sind hier die Koordinaten für den durch $\tilde{\mathbf{x}}_i$ gegebenen Fall im Raum der kanonischen Variablen. Man kann auch die Klassifikation eines neuen, nicht in X enthaltenen Falls betrachten: $\tilde{\mathbf{x}}$ ist denn einfach ein Vektor mit den p Messungen der Variablen und $\tilde{\mathbf{y}}$ ist der zugehörige Vektor auf den neuen Variablen. Die folgenden Betrachtungen gelten für diesen Fall; soll der i-te Fall der Matrix X betrachtet werden, muß nur der Index i zugefügt werden.

Es sei also

$$\bar{\mathbf{y}}_k = \begin{pmatrix} \bar{y}_{1k} \\ \bar{y}_{2k} \\ \vdots \\ \bar{y}_{sk} \end{pmatrix}$$

sei der Vektor des Centroids, also des Schwerpunkts der Punkte $\mathbf{y}$, die zur k-ten Gruppe bzw Kategorie gehören; $\bar{y}_{jk}$ ist der Mittelwert der y-Werte auf der j-ten kanonischen Variablen in der k-ten Gruppe. Das i-te Objekt aus der k-ten Gruppe hat die x-Werte

$$x_{i1k}, x_{i2k}, \ldots, x_{ipk},$$

und der zugehörige y-Wert auf der j-ten kanonischen Variablen ist durch

$$y_{ijk} = (x_{i1k}, x_{i2k}, \ldots, x_{ipk})\mathbf{u}_j = (x_{i1k}, x_{i2k}, \ldots, x_{ipk}) \begin{pmatrix} u_{1j} \\ u_{2j} \\ \vdots \\ u_{pj} \end{pmatrix}$$

gegeben. Dann ist

$$\bar{y}_{jk} = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{ijk} = \frac{1}{n_k} \left(\left(\frac{1}{n_k} \sum_{i=1}^{n_k} x_{ijk} \right) u_{11} + \cdots + \left(\frac{1}{n_k} \sum_{i=1}^{n_k} x_{ipk} \right) u_{jp} \right),$$

d.h.

$$\bar{y}_{jk} = \bar{x}_{1k}u_{1j} + \bar{x}_{2k}u_{2j} + \cdots + \bar{x}_{pj}u_{pj}, \quad j = 1, \ldots, s \tag{1.35}$$

Dann ist[2]

$$\tilde{\mathbf{y}} - \bar{\mathbf{y}}_k = U'\mathbf{x} - U'\bar{\mathbf{x}}_k = U'(\tilde{\mathbf{x}} - \bar{\mathbf{x}}_k). \tag{1.36}$$

Man hat dann den

Satz 1.4 *Es gilt*

$$\|\tilde{\boldsymbol{y}} - \bar{\boldsymbol{y}}_k\|^2 = (\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}_k)'W^{-1}(\tilde{\boldsymbol{x}} - \bar{\boldsymbol{x}}_k), \tag{1.37}$$

d.h. im Raum der kanonischen Variablen ist die euklidische Distanz des zu $\boldsymbol{x}$ gehörenden Punktes $\boldsymbol{y}$ zum Centroid $\bar{\boldsymbol{y}}_k$ der k-ten Gruppe ist gleich der Mahalanobis-Distanz des Punktes $\boldsymbol{x}$ zum Centroid der k-ten Gruppe im Raum der Messwerte.

Beweis: Es sei $W^{-1/2}$ die Wurzel der Matrix W^{-1}, so dass $W^{-1/2}W^{-1/2} = W^{-1}$ und $W^{-1/2}W^{1/2} = I$ (s.Anhang, Abschnitt 5.2). Aus der Gleichung (1.29) folgt dann

$$\begin{aligned} W^{-1}BU &= W^{-1/2}W^{-1/2}BW^{-1/2}W^{1/2}U = U\Lambda \\ \Rightarrow\ W^{-1/2}BW^{-1/2}\underbrace{W^{1/2}U}_{P} &= \underbrace{W^{1/2}U}_{P}\Lambda. \end{aligned}$$

P ist die Matrix der orthonormalen Eigenvektoren der symmetrischen Matrix $W^{-1/2}BW^{-1/2}$, so dass $P'P = PP' = I$ und $U = W^{-1/2}P$, so dass $UU' = W^{-1/2}PP'W^{-1/2} = W^{-1}$, und wegen (1.36) folgt

$$\|\mathbf{y} - \bar{\mathbf{y}}_k\|^2 = (\mathbf{x} - \bar{\mathbf{x}}_k)'UU'(\mathbf{x} - \bar{\mathbf{x}}_k) = (\mathbf{x} - \bar{\mathbf{x}}_k)'W^{-1}(\mathbf{x} - \bar{\mathbf{x}}_k),$$

und dies war zu zeigen. □

Anmerkung: Der Beziehung (1.37) liegt *nicht* die Annahme der multivariaten Normalverteilung zugrunde, die ja ebenfalls durch die Mahalanobis-Distanz $(\mathbf{x} - \bar{\mathbf{x}}_k)'W^{-1}(\mathbf{x} - \bar{\mathbf{x}}_k)$ definiert ist; die Beziehung (1.37) folgt rein algebraisch, d.h. ohne Bezug auf wahrscheinlichkeitstheoretische Annahmen. □

Das Objekt ω mit dem Vektor $\mathbf{x} = \mathbf{x}(\omega)$ und dem zugehörigen Vektor $\mathbf{y} = U'\mathbf{x}$ soll nun einer Klasse Ω_k zugeordnet werden. Es wird die folgende Entscheidungsregel eingeführt:

Regel:

$$\begin{aligned} \omega \to \Omega_k \Longleftrightarrow \|\mathbf{y} - \bar{\mathbf{y}}_k\|^2 &= \min_j \|\mathbf{y} - \bar{\mathbf{y}}_j\|^2 && (1.38) \\ &= \min_j (\mathbf{x} - \bar{\mathbf{x}}_j)'W^{-1}(\mathbf{x} - \bar{\mathbf{x}}_j), && (1.39) \end{aligned}$$

[2]Es müßte $\tilde{\bar{\mathbf{y}}}_k$ geschrieben werden, aber diese Schreibeweise ist doch ein wenig umständlich, so dass einfach $\bar{\mathbf{y}}_k$ geschrieben wurde; analog für $\bar{\mathbf{x}}_k$.

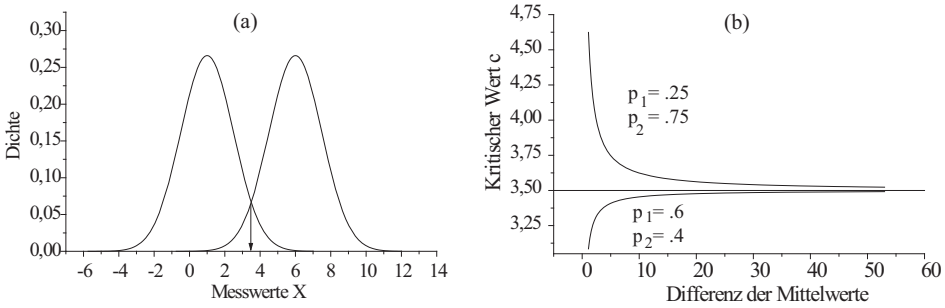
wobei $\Longleftrightarrow$ für "genau dann, wenn" steht. Man ordnet ω derjenigen Klasse G_k zu, für die $\|\mathbf{y} - \bar{\mathbf{y}}_k\|$ minimal ist. Das ist gleichzeitig diejenige Klasse, für die die Mahalanobis-Distanz zwischen $\mathbf{x}$ und $\bar{\mathbf{x}}_k$ minimal ist.

Spätestens bei der Berechnung von $\mathbf{y}$ für einen neuen Fall $\mathbf{x}$ stellt sich die Frage, wieviele kanonische Variablen betrachtet werden solle. Gibt es K Klassen. Kategorien oder Gruppen, werden auf jeden Fall $K-1$ kanonische Variable berechnet. Die im folgenden Abschnitt vorgestellte Beziehung zwischen Diskriminanzanalyse und Kanonischer Korrelation erleichtert die Beantwortung der Frage.

1.6 Beispiele I

Beispiel 1.1 Der einfachste Fall ergibt sich für $p = 1$, wenn also nur eine Größe gemessen wird, und wenn nur zwei Gruppen betrachtet werden, also $g = 2$ ist. x sei also normalverteilt, $N(\mu_k, \sigma^2)$, und $k = 1, 2$. Für einen gegebenen Messwert x soll entschieden werden, ob er von einem Objekt oder einer Person aus der Klasse Ω_1 oder aus der Klasse Ω_2 ist. Für μ_1 und μ_2 werden die Mittelwerte $\bar{x}_1$ und $\bar{x}_2$ eingesetzt, für σ^2 berechnet man die aus allen Daten berechnete Schätzung s^2.

Abbildung 3: (a) Gaussverteilungen mit gleicher Varianz; der Pfeil zeigt auf $c_0 = (\mu_1+\mu_2)/2$. (b) Kritische Werte für verschiedene Werte von p_1/p_2 in Abhängigkeit von der Differenz $\mu_1 - \mu_2$ bei konstantem c_0.



Der Variablenraum ist hier 1-dimensional, d.h. ist eine Gerade. Die Hyperebene, die die Bereiche R_1 und R_2 trennt, ist jetzt ein Punkt auf dieser Gerade, d.h. eine reelle Zahl c. Dieser Punkt entspricht der Lösung $x = c$ der Gleichung (2.56), in der x_0 ebenfalls ein Skalar und kein Vektor ist. Dann ist aber (2.56) genau dann erfüllt, wenn $c = x_0$ ist, und x_0 ist durch (2.55) gegeben. Man findet $\delta = (\bar{x}_1 - \bar{x}_2)/\sigma^2$ und damit

$$c = \frac{1}{2}(\bar{x}_1 + \bar{x}_2) - \frac{\sigma^2}{\bar{x}_1 - \bar{x}_2} \log\left(\frac{p(\Omega_2)}{p(\Omega_1)}\right), \tag{1.40}$$

und c trennt die Bereiche R_1 und R_2; beobachtet man einen $x(\omega)$-Wert größer als c so wird man ω der Klasse Ω_2 zuordnen, andernfalls Ω_2. Für $p(\Omega_1) = p(\Omega_2)$ wird

$\log(p(\Omega_k)/p(\Omega_j)) = 0$ und c halbiert gerade das Intervall zwischen $\bar{x}_1$ und $\bar{x}_2$. Der Effekt unterschiedlicher a priori-Wahrscheinlichkeiten hängt vom Wert der Differenz $\mu_1 - \mu_2$ ab; für größer werdende Differenz strebt c gegen $c_0 = (\mu_1 + \mu_2)/2$ (vergl. Abb. 3). □

Beispiel 1.2 Es sei nun $g = p = 2$, d.h. es werden zwei Variablen x_1 und x_2 und zwei Gruppen betrachtet. Ω_1 ist die Kaste der Brahmanen, und Ω_2 ist die Kaste der Handwerker, nach Rao (1948)

$$S = \begin{pmatrix} 32.948, & 10.743 \\ 10.743, & 10.24 \end{pmatrix}, \quad S^{-1} = 365 \begin{pmatrix} 0.046, & -.048 \\ -.048, & .148 \end{pmatrix}. \tag{1.41}$$

Bezeichnet man also mit $\bar{x}_B$ den Mittelwertsvektor für die Brahmanen und mit

Tabelle 1: Mittelwerte und Varianzen der Variablen Größe und Sitzhöhe für die beiden Gruppen

	Brahmanen	Handwerker	Varianz
Größe	164.51	160.53	32.948
Sitzhöhe	86.43	81.47	10.240

$\bar{x}_A$ den für die Handwerkers, so hat man

$$\bar{x}_B = \begin{pmatrix} 164.51 \\ 86.43 \end{pmatrix}, \quad \bar{x}_A = \begin{pmatrix} 160.53 \\ 81.47 \end{pmatrix}. \tag{1.42}$$

Nach (2.48) ist die Hyperebene, die die Bereiche trennt, durch

$$(\bar{x}_B - \bar{x}_A)' S^{-1} x = \frac{1}{2}(\bar{x}_B - \bar{x}_A)' S^{-1}(\bar{x}_B + \bar{X}_A) - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right). \tag{1.43}$$

Es ist $(\bar{x}_B - \bar{x}_A)' S^{-1} = (b_1, b_2)' = (.056, -.544)'$ und $(\bar{x}_B - \bar{x}_A)' S^{-1}(\bar{x}_B + \bar{X}_A)/2 = -36.460$, mithin ist die Hyperebene durch die Gerade

$$.056x_1 - .544x_2 = -36.460 - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right) \tag{1.44}$$

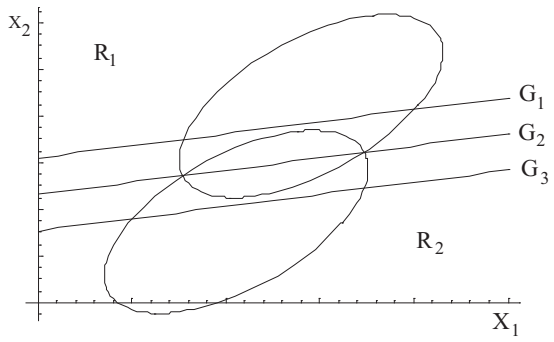
gegeben, bzw. durch

$$x_2 = \left(36.460 + \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right)\right)/.544 - (.056/.544)x_1. \tag{1.45}$$

Die Geraden, die zu verschiedenen Werten von $P(\Omega_B)/p(\Omega_A)$ korrespondieren, sind parallel zueinander.

Die ursprünglichen Messwerte sind nicht gegeben, aber man kann die Ellipsen, die den jeweiligen 2-dimensionalen Normalverteilungen der beiden Gruppen

Abbildung 4: Brahmanen und Handwerker (Beispiel 1.2): Trenn"flächen" für verschiedene a priori-Wahrscheinlichkeiten; G_1: $p_1/p_2 = .3/.7$; G_2: $p_1/p_2 = 1$, G_3: $p_1/p_2 = .7/.3$.



entsprechen, bestimmen, denn sie sind durch die Mahalanobis-Distanz $\delta(x, \bar{x}_i)$, $i = 1, 2$ gegeben.

Für eine zufällig gewählte Person findet man die Messerte $x = (x_1, x_2)'$ für die beiden betrachteten Variablen; liegt x oberhalb der Geraden G, so wird die Person als zu R_1, also zur Kaste der Brahmanen gehörig btrachtet, andernfalls als zur Kaste der Handwerker (R_2) gehörig. □

Beispiel 1.3 Nach Amthauer (1970) erreichen Ärzte, Juristen und Pädagogen in den Untertests Analogien (AN), Figurenauswahl (FA) und Würfelaufgaben (WÜ) des Intelligenz-Struktur-Tests (IST) (IST) Durchschnittswerte, die in Tabelle 2 angegeben werden. Für die durchschnittliche Varianz-Kovarianz-Matrix S und

Tabelle 2: Scores für verschiedene Berufe

	Ärzte	Juristen	Pädagogen
Analogien	114	111	105
Figurenauswahl	111	103	101
Würfelaufgaben	110	100	98

deren Inverse S^{-1} hat man

$$S = \begin{pmatrix} 100 & 30 & 32 \\ 30 & 100 & 44 \\ 32 & 44 & 100 \end{pmatrix}, \quad S^{-1} = \begin{pmatrix} .0115 & -.0023 & -.0027 \\ -.0023 & .0129 & -.0049 \\ -.0027 & -.0049 & .0130 \end{pmatrix} \tag{1.46}$$

Ein Abiturient hat in den gleichen Untertests die folgenden Scores erzielt: AN = 108, FA = 112, WÜ = 101. Die Frage ist, welcher Berufsgruppe der Abiturient

zuzuordnen ist, wenn alle drei Gruppen die gleiche a priori-Wahrscheinlichkeit haben.

Es müssen nur die Mahalanobis-Distanzen (2.45) (Seite 59) zwischen dem Score-Vektor des Abiturienten und den drei Berufgruppen berechnet werden; da $p(\Omega_k)$ konstant ist für $k = 1, 2, 3$, gibt der Wert $\log p(\Omega_k)$ keinerlei Information über die Gruppenzugehörigkeit und kann bei der Berechnung der Distanz weggelassen werden. Man entscheidet für diejenige Gruppe, für die die Mahalanobis-Distanz minimal ist. Für die Gruppe der Ärzte muß also

$$d_1 = (108 - 114, 112 - 111, 101 - 110)S^{-1} \begin{pmatrix} 108 - 114 \\ 112 - 111 \\ 101 - 110 \end{pmatrix} \tag{1.47}$$

berechnet werden; man findet $d_1 = .9441$. Analog findet man für die Gruppe der Juristen $d_2 = 1.1236$, und für die Pädagogen $d_3 = 1.1676$. Die geringste Distanz hat der Abiturient also zu den Medizinern, so dass man ihm empfehlen wird, Arzt zu werden. □

Beispiel 1.4 Bei gesunden Männern wurden die Variablen x_1: Alter, x_2: Blutdruck, und x_3 Cholesterinspiegel gemessen. Die Frage ist, ob sich aus diesen Werten das Risiko für eine spätere Herzgefäßkranzerkrankung feststellen läßt. Am Ende einer Beobachtungsperiode waren 71 der Männer erkrankt. Es ergaben sich die folgenden Messungen:

Tabelle 3: Mittelwerte und Varianzen für Gesunde und Erkrankte

	Arithm. Mittel		Standardabw.	
Variable	Gesunde	Kranke	Gesunde	Kranke
x_1 (Alter)	44.81	56.86	14.98	10.28
x_2 (Blutdruck)	86.99	95.62	14.50	15.37
x_3 (Cholesterin)	210.27	221.51	43.01	38.83

Für die ("Pooled") Varianz-Kovarianz-Matrix ergab sich

$$S = \begin{pmatrix} 214.26 & 72.37 & .61 \\ 72.73\ 985 & 212.44 & 175.53 \\ 195.61 & 175.53 & 1820.61 \end{pmatrix}. \tag{1.48}$$

Zur Illustration wird die inverse Matrix S^{-1} angegeben:

$$S^{-1} = \begin{pmatrix} 214.26 & .72.37 & 195.61 \\ 72.73 & 212.44 & 175.53 \\ 195.61 & 175.53 & 1820.61 \end{pmatrix}. \tag{1.49}$$

Tabelle 4: Ergebnisse

Variable	Koeffizienten u_i	partielle F-Werte	
x_1	.045	22.657	$F = 17.605$
x_2	.022	5.282	$\delta^2 = .815$
x_3	.004	1.675	
Konstante	-5.165		

Tabelle 5: Klassifikation nach der ML-Regel

	ML-Klassifizierung		Σ
Lernstichprobe	krank	gesund	
Kranke	51	20	71
Gesunde	272	489	761

Tabelle 6: Klassifikation nach der Bayes-Regel

	Klassifizierung		Σ
Lernstichprobe	krank	gesund	
Kranke	0	71	71
Gesunde	0	761	761

Die Gewichte der Variablen sind Es ergibt sich ein Gesamt-F-Wert von $F = 17.605$; bei $df = n_1033 + n_2 - 3 - 1 = 828$ ist er hochsignifikant. Dies bedeutet, daß sich die beiden Gruppen (Erkrankte und Gesund) anhand der Risikofaktoren x_1 , x_2 und x_3 gut trennen lassen. Da der partielle F-Wert für x_3 relativ klein ist, ist es möglich, daß der Cholesteringehalt kaum zur Trennung der Gruppen beiträgt.

Nach der Ml-Regel ergeben sich die folgenden Klassifizierungen: Das Verhältnis von als "gesund" klassifizierten Kranken beträgt $20/71 = .282$ oder 28.2%; dies ist der *Resubstitutionsfehler* für die Gruppe der Kranken. Für die Gruppe der Gesunden ergibt sich der entsprechende Fehler als Verhältnis der als "krank" klassifizierten Gesunden, also $272/767 = .357$, oder 35.7%.

Die Mahalanobis-Distanz ist gemäß Tabelle 4 $\delta^2 = .815$. Die plug-in-Schätzung für die ML-Regel ergibt

$$\hat{\epsilon}_{12} = \hat{\epsilon}_{21} = \Phi(-D/2) = \Phi(-\sqrt{.815}/2) = .326.$$

Will man die Bayes-Regel anwenden, so muß man die a-priori-Wahrscheinlichkeiten für die Gruppenzugehörigkeit schätzen. Man erhält

$$\hat{\pi}_1 = \frac{n_1}{n} = \frac{71}{832} = .0853, \quad \hat{\pi}_2 = \frac{n_2}{n} = \frac{761}{832} = .915, \log n_2/n_1 = 2.372$$

Man erhält die folgende Klassifikation

Das Bemerkenswerte ist hier, daß alle Kranken falsch klassifiziert werden. Für die Plug-in-Schätzungen ergeben sich die Werte

$$\hat{\epsilon}_{12} = \Phi\left(\frac{2.372 - .815/2}{\sqrt{.815}}\right) = .985, \quad \hat{\epsilon}_{21} = \Phi\left(-\frac{2.372 + .815/2}{\sqrt{.815}}\right) = .001.$$

Die Bayes-Regel gilt als optimal, führt hier aber zu deutlich schlechteren Vorhersagen als die ML-Regel. Die Ursache dafür ist hier, daß die Bayes-Regel den *Gesamtfehler* minimiert, - und der wird hier minimiert, wenn man eben alle Kranken als gesund klassifiziert. Tatsächlich ist also im vorliegenden Fall die ML-Regel besser. □

Beispiel 1.5 Die Angestellten einer Fluglinie wurden hinsichtlich ihrer Freizeitinteressen getestet; es wurden Werte auf drei Skalen des *Activity Preference Inventory* (API) erhoben: $X_1 =$ "Outdoor", $X_2 =$ "Convivial", und $X_3 =$ "Conservative". Die Angestellten wurden in drei Klassen eingeteilt: p "Passenger Agents", m "Mechanics". und o "Operations Control Agents", vergl. Tabelle 7. Ein hoher Wert auf einer Skala reflektiert eine hohe Präferenz für die entsprechende Aktivität.Die Tabelle 8 gibt die Mittelwerte der drei Variablen für die einzelnen Gruppen.

Tabelle 7: Die Freizeitinteressen von Angestellten einer Fluglinie, p: Passenger Agents, m Mechanics, o Operations control (Beispiel 1.5)

Person	outdoor (X_1)	convivial(X_2)	conservative (X_3)	Klasse
1	10	22	13	p
2	20	25	12	p
3	10	24	5	p
4	13	21	11	p
5	11	22	11	p
6	8	29	14	p
7	22	22	6	p
8	15	21	4	p
9	11	23	5	p
10	12	26	9	p
11	18	26	10	m
12	12	16	10	m
13	17	24	5 5	m
14	15	22	13	m
15	17	19	12	m
16	20	19	11	m
17	17	24	11	m
18	16	19	8	m
19	14	24	7	m
20	16	22	5	m
21	24	14	7	m
22	11	25	12	m
23	17	19	11	m
24	4	12	11	o
25	13	20	16	o
26	13	15	18	o
27	13	16	7	o
28	17	15	10	o
29	11	12	19	o
30	15	16	14	o
31	15	18	14	o
32	4	10	15	o
33	10	12	9	o
34	17	18	9	o
35	15	18	14	o
36	20	13	19	o
37	18	11	19	o

Tabelle 8: Mittelwerte der drei Variablen für die drei Gruppen

	outdoor	convivial	conservative
p	13.200	23.500	9.00
m	16.461	21.000	9.423
o	13.214	14.714	13.857

Tabelle 9: Varianz-Kovarianz-Matrix W

	X_1	X_2	X_3
outdoor (X_1)	609.188	-10.143	13.044
convivial (X_2)	-10.143	343.357	148.429
conservative (X_3)	13.044	-217.996	154.086

Tabelle 10: Varianz-Kovarianz-Matrix B

	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$
$\bar{x}_1$	89.244	71.278	38.740
$\bar{x}_2$	71.278	508.373	-217.996
$\bar{x}_3$	38.740	-217.996	154.086

Die zusammengefasste ("pooled") Varianz-Kovarianz-Matrix wird in der Tabelle 9 angegeben. 181Die Matrix B der Varianzen-Kovarianzen zwischen den Mittelwerten findet man in der Tabelle 10. Die Tabelle 11 enthält die von Null verschiedenen Eigenwerte (Diskriminanzkriteria) λ_1 und λ_2 sowie die zugehörigen Eigenvektoren u_1 und u_2, deren Komponenten die Regressionsgewichte zur Vorhersage auf den maximal diskriminierenden Skalen Y_1 und Y_2 sind. Hier Be-

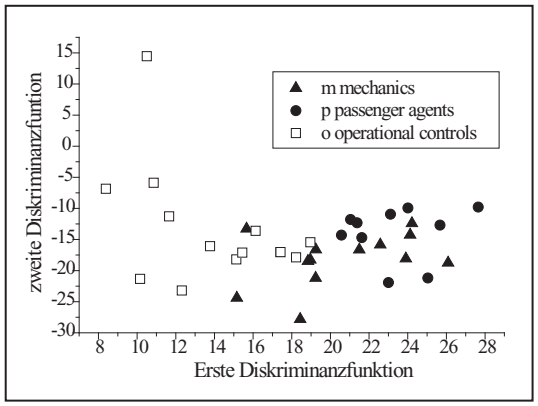
Tabelle 11: Eigenvektoren **u** und Eigenwerte λ

Variable	u_1	u_2
X_1	.091	-.974
X_2	.988	.0386
X_3	-.124	-.221
λ	$\lambda_1 = 1.675$	$\lambda_2 = .155$

merkungen über die Signifikanz der einzelnen Eigenwerte machen!

Der Abbildung 5 entsprechend kann man folgern, dass in erster Linie die Gruppe o, also die Operational Control Agents, von den übrigen beiden Grup-

Abbildung 5: Personal der Fluglinie: Projektion auf die maximal diskriminierende Achse u_1, I



pen unterscheidet, während sich die Gruppen p (Passenger Agents) und m (Mechanics) kaum voneinander unterscheiden. Nach Tabelle 11 hat die Variable X_2 (convivial = heiter, gesellig) bei u_1 mit $u_{12} = .98$ das größte Gewicht, während X_2 auf u_2 mit $u_{21} = -.974$ "lädt". □

1.7 Alternative Herleitung der Lösung für u und λ

Dieser Abschnitt behandelt die Diskriminanzanalyse standardisierter Datenmatrizen; diese Version der DA liegt liegt dem Programm lda aus dem R-Paket MASS[3] zugrunde.

Dem Fisherschen Ansatz zufolge wird $\mathbf{u}$ so bestimmt, dass der Quotient QS_{zw}/QS_{inn} maximiert wird; aus der Definition von QS_{zw} und QS_{inn} folgt, dass die Daten nicht standardisiert in die Analyse eingehen. Es wird der Vektor $\mathbf{u}$ bestimmt, der den Quotienten $\lambda(\mathbf{u}) = \mathbf{u}'B\mathbf{u}/\mathbf{u}'W\mathbf{u}$ maximiert, vergl. (1.26). Dieser Ansatz setzt stillschweigend voraus, dass die Skalen der verschiedenen Prädiktoren vergleichbar sind. Eine mögliche Erweiterung des Ansatzes besteht darin, Die Matrix X der Prädiktoren so umzuskalieren, d.h. mit einer geeignet gewählten Matrix S zu multiplizieren, so dass die zu XS korrespondierende "Innerhalb"-Matrix W zur Einheitsmatrix I wird. $\mathbf{u}$ wird dann so bestimmt, dass

$$\lambda^* = \frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'\mathbf{u}}$$

maximal wird. Da $\mathbf{u}'\mathbf{u}$ ein Skalar ist, läuft die Bestimmung von $\mathbf{u}$ nun auf die Bestimmung der Eigenwerte und -vektoren von B hinaus; der Quotient wird maximal, wenn $\mathbf{u}$ der erste Eigenvektor von B ist. Sollte allerdings B nicht den

[3]The R Project for Statistical Computing, http://www.r-project.org

vollen Rang haben, können sich Schwierigkeiten ergeben. Eine alternative Lösung ergibt sich, wenn X spaltenweise standardisiert wird. Die folgende Betrachtung ist aus Ripley (1996), p. 94.

Es sei also Z die Matrix, die aus X durch Spaltenstandardisierung hervorgeht. Weiter sei

$$Z = U\Lambda V' \tag{1.50}$$

die Singularwertzerlegung (SVD) von Z: man sucht orthogonale Basisvektoren $\mathbf{L}$ für die Spaltenvektoren von Z, so dass diese als Linearkombinationen der Basisvektoren dargestellt werden können, d.h. es soll $Z = LV'$ gelten, wobei V' die Matrix der Koeffizienten für die Spaltenvektoren von L ist. Die vorausgesetzte Orthogonalität von L impliziert $Z'Z = VL'LV' = V\Lambda^2V'$, woraus folgt, dass V die Matrix der Eigenvektoren von $Z'Z$ ist und Λ^2 die Diagonalmatrix der Eigenwerte von $Z'Z$. Die Spaltenvektoren von L werden auf die Länge 1 normiert, wenn L von rechts mit Λ^{-1} multipliziert wird; sei $U = L\Lambda^{-1}$. Dann muß $Z = LV' = U\Lambda V'$, also (1.50) gelten. Da $ZZ' = U\Lambda V'V\Lambda U' = U\Lambda^2U'$ gilt folgt, dass U die normierten Eigenvektoren von ZZ' enthält, die zu den von Null verschiedenen Eigenwerten in Λ^2 korrespondieren. Die Elemente in Λ heißen *Singularwerte*; sie sind die Wurzeln aus den Eigenwerten von $Z'Z$, die identisch mit den von Null verschiedenen Eigenwerten von ZZ' sind.

Die folgende Transformation von Z vereinfacht die Herleitung der Diskriminanzanalyse: Es sei

$$A = \sqrt{n}ZV\Lambda^{-1} \tag{1.51}$$

Während nämlich $Z'Z = pR$, R die Matrix der Korrelationen zwischen den Prädiktoren ist, gilt wegen $Z = U\Lambda V'$

$$A'A = n\Lambda^{-1}V'V\Lambda U'U\Lambda V'V\Lambda^{-1} = nI_p, \tag{1.52}$$

I_p die $(p \times p)$-Einheitsmatrix. Die Nützlichkeit dieses Sachverhalts erweist sich in Gleichung (1.57), Seite 25.

Es sei G die in Gleichung (1.73), Seite 30 definierte Matrix. Dann ist $G'G$ eine Diagonalmatrix mit n_k in den Diagonalzellen; n_k ist die Anzahl der Fälle (Beobachtungen) in der k-ten Gruppe (Kategorie oder Klasse). Weiter sei $T = \text{diag}(\sqrt{n_k/n})$. Dann folgt

$$TG'GT' = nI_p \tag{1.53}$$

I_n die $(p \times p)$-Matrix.

Die Mittelwerte der Spalten von A sind gleich Null. Denn es sei $N = V\Lambda^{-1}$; dann ist $A = ZN$. Sei $\mathbf{n}_j = (n_{1j}, n_{2j}, \ldots, n_{pj})'$ der j-te Spaltenvektor von N, und $\mathbf{a}_j$ der j-te Spaltenvektor von A: dann gilt

$$\mathbf{a}_j = n_{1j}\mathbf{z}_1 + n_{2j}\mathbf{z}_2 + \cdots + n_{pj}\mathbf{z}_p.$$

Da die Summen der Komponenten der $\mathbf{z}_1, \ldots, \mathbf{z}_p$ alle gleich Null sind, muß auch die Summe der Komponenten von $\mathbf{a}_j$ gleich Null sein, d.h. die Mittelwerte der Komponenten der $\mathbf{a}_j$ sind gleich Null.

Mittelt man aber nur für die einzelnen Gruppen, so sind die Gruppenmittelwerte ungleich Null. Die $(K \times p)$-Matrix der Gruppenmittelwerte ist nun durch

$$M = (G'G)^{-1}G'A = \frac{1}{n}S^2G'A \tag{1.54}$$

gegeben. Es sei r der Rang von M. Weiter werde die SVD der Matrix $T^{-1}M$ betrachtet:

$$T^{-1}M = Q\Sigma P', \tag{1.55}$$

wobei natürlich wieder $Q'Q = I$, $P'D = I$, und Σ ist die Diagonalmatrix der Singularwerte von $T^{-1}M$.

B und W seien wieder die "between"- und "within"-Varianz-Kovarianzmatrizen, jetzt allerdings für die Matrix A. Dann hat man

$$\begin{aligned}(K-1)B &= (GM)'(GM)\\ &= (GTQ\Sigma P')'(GTQ\Sigma P')\\ &= D\Sigma Q'T'G'GTQ\Sigma P'\\ &= nD\Sigma Q'Q\Sigma P' = nQD\Sigma^2P'\end{aligned} \tag{1.56}$$

denn $M = TQ\Sigma P'$, $T'G'GT = nT^{-2}$ wegen (1.54) und (1.55) und $Q'Q = I$ wegen der Orthonormalität der Eigenvektoren symmetrischer Matrizen.

$Z'Z$ ist die Matrix der standardisierten Varianzen und Kovarianzen (Korrelationen) "gesamt", entspricht also QS_{ges}, repräsentiert in der Matrix T, für die die Aussage $T = W + B$ gilt (vergl. (1.25), Seite 9).

Für die Matrix W für die Variation "innerhalb" erhält man

$$\begin{aligned}(n-K)W &= A'A - (K-1)B\\ &= nI_n - nD\Lambda^2P' = nV(I_n - \Sigma^2)P'\end{aligned} \tag{1.57}$$

denn $DI_nP' = I_n$ wegen der Orthonormalität von V. W ist symmetrisch, so dass alle Eigenwerte von W positiv sein mussen, so dass alle Diagonalelemente von $I_n - \Lambda^2$ nicht kleiner als Null sein können, woraus wiederum folgt, dass die Elemente von Λ^2 kleiner oder höchstens gleich 1 sein müssen. Gesucht ist nun ein Vektor $\mathbf{r}$, der $\mathbf{r}'B\mathbf{r}/\mathbf{r}'W\mathbf{r}$ maximiert, wobei B und W nun in (1.56) und (1.57) definiert werden. Man erhält

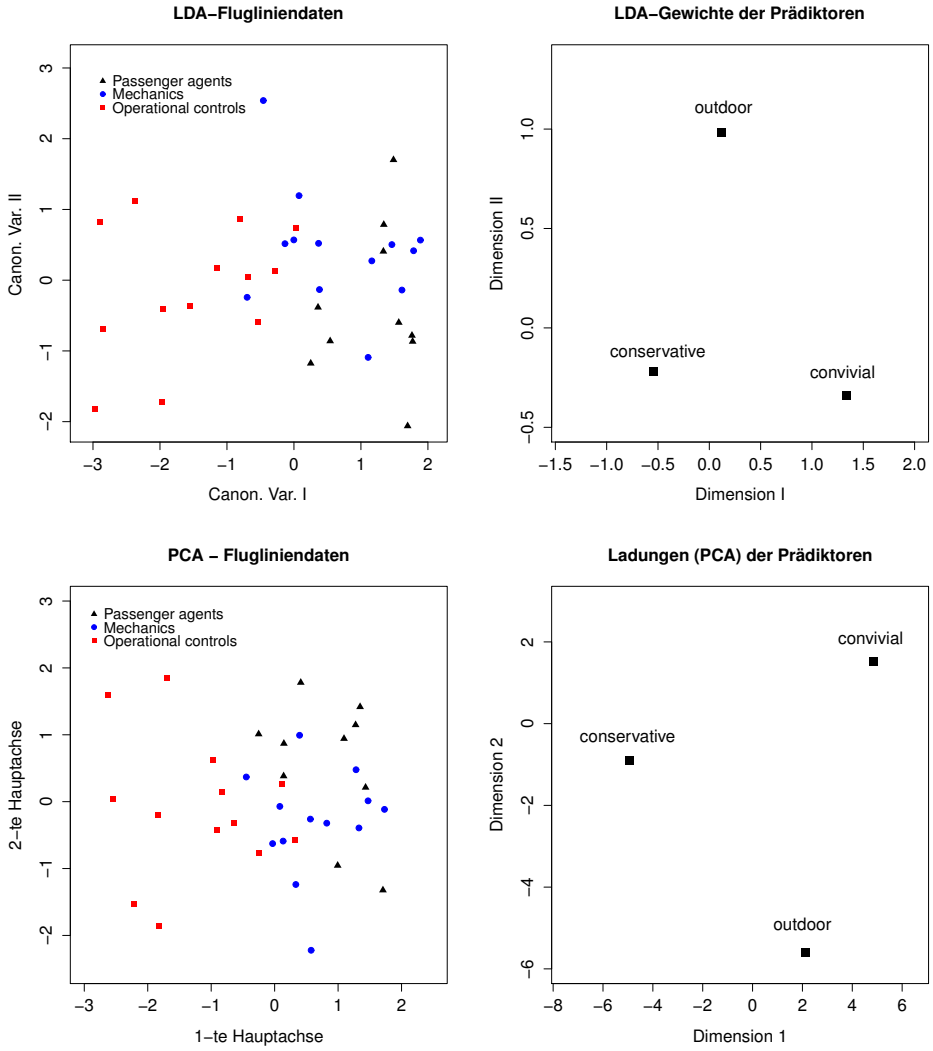
$$\frac{\mathbf{r}'B\mathbf{r}}{\mathbf{r}'W\mathbf{r}} = \frac{\mathbf{r}'P\Sigma^2P'\mathbf{r}}{\mathbf{r}'P(I_n - \Sigma^2)P'\mathbf{r}} = \frac{\mathbf{b}'\Sigma^2\mathbf{b}}{\mathbf{b}'(I_p - \Sigma^2)\mathbf{b}}, \quad \mathbf{b} = P'\mathbf{r}. \tag{1.58}$$

Es ist

$$\frac{\mathbf{b}'\Sigma^2\mathbf{b}}{\mathbf{b}'(I_p - \Sigma^2)\mathbf{b}} = \frac{\sum_i \lambda_i^2 b_i^2}{\sum_i (1-\lambda_i^2)b_i^2}, \tag{1.59}$$

und dieser Quotient wird maximal, wenn man nur $b_1 \neq 0$ berücksichtigt (vergl. auch Abschnitt 5.1). Abbildung 6 zeigt die Diskriminanzanalyse (LDA) der in

Abbildung 6: LDA und PCA des Fluglinienpersonals



Beispiel 1.5 vorgestellten Daten. Die erste kanonische Variable erklärt 79.33% der Variation, die zweite dementsprechend 20.87% der Variation. Der Vergleich mit Abbildung 5 zeigt, dass die erste kanonische Variable wieder in erster Linie zwischen den "Operational controls" und den übrigen Angestellten unterscheided. Die Skalen der kanonischen Variablen unterscheiden sich aber auf Grund der Standardisierung von den üblichen Skalen, die der Abbildung 5 unterliegen.

In Abschnitt 1.11 wird eine Beziehung zwischen LDA und der Hauptachsentransformation (Principal Component Analysis, PCA) diskutiert. In Abb. 6 wird gleichwohl schon das Ergebnis einer PCA der Fluglinendaten gezeigt, um einen unmittelbaren Verglich von LDA und PCA zu ermöglichen: die Ergebnisse sind

nicht identisch, aber doch insofern äquivalent, als die PCA eine Trennung der Gruppe der "operational controls" und den übrigen Angestelltengruppen nahelegt. Diese Trennung mittels PCA gelingt hier, weil die Unterschiede zwischen den "operational controls" und den übrigen Gruppen eine Richtung mit maximaler Varianz erzeugt. Im Allgemeinen vermag die PCA aber nicht verschiedene Gruppen so optimal zu trennen wie die LDA. Die Tabelle 12 zeigt die Korrelationen zwischen den Prädiktoren sowie die zugehörigen Eigenwerte. Während die Korrelation zwischen den Prädiktoren "outdoor" und "convivial" vernachlässigbar ist, ist die Korrelation zwischen diese Prädiktoren auf der einen Seite und dem Prädiktor "conservative" auf der anderen Seite jeweils negativ und betragsmäßig zwischen "convivial" und "conservative" am höchsten. Diese Werte legen nahe, dass sich Angestellte, die hohe Werte auf dem "conservative"-Prädiktor haben, sich von denjenigen Angestellten, die eher größere Werte bei den Prädiktoren "outdoor" und/oder "convivial" haben, unterscheiden. Die Angestellten, die eher höhere Werte auf dem "conservative"-Prädiktor haben, scheinen in erster Linie zur Gruppe der "operational controls" zu gehören, so dass man erwarten kann, dass die PCA diese Gruppe von den übrigen Gruppen trennt.

Tabelle 12: Korrelationen zwischen den Prädiktoren und zugehörige Eigenwerte

	outdoor	convivial	conservative
outdoor	1.000	.079	-.121
convivial	.079	1.000	-.418
conservative	-.121	-.418	1.000
Latente Variable	Dim. 1	Dim. 2	Dim. 3
Eigenwerte	52.630	34.514	20.859

Die erste Hauptachse erklärt 41%, und die zweite Hauptachse erklärt 33% der Gesamtvarianz. Die "operational controls" werden ungefähr gleich gut von den übrigen Gruppen getrennt wie bei der LDA, auch wenn sich die Konfigurationen leicht voneinander unterscheiden. Diese Vorraussage entspricht sowohl den Ergebnissen der LDA wie auch der PCA, wobei es die erste Dimension bzw Hauptachse ist, in Bezug auf die sich die Gruppen unterscheiden. Die Gewichte der Prädiktoren – sie entsprechen den Ladungen bei der PCA – korrespondieren bei der PCA zu diesem Befund: Für die erste Hauptachse ist die Ladung für "conservative" ist wohlsepariert von denn Ladungen für die beiden anderen Prädiktoren.

1.8 Diskriminanzanalyse und Kanonische Korrelation

Die Beziehung zwischen diesen beiden Verfahren wurde zum ersten Mal von Bartlett (1938) hergestellt.

1.8.1 Kanonische Korrelation

Bei der Kanonischen Korrelation (CCA – Canonical Correlation Analysis) sind zwei Datensätze X und Y gegeben, und man versucht, die latenten Variablen von Y aufgrund der latenten Variablen von X in bestmöglicher Weise vorauszusagen; das Verfahren kann als Verallgemeinerung der multiplen Regression verstanden werden. X bestehe aus p Spaltenvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_p$, und Y aus q Spaltenvektoren $\mathbf{y}_1, \ldots, \mathbf{y}_q$, Es sollen zwei Vektoren $\mathbf{u}$ und $\mathbf{v}$ bestimmt werden derart, dass

$$\begin{aligned} \mathbf{u} &= a_1\mathbf{x}_1 + \cdots + a_p\mathbf{x}_p & (1.60)\\ \mathbf{v} &= b_1\mathbf{y}_1 + \cdots + b_q\mathbf{y}_q, & (1.61) \end{aligned}$$

und $\mathbf{u}'\mathbf{v} = \max$ gilt; bei geeigneter Normalisierung kann $\mathbf{u}'\mathbf{v}$ als Korrelation R_{uv} angesehen werden. Setzt man $\mathbf{a} = (a_1, \ldots, a_p)'$ und $\mathbf{b} = (b_1, \ldots, b_q)'$, so kann man diese Gleichungen auch in der Form

$$\begin{aligned} \mathbf{u} &= X\mathbf{a} & (1.62)\\ \mathbf{v} &= Y\mathbf{b} & (1.63) \end{aligned}$$

schreiben. Dann ist

$$R_{uv} = \mathbf{u}'\mathbf{v} = \mathbf{a}'X'Y\mathbf{b} = \mathbf{a}'R_{xy}\mathbf{b}. \quad (1.64)$$

Wegen $(X'Y)' = Y'X = R_{yx}$ folgt, dass $R'_{xy} = R_{yx}$. R_{uv} hängt von den Vektoren $\mathbf{a}$ und $\mathbf{b}$ ab. Die Maximierung von R_{uv} ist unbestimmt, wenn keine Neben- oder Randbedingungen gesetzt werden, weshalb die Randbedingungen $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$ eingeführt werden. Sie stellen keine Beschränkung der Allgemeinheit dar, sondern nur eine Skalierung. Wegen $\mathbf{u} = X\mathbf{a}$ und $\mathbf{v} = Y\mathbf{b}$ hat man $\|\mathbf{b}\|^2 = \mathbf{a}'X'X\mathbf{a}$ und $\|\mathbf{v}\|^2 = \mathbf{b}'Y'Y\mathbf{b}$. Für die Maximierung unter Nebenbedingungen hat man die Lagrangesche Multiplikatorenregel: man betrachtet

$$Q(\mathbf{a}, \mathbf{b}) = \mathbf{a}'X'Y\mathbf{b} - \lambda(\mathbf{a}'X'X\mathbf{a} - 1) - \mu(\mathbf{b}'Y'Y\mathbf{b} - 1), \quad (1.65)$$

wobei λ und μ die Lagrange-Faktoren sind. Die partiellen Ableitungen nach $\mathbf{a}$ und $\mathbf{b}$ werden gleich Null gesetzt:

$$\begin{aligned} \frac{\partial Q}{\partial \mathbf{a}} &= R_{xy}\mathbf{b} - \lambda R_{xx}\mathbf{a} = 0 & (1.66)\\ \frac{\partial Q}{\partial \mathbf{b}} &= R_{yx}\mathbf{b} - \mu R_{yy}\mathbf{b} = 0 & (1.67) \end{aligned}$$

Multipliziert man die erste Gleichung mit $\mathbf{a}$ und die zweite mit $\mathbf{b}$, so erhält man

$$\begin{aligned} \mathbf{a}'R_{xy}\mathbf{b} - \lambda\mathbf{a}'R_{xx}\mathbf{a} &= 0 & (1.68)\\ \mathbf{b}'R_{yx}\mathbf{b} - \mu\mathbf{b}'R_{yy}\mathbf{b} &= 0 & (1.69) \end{aligned}$$

Es ist aber $R_{uv} = \mathbf{a}'R_{xy}\mathbf{b} = \mathbf{b}'R_{yx}\mathbf{a}$ und Wegen der Normierung $\|\mathbf{u}\| = \|\mathbf{v}\| = 1$ muß $\mathbf{a}'R_{xx}\mathbf{a} = \mathbf{b}'R_{yy}\mathbf{b} = 1$ gelten. Daraus folgt

$$R_{uv} = \lambda = \mu, \quad (1.70)$$

Für $\mathbf{a}$ und $\mathbf{b}$ erhält man die Lösungen

$$\lambda^2 \mathbf{a} = R_{xx}^{-1} R_{xy} R_{yy}^{-1} R_{yx} \mathbf{a}. \quad (1.71)$$

$$\lambda^2 \mathbf{b} = R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy} \mathbf{b}. \quad (1.72)$$

$\mathbf{a}$ und $\mathbf{b}$ sind also Eigenvektoren von $R_{xx}^{-1} R_{xy} R_{yy}^{-1} R_{yx}$ bzw. $R_{yy}^{-1} R_{yx} R_{xx}^{-1} R_{xy}$ mit dem Eigenwert λ^2. Die Herleitung dieser Gleichung wird hier übergangen. Im Allgemeinen wird es mehr als ein Paar von Eigenvektoren geben, und es läßt sich zeigen, dass verschiedene Eigenvektoren $\mathbf{a}_1, \ldots, \mathbf{a}_r$ paarweise orthogonal sind; ebenso sind die Eigenvektoren $\mathbf{b}_1, \ldots, \mathbf{b}_r$ paarweise orthogonal, und $r \leq \min(p, q)$. Für jedes Paar $(\mathbf{a}_j, \mathbf{b}_j)$ existiert ein Eigenwert $\lambda_j^2 = R_j^2$, wobei R_j^2 eine abkürzende Schreibweise für $R_{u_j v_j}^2$ sein soll: jedes Paar $(\mathbf{a}_j, \mathbf{b}_j)$ definiert ja ein Paar von Skalen $\mathbf{u}_j$ und $\mathbf{v}_j$. Diese Skalen kann man als latente Variablen für den Datgensatz X bzw. Y ansehen, deren Orientierung so gewählt wird, dass $\mathbf{u}_j$ und $\mathbf{v}_j$ jeweils maximal miteinander korrelieren. Im Allgemeinen sind sie nicht identisch mit den Hauptachsen des zu $X'X$ bzw $Y'Y$ korrespondierenden Ellipsoids.

1.8.2 Beziehung der CCA zur LDA

Man betrachte die Matrix G in (1.73), Seite 30. G ist eine *Indikatormatrix*: die Spalten repräsentieren die verschiedenen Gruppen, und eine 1 zeigt an, in welche Gruppe eine Person gehört. Personen, die zu einer Gruppe gehören, sind zusammengefasst worden: die horizontalen Geraden trennen die Meßwerte der verschiedenen Gruppen. Korrespondierend zu den ersten n_1 Zeilen der ersten Gruppe in der Matrix X enthalten die ersten n_1 Zeilen von G den K-dimensionalen Einheitsvektor $(1, 0, \cdots, 0)'$; die folgenden n_2, zur zweiten Gruppe korrespondierenden Zeilen enthalten die 1 an der zweiten Stelle (d.h. in der zweiten Spalte), etc. Allgemein zeigt für eine gegebene Zeile von G die 1 in der k-ten Spalte an, daß das Objekt oder die Person in der entsprechenden Zeile von X zur k-ten Gruppe gehört. Die Elemente der Matrix G sind "Meßwerte", die einfach nur die Zugehörigkeit zu einer der K Gruppen anzeigen, (d.h. sie sind "Dummy"-

Variablen).

$$G = \left(\begin{array}{ccccc} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ \hline 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & \cdots & \vdots \\ \hline 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{array}\right) \tag{1.73}$$

Man kann G knapp in der folgenden Weise definieren:

$$G = (g_{ik}), \quad g_{ik} = \begin{cases} 1, & \text{Fall } i \text{ ist in Klasse } k, \\ 0, & \text{Fall } i \text{ ist nicht in Klasse } k, \end{cases} \tag{1.74}$$

$1 \le i \le N,\ 1 \le k \le K,\ N = \sum_j n_j$.

Berechnet man nun für X und Y kanonischen Variablen, so bestimmt man Variablen, die eine optimale Vorhersage der Gruppenzugehörigkeit erlauben. Damit liegt nahe, daß die Diskriminanzanalyse und die Kanonische Korrelation äquivalente Verfahren sind, wenn man $Y = G$ setzt. Dieses Argument soll etwas expliziter ausgeführt werden, damit auch die Beziehung zwischen dem Diskriminanzkriterium λ und der Kanonischen Korrelation R deutlich wird. Zur Vereinfachung der Schreibweise soll dabei wieder auf die "zentrierte" Matrix $x = X - 1\bar{x}'$ übergegangen werden.

Es sei

$$M = \begin{pmatrix} \bar{x}_{11} & \bar{x}_{12} & \cdots & \bar{x}_{1p} \\ \bar{x}_{21} & \bar{x}_{22} & \cdots & \bar{x}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_{K1} & \bar{x}_{K2} & \cdots & \bar{x}_{Kp} \end{pmatrix} \tag{1.75}$$

die Matrix der Mittelwerte. Weiter gilt

$$G'G = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_K \end{pmatrix}, \quad (G'G)^{-1} = \begin{pmatrix} 1/n_1 & 0 & \cdots & 0 \\ 0 & 1/n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/n_K \end{pmatrix}, \tag{1.76}$$

$n_1, \ldots, n_K$ die Anzahl der Messwerte in den verschiedenen Gruppen. Die in (1.75) eingeführte Matrix M der Mittelwerte läßt sich dann wie folgt darstellen: die

Matrix $G'X$ ist offenbar eine $(K \times p)$-Matrix, deren k-te Zeile und j-te Spalte gerade die Summe $\sum_{i=1}^{n_k} x_{ijk}$ enthält. Also ist

$$G'X = (G'G)M, \quad X'G = M'(G'G), \tag{1.77}$$

und damit

$$M = (G'G)^{-1}G'X. \tag{1.78}$$

Die Gleichung für die Kanonische Korrelation ist nach (1.71)

$$(\mathbf{x}'\mathbf{x})^{-1}(\mathbf{x}'G)(G'G)^{-1}(G'\mathbf{x})\mathbf{u} = \lambda\mathbf{u} \tag{1.79}$$

wobei G für y substituiert wurde. Offenbar gilt (vergl. (1.29), Seite 11)

$$B = (\mathbf{x}'G)(G'G)^{-1}(G'\mathbf{x}) \tag{1.80}$$

Es sei $\underline{1} = (1, 1, \cdots, 1)'$ ein Vektor mit $n = \sum_k n_k$ Komponenten, die alle gleich 1 sind. Dann ist $\underline{1}\bar{x}'$ eine $(n \times p)$-Matrix, deren j-te Spalte den Mittelwert $\bar{x}_j$ (also den Mittelwert *aller* n Werte der j-ten Variablen) enthält. Man rechnet leicht nach, daß die Definitionen von W und B den Definitionen

$$W = (X - GM)'(X - GM), \quad B = (GM - 1\bar{\mathbf{x}}')'(GM - 1\bar{\mathbf{x}}') \tag{1.81}$$

entsprechen. Mit

$$T = (X - \underline{1}\bar{\mathbf{x}}')'(X - \underline{1}\bar{\mathbf{x}}') \tag{1.82}$$

gilt[4]

$$T = W + B, \quad QS_{ges} = \mathbf{u}'T\mathbf{u} = \mathbf{u}'(W + B)\mathbf{u} \tag{1.83}$$

Es sei $x = X - \underline{1}\bar{\mathbf{x}}'$; die Matrix x heißt *zentriert*. Für den Fall zentrierter Daten vereinfachen sich diese Ausdrücke etwas. Da $\bar{\mathbf{x}} = \vec{0}$ ergeben sich die Ausdrücke für den zentrierten Fall, indem man $\underline{1}\bar{\mathbf{x}}'$ gleich Null setzt bzw. einfach fortläßt. Dementsprechend erhält man aus (1.81)

$$W = (\mathbf{x} - GM)'(\mathbf{x} - GM), \quad B = (GM)'(GM) = M'G'GM \tag{1.84}$$

[4]Die Gültigkeit von $T = W + B$ läßt sich in Matrixschreibweise leicht zeigen:

$$T = (x - 1\bar{x}')'(x - 1\bar{x}') = (x - GM + GM - 1\bar{x}')'(x - GM + GM - 1\bar{x}')$$

Ausmultipliziert ergibt sich

$$T = (X-GM)'(X-GM)+(GM-1\bar{x})'(GM-1\bar{x}')+(X-GM)'(GM-1\bar{x}')+(GM-1\bar{x}')'(X-GM)$$

Es ist aber

$$(x - GM)'(GM - 1\bar{x}') = x'GM - M'G'GM - x'1\bar{x}' - M'G'1\bar{x}'$$

Aus $x'G = M'G'G$ (vergl. (1.78)) folgt durch Multiplikation mit G^{-1} von rechts $x' = M'G'$, so daß $x'GM - M'G'GM = M'G'GM - M'G'GM = 0$, und ebenso $x'1\bar{x}' - M'G'1\bar{x}' = M'G'1\bar{x}' - M'G'1\bar{x}' = 0$. Der letzte Term in der Zerlegung für T ist dann ebenfalls gleich Null, da $(x - 1\bar{x}')'(x - GM) = ((x - GM)'(x - 1\bar{x}))'$; so daß tatsächlich $T = W + B$.

wobei M natürlich anhand von $\mathbf{x}$, nicht von X berechnet wird. (1.82) liefert

$$T = \mathbf{x}'\mathbf{x} \tag{1.85}$$

und (1.83) gilt natürlich nach wie vor. Nach (1.78) ist aber $\mathbf{x}'\mathbf{y} = x'G = M'(G'G)$ und $\mathbf{y}'\mathbf{x} = G'\mathbf{x} = (G'G)M$ und natürlich $(\mathbf{y}'\mathbf{y})^{-1} = (G'G)^{-1}$. In (**??**) eingesetzt ergibt sich

$$T^{-1}M'(G'G)(G'G)^{-1}(G'G)M\,\mathbf{a} = T^{-1}M'G'GM\,\mathbf{a} = R^2\mathbf{a}$$

d.h. aber

$$T^{-1}B\,\mathbf{a} = R^2\mathbf{a} \tag{1.86}$$

Wegen $T = W + B$ folgt hieraus

$$B\,\mathbf{a} = R^2T\,\mathbf{a} = R^2(W + B)\mathbf{a} = R^2W\,\mathbf{a} + R^2B\,\mathbf{a}$$

so daß

$$B\,\mathbf{a} - R^2B\,\mathbf{a} = R^2W\,\mathbf{a}$$

Multiplikation von links mit $\mathbf{a}'$ ergibt dann

$$\mathbf{a}'B\,\mathbf{a} - R^2\mathbf{a}'B\,\mathbf{a} = \mathbf{a}'B\,\mathbf{a}(1 - R^2) = R^2\mathbf{a}'W\,\mathbf{a}$$

(man beachte, daß $\mathbf{a}'B\,\mathbf{a}$ und $\mathbf{a}'W\,\mathbf{a}$ Skalare sind). Dividiert man nun einerseits durch $\mathbf{a}'W\,\mathbf{a}$ und andererseits durch $1 - R^2$, so erhält man

$$\frac{\mathbf{a}'B\,\mathbf{a}}{\mathbf{a}'W\,\mathbf{a}} = \lambda = \frac{R^2}{1 - R^2}, \quad \text{d.h. } R^2 = \frac{\lambda}{1 + \lambda} \tag{1.87}$$

wobei λ natürlich das Diskriminanzkriterium ist, vergl. (1.26), p. 10; dies ist Gleichung (1.88). Der Vergleich mit (1.26) zeigt weiter, daß sich der Gewichtevektor $\mathbf{a} = \mathbf{u}$ als Eigenvektor der Matrix $(\mathbf{x}'\mathbf{x})^{-1}(\mathbf{x}'G)(G'G)^{-1}(G'\mathbf{x})$ ergibt. Das Ergebnis wird im folgenden Satz zusamengefasst:

Satz 1.5 *Zwischen den Kanonischen Korrelation R und dem Diskriminanzkriterium λ besteht die als* Roysches Kriterium *bekannte Beziehung*

$$\frac{\boldsymbol{u}'B\,\boldsymbol{u}}{\boldsymbol{u}'W\,\boldsymbol{u}} = \lambda = \frac{R^2}{1 - R^2}, \quad \textit{d.h. } R^2 = \frac{\lambda}{1 + \lambda} \tag{1.88}$$

Der Gewichtevektor $\boldsymbol{u}$ ergibt sich als Eigenvektor der Matrix $(\boldsymbol{x}'\boldsymbol{x})^{-1}(\boldsymbol{x}'G)(G'G)^{-1}(G'\boldsymbol{x})$, d.h. es gilt

$$(\boldsymbol{x}'\boldsymbol{x})^{-1}(\boldsymbol{x}'G)(G'G)^{-1}(G'\boldsymbol{x})\boldsymbol{u} = \lambda\boldsymbol{u} \tag{1.89}$$

wobei G für $\boldsymbol{y}$ substituiert wurde.

λ ist monoton wachsend mit R^2: Für $R^2 \to 1$ folgt $\lambda \to \infty$, und für $R^2 \to 0$ folgt $\lambda \to 0$, d.h. λ ist eine monoton wachsende Funktion von R^2. $\lambda \to \infty$ bedeutet, dass QS_{inn} beliebig klein im Vergleich zu QS_{zw} wird, d.h. die Fehlervarianz wird beliebig klein im Vergleich zu den systematischen Unterschieden zwischen den Klassen.

Anteile erklärter Varianz: Aus der Regressionsrechnung ist bekannt, dass das Quadrat eines Korrelationskoeffizienten dem Anteil der durch den Prädiktor erklärten Varianz entspricht. Dies gilt auch für R^2. Aus der Beziehung (1.86), d.h. $T^{-1}B\mathbf{a} = R^2\mathbf{a}$, folgt durch Multiplikation von links mit T

$$B\,\mathbf{a} = R^2 T\,\mathbf{a}$$

Multipliziert man noch einmal von links mit $\mathbf{a}'$, so erhält man

$$\mathbf{a}'B\,\mathbf{a} = R^2\mathbf{a}'T\,\mathbf{a}.$$

Daraus folgt ein Ausdruck für R^2, der zusammen mit dem für λ vorgestellt wird, um den Unterschied zu verdeutlichen:

$$R^2 = \frac{\mathbf{a}'B\,\mathbf{a}}{\mathbf{a}'T\,\mathbf{a}} = \frac{QS_{zw}}{QS_{ges}}, \quad \lambda = \frac{QS_{zw}}{QS_{inn}} \tag{1.90}$$

Diese Beziehung für R^2 entspricht der für Korrelationen bereits bekannten Beziehung $r^2 = QS_{zw}/QS_{ges}$. (Zur Erinnerung: $\lambda = QS_{zw}/QS_{inn}$.) R^2 entspricht also einem Determinationskoeffizienten. R^2 bezieht sich auf $\mathbf{u}$, und da es mehr als eine kanonische Variable $\mathbf{u}$ geben kann, gibt R^2 den Varianzanteil an, der an der Varianz zwischen den Gruppen durch $\mathbf{u}$ erklärt wird. Die Interpretation überträgt sich auf λ, da λ eine monotone Funktion von R^2 ist. So sei λ_j das Diskiminanzkriterium für die Skala $\mathbf{u}_j$. Dann entspricht

$$q_j = \frac{\lambda_j}{\sum_{k=1}^{r}\lambda_k} \tag{1.91}$$

dem Anteil, der durch die j-te Skala an der Varianz zwischen den Gruppen erklärt wird, wobei r die Anzahl der betrachteten Skalen (kanonischen Variablen) ist.

1.9 Zur Anzahl der kanonischen Variablen

Es gibt so viele kanonische Variablen (Diskriminanten), wie es von Null verschiedene Eigenwerte der $(p \times p)$-Matrix $W^{-1}B$ gibt. Es sei also $s \leq p$ die Anzahl der $\lambda_k > 0$. Für jede Klasse Ω_k, $k = 1, 2, \ldots, K$ existiert ein Vektor $\bar{\mathbf{y}}_k$, dessen Komponenten die Mittelwerte über die Variablen für die k-te Klasse sind. Weiter sei

$$\bar{\mathbf{y}} = \frac{1}{g}\sum_{k=1}^{g}\bar{\mathbf{y}}_k. \tag{1.92}$$

Es werden nun die g Vektoren

$$\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}, \bar{\mathbf{y}}_2 - \bar{\mathbf{y}}, \ldots, \bar{\mathbf{y}}_K - \bar{\mathbf{y}} \tag{1.93}$$

betrachtet. Dann folgt

$$(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}) + (\bar{\mathbf{y}}_2 - \bar{\mathbf{y}}) + \cdots + (\bar{\mathbf{y}}_K - \bar{\mathbf{y}}) = K\bar{\mathbf{y}} - K\bar{\mathbf{y}} = \bar{0}.$$

Also gilt z.B.

$$(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}) = -(\bar{\mathbf{y}}_2 - \bar{\mathbf{y}}) - \cdots - (\bar{\mathbf{y}}_K - \bar{\mathbf{y}}),$$

d.h. $(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}})$ kann als Linearkombination der restlichen Differenzen dargestellt werden. Linearkombinationen der Vektoren (1.93) definieren Hyperebenen der Dimension $q \leq K-1$. Es sei $\mathbf{v}$ ein Vektor, der senkrecht auf jedem Vektor $\bar{\mathbf{y}}_k - \bar{\mathbf{y}}$ und damit auf den Hyperebenen steht. Dann hat man

$$B\mathbf{v} = \sum_{k=1}^{K} (\bar{\mathbf{y}}_k - \bar{\mathbf{y}})(\bar{\mathbf{y}}_k - \bar{\mathbf{y}})'\mathbf{v} = \vec{0}, \tag{1.94}$$

denn $(\bar{\mathbf{y}}_k - \bar{\mathbf{y}})'\mathbf{v} = 0$ nach Voraussetzung. Daraus folgt

$$W^{-1}B\mathbf{v} = \mathbf{v}, \tag{1.95}$$

Es gibt $p - q$ orthogonale Eigenvektoren, die zum Eigenwert 0 korrespondieren. Also gilt für die Anzahl s der Eigenwerte ungleich Null $s \leq \min(p, K-1)$. Für die Maximalzahl der zu betrachtenden Diskriminanten ergibt sich demnach die folgende Übersicht:

Tabelle 13: Maximalzahl zu betrachtender Diskriminanten

Anzahl d. Variablen	Anzahl der Klassen	Maximalzahl der Diskriminanten
Beliebiges p	$K = 2$	1
Beliebiges p	$K = 3$	2
$p = 2$	Beliebiges K	2

1.10 Kreuzvalidierung und Inferenz

Kreuzvalidierung Kann die multivariate Normalverteilung mit homogenen Varianzen angenommen werden, so legen die Definitionen von R^2 und λ nahe, dazu korrespondierende F-Tests zu entwickeln. Diese Annahmen sind aber selten gerechtfertigt. Dann stellt sich die Frage, wie getestet werden kann, ob die Ergebnisse der Diskriminanzanalyse für die Anwendung auf Klassifikationsfragen geeignet sind oder nicht.

Wie bei der Regression bietet sich die Kreuzvalidierung an. Auf der Basis der anhand der Trainingsstichprobe geschätzten $\mathbf{u}_i$ werden neue Fälle den Klassen zugeordnet. Dabei wird es zu Fehlklassifikationen kommen, weil die zu den einzelnen Klassen korrespondierenden Punktekonfigurationen sich oft mehr oder weniger überlappen und auch bei optimaler Schätzung der $\mathbf{u}_i$ Fehler unvermeidbar sind. Man kann eine neue Stichprobe von Fällen für die Kreuzvalidierung erheben, – nur ist dies häufig schon auspraktischen Gründen kaum möglich. Man kann nur einen Teil der Stichprobe für die Schätzung verwenden und den Rest für die Validierung. Aber die Schätzungen werden um so besser, je größer die Trainingsstichprobe ist, so dass man alle beobachteten Fälle in der Trainingsstichprobe zusammenfassen wird. So kommt es, dass im Allgemeinen die *Leave-one-out*- oder *Jackknife*-Validierung verwendet wird. Da wird ein Fall (ein Objekt, eine Person, etc) aus der Trainingsstichprobe herausgenommen, die $\mathbf{u}_i$ werden anhand der Reststichprobe geschätzt und es wird eine Klassifikation des herausgenommenen Falls vorgenommen. Dieses Vorgehen wird der Reihe nach für alle Fälle der Trainingsstichprobe durchgeführt. Die Ergebnisse werden in einer Konfusionstabelle zusammengefasst, deren Zeilen für die vorausgesagten Klassen und deren Spalten für die wahren Klassen stehen. Das Element n_{ij} in der i-ten Zeile und j-ten Spalte gibt an, wie häufig die i-te und die j-te Klasse miteinander verwechselt wurden. Im Zweifel hilft dann ein chi^2-Test, zu entscheiden, ob die n_{ij} der Nullhypothese H_0 entsprechen, derzufolge die Klassifikationen zufällig oder nicht getroffen wurden. Zusammen mit den Anteilen $q_j = \lambda_i / \sum_k \lambda_k$ ergibt sich dann ein Bild über die Güte der Klassifikationsleistung; in den folgenden Beispielen wird dieses Verfahren vorgestellt.

Statistische Tests Sind das Kriterium λ und die Gewichte $\mathbf{u}$ gegeben, so ist es von Interesse, zu entscheiden, ob alle oder nur einige der Variablen x_i diskriminatorische Relevanz haben. Weiter wird man an einer Schätzung der Fehlerrate für die gewählte Entscheidungsregel interessiert sein. Es müssen die folgenden Annahmen gemacht werden:

1. Die Variablen sind in den verschiedenen Gruppen normalverteilt,
2. Für die Varianz-Kovarianzmatrizen gilt

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K, \tag{1.96}$$

   d.h. es muß gefordert werden, daß die Varianzen und Kovarianzen zwischen den Variablen in den verschiedenen Gruppen gleich sind.

Es ergeben sich zwei Deutungen:

1. Generell kann man die Menge der $\mathbf{y}$ betrachten, für die

$$\|\mathbf{y} - \bar{\mathbf{y}}_K\|^2 = \sum_{j=1}^{s} (y_j - \bar{y}_{jk})^2 = \text{konstant}$$

gilt. Offenbar liegen die Endpunkte all dieser Vektoren auf einer Hyperkugel. Betrachtet man die zu den Y korrespondierende Menge der $\mathbf{x}$, für die die Mahalanobis-Distanzen $(\mathbf{x} - \bar{\mathbf{x}}_k)' W^{-1} (\mathbf{x} - \bar{\mathbf{x}}_k)$ konstant sind, so liegen die Endpunkte der $\mathbf{x}$ auf einem Ellipsoid.

2. Nimmt man die multivariate Normalverteilung an so kann man die Mahalanobis-Distanz als Ort gleicher Wahrscheinlichkeit deuten: alle Punkte, die nach der multivariaten Normalverteilung gleiche Wahrscheinlichkeit haben, liegen auf einem Ellipsoid. Ein Ellipsoid entsteht im Übrigen, wenn die Hauptachsen, die ja als latente Dimensionen interpretiert werden können, unterschiedlich große Varianzanteile haben; sind diese Anteile gleich, so definiert die Mahalanobis-Distanz eine Menge von Hyperkugeln.

3. Die Beziehung (1.37) gilt andererseits unabhängig von der Annahme der Normalverteilung, denn sie besagt ja nur, daß $\|\bar{\mathbf{y}} - \bar{\mathbf{y}}_k\|^2$ gleich der Mahalanobis-Distanz des durch $\mathbf{x}$ definierten Punktes von $\bar{\mathbf{y}}_k$ ist. Nimmt man diese Verteilung *nicht* an, so kann man der Mahalanobis-Distanz auch eine andere Deutung geben. Durch eine geeignete Koordinatentransformation kann man die Endpunkte der $\mathbf{x}$ auch durch die Projektionen auf die Hauptachsen dieses Ellipsoids definieren; die Hauptachsen korrespondieren zu den latenten Dimensionen, die man etwa in der Faktorenanalyse betrachtet. Man kann dann sagen, daß die ellipsoide Punktekonfiguration durch unterschiedliche Gewichtung der Koordinatenachsen entsteht; im 2-dimensionalen Fall hat ein Punkt dann die Koordinaten (x_1, x_2), die der Gleichung $x_1^2/a^2 + x_2^2/b^2 = k$ eine Konstante genügen, wbei $a \neq b$. Für $a = b$, also gleicher Gewichtung, liegen alle Endpunkte der $\mathbf{x}$ auf einer Hyperkugel. a und b reflektieren die Ausmaße, mit denen die latenten Variablen in die Messung der x_1, x_2 eingehen.

4. Die vorangegangene Deutung ist mit der Annahme der multivariaten Normalverteilung kompatibel; a^2 und b^2 entsprechen dann den Varianzen der beiden Meßgrößen. Die Länge der Hauptachse ist proportional zu a, d.h. zur Streuung σ; die unterschiedlichen Gewichtungen lassen sich dann durch unterschiedliche Streuungen, und die unterschiedlichen Streuungen lassen sich durch unterschiedliche Gewichtungen interpretieren; welche Implikationsrichtung man wählt, hängt vom theoretischen Ansatz ab, von dem man bei der Interpretation ausgeht.

Diskriminanz: Mittelwertsunterschiede: Da $\lambda = QS_{zw}/QS_{ges}$ gilt (und die Mittelwerte der Gruppen so bestimmt werden, daß λ maximal ist), liegt es nahe, die aus der Varianzanalyse bekannten Statistiken bzw. Prüfgrößen zu verwenden. Zunächst einmal läßt sich auf diese Weise testen, ob die Klassenmittelwerte sich tatsächlich signifikant voneinander unterscheiden. Unterscheiden sie sich nicht, so läßt sich sagen, daß trotz der Maximierung von QS_{zw} relativ zu QS_{ges} keine

Diskriminierung der Gruppenmitglieder anhand der Meßwerte x_i möglich ist. Dementsprechend hat man

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_K, \tag{1.97}$$

$$H_1: \quad \mu_i \neq \mu_j, \text{ für mindestens ein Paar } (i,j) \text{ mit } i \neq j \tag{1.98}$$

In der einfachen Varianzanalyse hat man den bekannten Test

$$F = \frac{QS_{zw}/(K-1)}{QS_{ges}/K(j-1)}, \quad df = K-1, K(j-1)$$

Für die Diskriminanzanalyse hat man den entsprechenden Test für die multivariate Varianzanalyse

$$\Lambda = \frac{|W|}{|B+W|} = |I + W^{-1}B|^{-1}, \tag{1.99}$$

Wilk's Λ; unter H_0 gilt

$$\Lambda \sim \Lambda(q, N-K, K-1) \tag{1.100}$$

(Λ-Verteilung von Wilks).

Schätzung der Fehlerraten: Es der Fall zweier Gruppen betrachtet. Die Gesamtfehlerrate ist durch

$$\epsilon = p(\Omega_1)\epsilon_{12} + p(\Omega_2)\epsilon_{21} \tag{1.101}$$

gegeben. ϵ_{12} und ϵ_{21} sind die individuellen Fehlerraten; Zur Vereinfachung werde für die a-priori-Wahrscheinlichkeiten $P(\Omega_1) = \pi_1$ und $p(\Omega_2) = \pi_2$ gesetzt:

$$\epsilon_{12} = \Phi\left(\frac{\log(\pi_1/\pi_2) - \delta^2/2}{\delta}\right) \tag{1.102}$$

$$\epsilon_{21} = \Phi\left(-\frac{\log(\pi_2/\pi_1) + \delta^2/2}{\delta}\right), \tag{1.103}$$

wobei

$$\delta = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \tag{1.104}$$

die Mahalanobis-Distanz ist. (Φ bedeutet die Verteilungsfunktion der Gauss-Verteilung.)

Für die ML-Regel ergeben sich die Fehlerraten gemäß

$$\epsilon_{12} = \epsilon_{21} = \Phi\left(-\frac{\delta}{2}\right). \tag{1.105}$$

Die tatsächlichen Fehlerraten ergeben sich, wenn man zur geschätzten Diskriminanzfunktion $\hat{d}$ mit der geschätzten Kovarianzmatrix $S = \hat{\Sigma}$ übergeht:

$$\hat{d}(x) = (x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2))'S^{-1}((\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2) \tag{1.106}$$

übergeht.

Eine sogenannte *plug-in*-Schätzung erhält man, wenn man für μ_1, μ_2 und Σ die Schätzungen $\bar{x}_1$, $\bar{x}_2$ und S einsetzt:

$$\hat{d}(\bar{x}_1) = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2) \quad (1.107)$$

$$\hat{d}(\bar{x}_2) = -(\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2). \quad (1.108)$$

Dann ist für die Bayes-Regel

$$\hat{\epsilon} = \pi_1 \hat{\epsilon}_{12} + \pi_2 \epsilon_{21} \quad (1.109)$$

mit

$$\hat{\epsilon}_{12} = \Phi\left(\frac{\log(\pi_2/\pi_1) - D^2/2}{D}\right), \quad \hat{\epsilon}_{21} = \Phi\left(\frac{-\log(\pi_2/\pi_1) - D^2/2}{D}\right) \quad (1.110)$$

mit $D^2 = (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$. Für die ML-Regel gilt

$$\hat{\epsilon}_{12} = \hat{\epsilon}_{21} = \Phi(-D/2). \quad (1.111)$$

1.11 PCA und LDA

Es sei X eine $(n \times p)$-Matrix; sowohl die LDA wie auch die PCA (Principal Components Analysis – Hauptachsentransformation) repräsentieren die ursprünglichen p Variablen (Prädiktoren im Falle der LDA) in einem Raum mit einer Dimension $r < p$. Bei der PCA werden orthogonale Basisvektoren gesucht, aus denen sich die die Variablen repräsentierenden Spaltenvektoren als Linearkombination ergeben, d.h. es wird eine Matrix P gesucht derart, dass $XP = L$ und die Spaltenvektoren $\mathbf{L}_1, \mathbf{L}_2, \ldots, \mathbf{L}_p$ der $(n \times p)$-Matrix L sind die gesuchten orthogonalen Basisvektoren. Die Komponenten der $\mathbf{L}_j$ sind die Koordinaten der "Fälle" (z.B. Personen) auf den Achsen, die durch die $\mathbf{L}_j$ definiert werden. Dabei soll die Varianz der Koordinaten auf $\mathbf{L}_1$ maximal sein, die der Koordinaten auf $\mathbf{L}_2$ soll zweitmaximal sein, etc. Wegen der geforderten Orthogonalität der $\mathbf{L}_j$ hat man

$$L'L = P'X'XP = \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p). \quad (1.112)$$

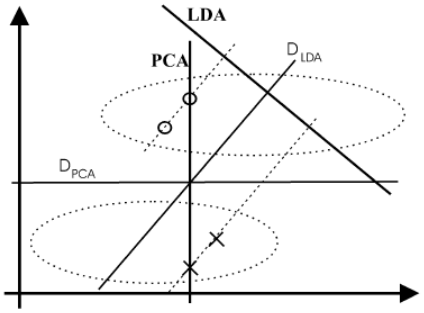
Offenbar ist $\lambda_1 = \mathbf{L}_1' \mathbf{L}_1$. P muß also so bestimmt werden, dass λ_1 maximal ist. Offenbar ist λ_1 maximal, wenn $\lambda_1 = \infty$, aber das ist keine vernünftige Lösung. Man muß eine Nebenbedingung einführen, und die besteht darin, dass man $P'P = I$, I die Einheitsmatrix, fordert, d.h. man fordert, dass die Spaltenvektoren von P auf die Länge 1 normiert sein sollen. Demensprechend soll der Rayleigh-Quotient

$$\frac{P'(X'X)P}{P'P} \stackrel{!}{=} \lambda_{\max}, \quad P'P = I \quad (1.113)$$

maximiert werden. Nach dem Satz 5.1 von Courant-Fischer gilt

$$\max_{\mathbf{x} \neq 0} \frac{\mathbf{x}'(X'X)\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \frac{\mathbf{P}_1'(X'X)\mathbf{P}_1}{\mathbf{P}_1'\mathbf{P}_1} = \lambda_1, \quad (1.114)$$

Abbildung 7: PCA versus LDA (nach Martinez & Kak (2001)



$\mathbf{P}_1$ der erste Eigenvektor aus P und λ_1 der zugehörige Eigenvektor, und natürlich $\|\mathbf{P}_1\| = \mathbf{P}_1'\mathbf{P}_1 = 1$. Die SVD liefert also die Achsen (Orientierungen) im p-dimensionalen Raum mit der jeweils maximal möglichen Varianz; es sind die Orientierungen der Hauptachsen des durch $X'X$ definierten Ellipsoids. Die Spaltenvektoren von L werden normalisiert, wenn man die jeweiligen Komponenten durch die Länge des Vektors multipliziert, d.h. man bildet $Q = L\Lambda^{-1/2}$, so dass $L = Q\Lambda^{1/2}$, und der Ansatz $XP = L$ führt wegen der Orthonormalität von P auf $X = LP'$, d.h. auf

$$X = Q\Lambda^{1/2}P'; \tag{1.115}$$

dies ist die *Singularwertzerlegung* (SVD = Simgular Value Decomposition) der Matrix X.

Das Maximierungsproblem ist bei der LDA analog: es soll ja $QS_{zw}/QS_{inn} = \mathbf{u}'B\mathbf{u}/\mathbf{u}'W\mathbf{u}$ maximiert werden,d.h. man maximiert die quadratische Form $\mathbf{u}'B\mathbf{u}$ unter der Nebenbedingung $\mathbf{u}'W\mathbf{u} = 1$. Die Lösung ist äquivalent einer PCA, die aber in Bezug auf die Matrix B (die der Varianz-Kovarianz-Matrix "zwischen" entspricht) durchgeführt wird. Die maximale Varianz der Projektionen (Koordinaten) der Fälle bei der PCA bedeutet hier, dass die Varianz der Fälle nach Maßgabe der Zugehörigkeit zu den verschiedenen Kategorien oder Klassen maximiert wird. Das Ergebnis dieser PCA *kann* den Ergebnissen der PCA von $X'X$ sehr ähnlich sein, aber im Allgemeinen wird dies nicht der Fall sein, denn bei der PCA von X bzw. $X'X$ werden die Klasssenzugehörigkeiten nicht berücksichtigt. Abbildung 7 illustriert die Unterschiede zwischen den beiden PCAs.

Auf den Hauptachsen $\mathbf{L}_1, \ldots, \mathbf{L}_r$, $r \leq p$ von $X'X$ werden die Klassen nicht notwendigerweise voneinander getrennt. Abbildung 7 illustriert den Sachverhalt. Die zwei Ellipsen repräsentieren zwei Teilpopulationen, die zusammengefasst einer PCA unterzogen werden. Betrachtet man die Projektionen der Ellipsen auf die erste Hauptachse (x-Achse), so sieht man, dass sich die beiden Populationen auf dieser Achse stark überlappen, d.h. die erste Hauptachse trennt nicht zwischen den beiden Populationen. Allerdings trennt in diesem Fall die zweite

Hauptachse zwischen den beiden Populationen – aber dies ist keineswegs der allgemeine Fall. Martinez & Kak (2001) haben vielmehr einen Spezialfall betrachtet, bei dem bei *kleinen* Stichproben die PCA der LDA überlegen sein kann, wenn es um die Trennung von verschiedenen Klassen oder Populationen geht.

1.12 Eigenschaften der Schätzung

Die Klassifikation neuer Fälle kann nach (1.37) gemäß der Beziehung

$$\|\mathbf{y} - \bar{\mathbf{y}}_k\|^2 = (\mathbf{x} - \bar{\mathbf{x}}_k)' W^{-1} (\mathbf{x} - \bar{\mathbf{x}}_k),$$

vorgenommen werden: man entscheidet für die j-te Kategorie, wenn

$$\|\mathbf{y} - \bar{\mathbf{y}}_j\| = \min_k \|\mathbf{y} - \bar{\mathbf{y}}_k\|. \tag{1.116}$$

Die Güte der Vorhersage hängt von W^{-1} ab, wobei W die Schätzung der Varianz-Kovarianz-Matrix "innerhalb" ist. W ist eine symmetrische Matrix mit dem Spektrum $W = P\Lambda P'$, wobei P die Matrix der Eigenvektoren vn W ist und Λ die Diagonalmatrix der Eigenwerte von W. Es ist dann

$$W^{-1} = (P\Lambda P')^{-1} = P\Lambda^{-1}P',$$

denn P is orthonormal. Ist $\mathbf{p}_k$ der k-te Eigenvektor von W (k-te Spalte von P), so gilt

$$W^{-1} = P\Lambda^{-1}P' = \sum_{k=1}^{p} \frac{\mathbf{P}_k \mathbf{P}_k'}{\lambda_k}.$$

Existieren deutlich von 0 verschiedene Kovarianzen in W, so werden zumindest einige Eigenwerte klein und die Elemente von W^{-1} werden groß. Dies impliziert Fehler in den Schätzungen der Vektoren $\mathbf{u}$ und damit eine erhöhte Wahrscheinlichkeit von Fehlklassifikationen. Dieses Problem tritt auf (i) bei korrelierenden Prädiktoren und (ii) bei kleinen Fallzahlen relativ zur Anzahl p der Prädiktoren.

Einen Ausweg aus dem hier entstehenden Problem liefern Shrinkage-Methoden, bei denen überschätzte Komponenten von $\mathbf{u}$ gewissermaßen "geschrumpft" werden. Friedman (1986) entwickelte hierzu die *regularisierte Diskriminanzanalyse*, die in Abschnitt 4.2.2, Seite 71, vorgestellt wird.

1.13 Beispiele II

Beispiel 1: Fishers Irisdaten: Fishers (1936) eigenes Beispiel – die Klassifikation von Pflanzen – gehört zu den Standardbeispielen für die Anwendung der Linearen Diskriminanzanalyse. Tabelle 14 enthält zur Illustration einen Ausschnitt aus diesem berühmten Datensatz. Es gibt vier Prädiktorvariablen: Die Kelchblatt (sepal)- sowie die Blütenblatt (petal)-Länge sowie die entprechenden Breiten in cm, und drei Kategorien (Arten: setosa, versicolor und virginica).

Tabelle 14: Fishers Irisdaten

	Sepal Length	Sepal Width	Petal Length	Petal Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
50	5.0	3.3	1.4	0.2	setosa
51	7.0	3.2	4.7	1.4	versicolor
52	6.4	3.2	4.5	1.5	versicolor
53	6.9	3.1	4.9	1.5	versicolor
⋮	⋮	⋮	⋮	⋮	⋮
100	5.7	2.8	4.1	1.3	versicolor
101	6.3	3.3	6.0	2.5	virginica
102	5.8	2.7	5.1	1.9	virginica
103	7.1	3.0	5.9	2.1	virginica
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3.0	5.1	1.8	virginica

Die Daten wurden mit dem Program lda aus dem Paket MASS Des Statistikprogramms R (http://www.r-project.org/) berechnet. Es werden zwei Kanonische Variablen ausgegeben, wobei die erste 99.1 % der Varianz von QS_{zw} erklärt und die zweite .9 %, – die Daten werden also im Wesentlichen durch eine kanonische Variable erklärt, die zweite dient mehr der Erhöhung der visuellen Deutlichkeit bei der Präsentation der Ergebnisse. Offenbar gelingt bei diesen Daten eine na-

Tabelle 15: Kreuzvalidation: Konfusionen

	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	48	2
virginica	0	2	49

hezu perfekte Klassifikation anhand der Prädiktoren.

Abbildung 8: Beispiel: Fishers Analyse der Irisdaten

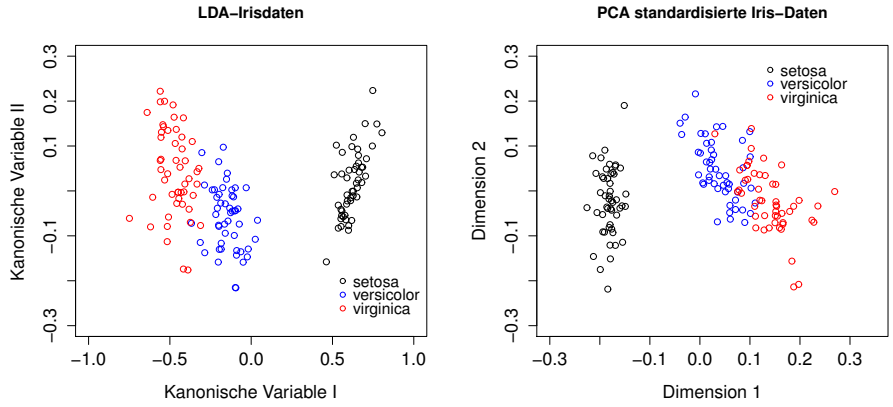
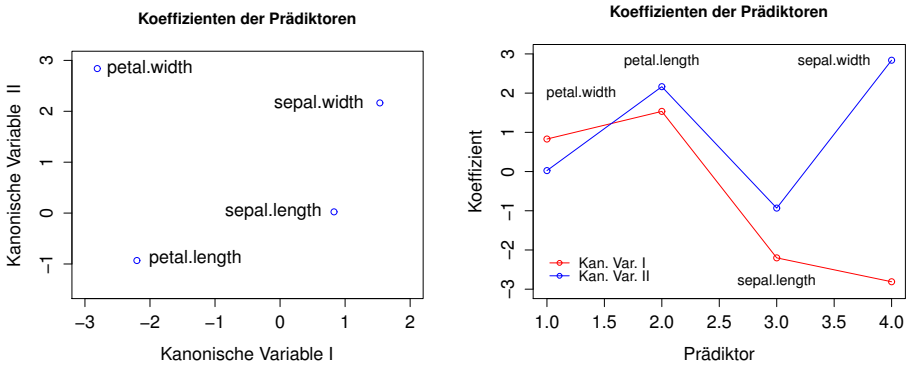


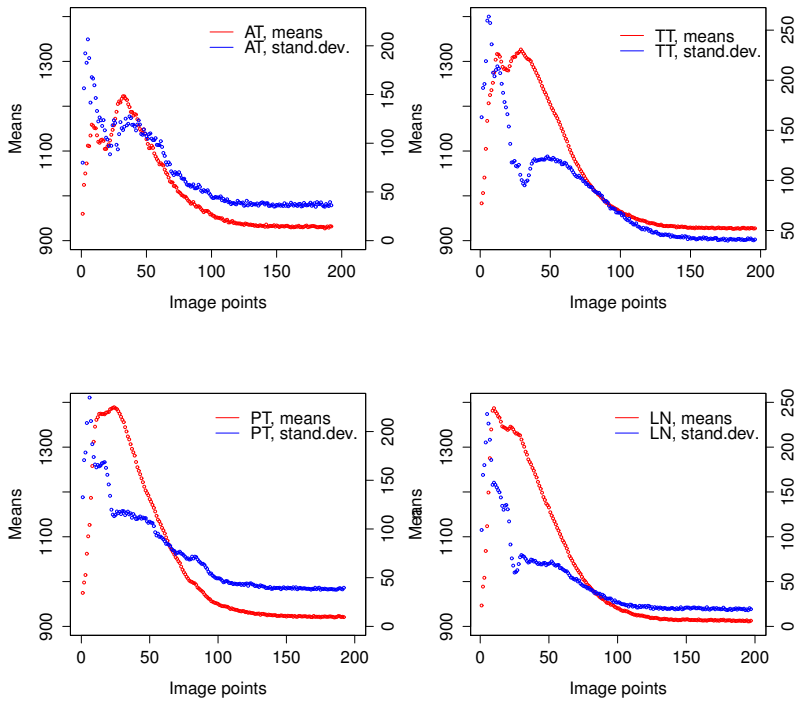
Abbildung 9: Koeffizienten der Prädiktoren



Klassifikation von Schilddrüsengeweben: In der Medizin müssen vielfach Gewebeproben klassifiziert werden. Für diese Aufgabe können OCT-Bilder (OCT – Optical Coherence Tomography) hilfreich sein. Bei dem hier betrachteten Gewebeproben handelt es sich um Schilddrüsengewebe. Die OCT-Bilder sind 2-dimensional. Es wurden 1-dimensionale Profile der Bilder angefertigt, die den Helligkeitsverlauf der Bilder über 190 bis 200 Pixel beschreiben. Die Frage war, ob diese Profile die für die Klassifikation notwendige Information enthalten. Dementsprechend gibt es 192 Prädiktoren, deren Helligkeitswerte als Prädiktorwerte in die Analyse eingingen

Da die Anzahl der Prädiktoren für alle Klassen gleich sein muß, wurde die Zahl der berücksichtigten Pixel auf die kleinste Anzahl (= 192) begrenzt. Die Helligkeitswerte für die ersten 192 Pixel waren also Prädiktorwerte. Insgesamt standen 291 "Fälle", d.h. Gewebeproben zur Verfügung: 26 Fälle für die Kategorie

Abbildung 10: Mittlere Helligkeiten (Profile) und Standardabweichungen (rechte Skala) von OCT-Bildern (Schilddrüsengewebe)



AT, 102 für die Kategorie TT, 89 für die Kategorie PT und 74 für die Kategorie LN.

Das Program berechnet für K Kategorien $K-1$ kanonische Variablen, in diesem Fall also drei. Die erste dieser Variablen erklärt 57.3 % der QS_{zw}, die zweite erklärt 23.9 % und die dritte erklärt die restlichen 18.8 % von QS_{zw}. Diese Werte legen nahe, dass alle drei kanonischen Variablen für die Klassifikation von Bedeutung sind. Offenbar unterscheiden sich die Profile hauptsächlich im Bereich der ersten 50 Pixel. Darüber hinaus sind die Standardabweichungen der Helligkeitswerte für die verschiedenen Pixel keineswegs konstant, so dass die oft geforderte Annahme der multivariaten Normalverteilung mit homogenen Varianzen bei diesen Daten keinen Sinn macht.

Die Jackknife-Kreuzvalidierung lieferte die folgende Konfusiontabelle: Die Berechnung eine χ^2-Wertes erübrigt sich eigentlich, aber der Vollständigkeit halber sei er genannt: $\chi^2 = 783.458$ bei $df = 9$ Freiheitsgraden; diesem Wert entspricht ein p-Wert mit 16 Nullen nach dem Dezimalpunkt, dann kommt eine 2.

PCA und MDS der Schilddrüsendaten Es ist von Interesse, die PCA der

Abbildung 11: LDA-Ergebnisse

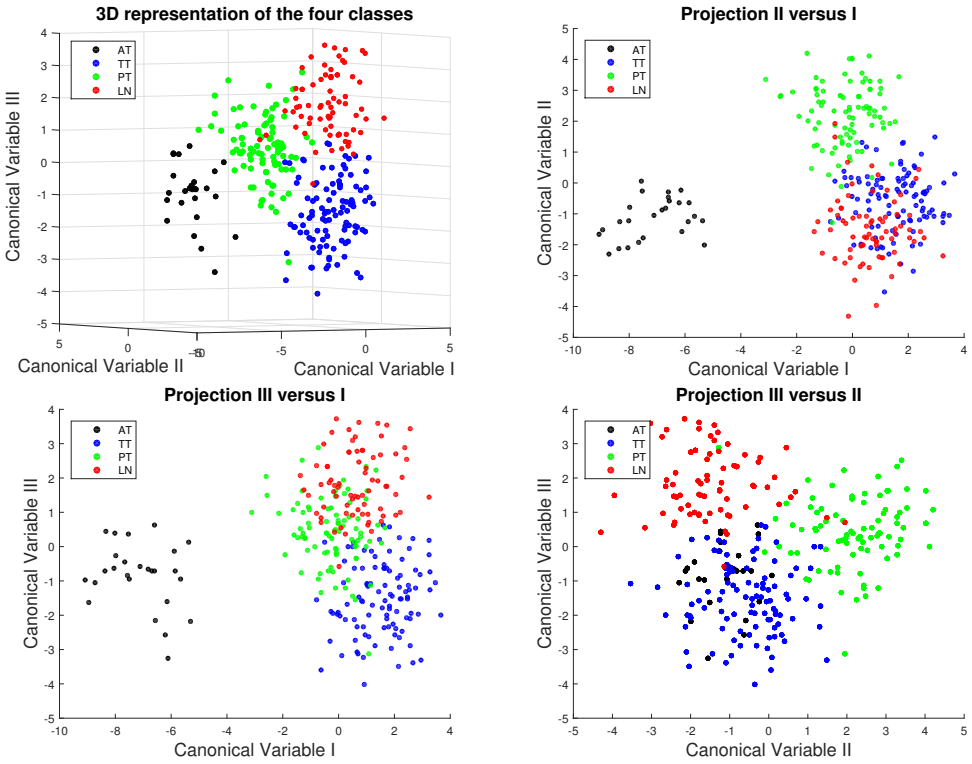
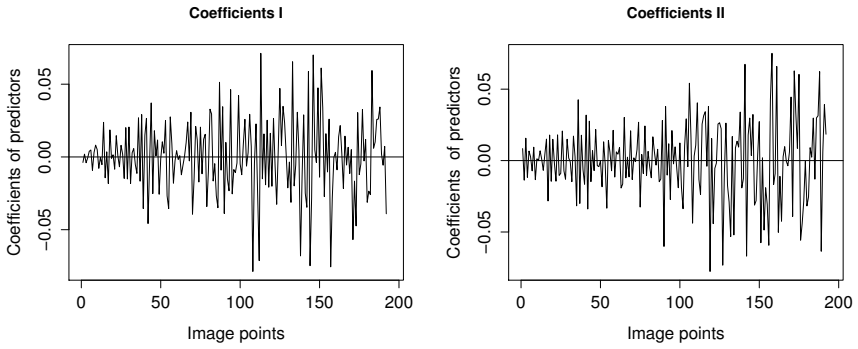


Abbildung 12: LDA – Koeffizienten der Prädiktoren (Pixel)



Daten mit der LDA zu vergleichen.

Abb. 14 zeigt die Resultate (i) einer PCA und (ii) einer multidimensionalen

Abbildung 13: Schilddrüsendaten: PCA-Ergebnisse

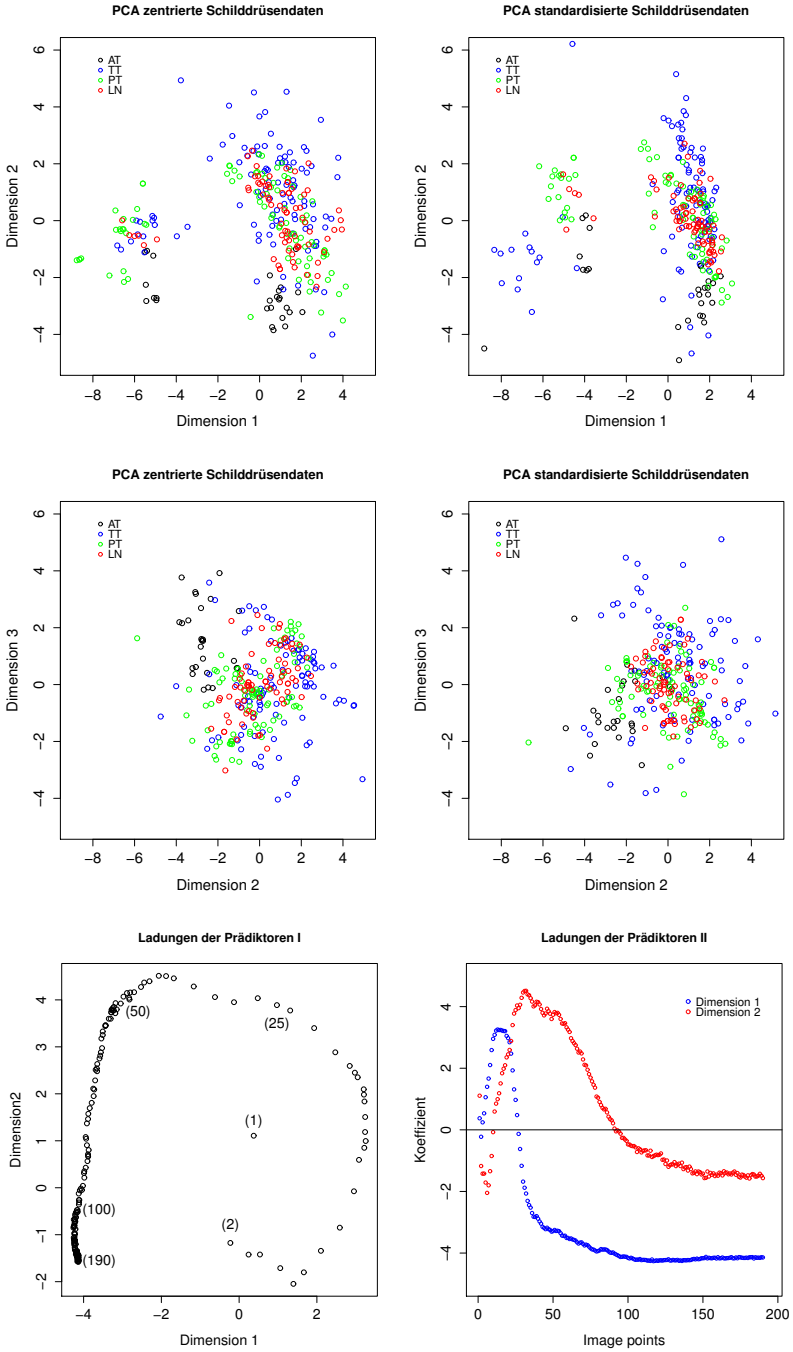


Abbildung 14: Schilddrüsendaten: PCA versus MDS

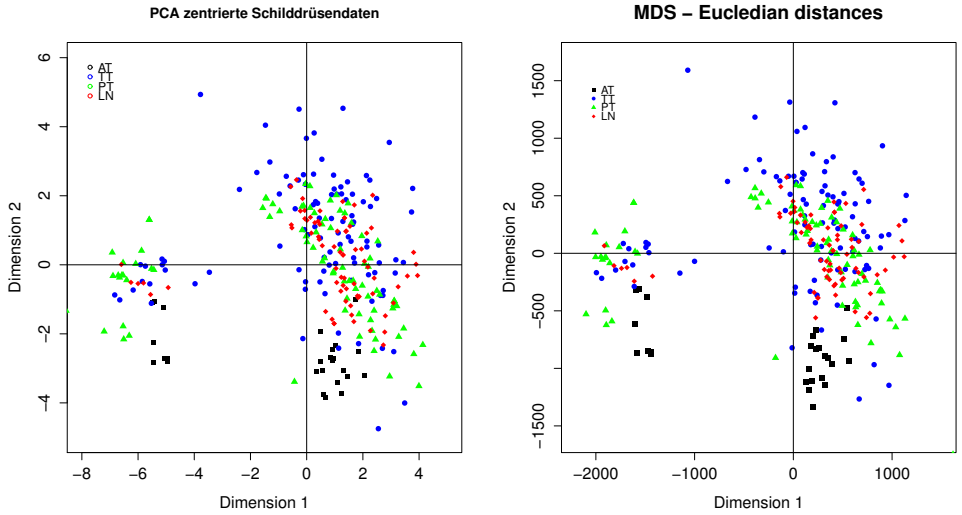


Tabelle 16: Konfusiontabelle

	AT	PT	TT	LN	Kanon. Var.	Anteil an QS_{zw}
AT	26	0	0	0	I	57.3 %
PT	0	96	2	1	II	23.9 %
TT	0	3	85	3	III	18.8 %
LN	0	3	2	70		100 %

Skalierung (MDS), bei der die Distanzen zwischen den Fällen als euklidische Distanzen berechnet werden:

$$d_{ij} = (\sum_{k=1}^{n}(x_{ik} - x_{jk})^2)^{1/2}, \tag{1.117}$$

wobei x_{ik} der Helligkeitswert für das k-te Pixel beim i-ten Fall ist; x_{jk} ist analog definiert. Auf diese Weise werden

$$\binom{291}{2} = \frac{291 \cdot 290}{2} = 291 \cdot 145 = 42195$$

Distanzen zwischen allen möglichen Paaren von Fällen berechnet. Damit wird eine Konfiguration von Punkten, die jeweils einen Fall repräsentieren bestimmt. Durch diese Konfiguration kann ein orthogonales Koordinatensystem gelegt werden, dessen Ursprung mit dem Schwerpunkt der Konfiguration übereinstimmt.

Die Frage ist, welche Orientierung dieses System habenb soll. Das R-PRogramm cmdscale stats bestimmt die Orientierung im Sinne der PCA: die erste Achse wird so gelegt, dass die Varianz der Projektionen der Punkte auf diese Achse maximal wird, die Varianz der Projektionen auf die zweite Achse wird zweitmaximal, etc. Der Vergleich der PCA- mit dem MDS-Resultat zeigt, dass PCA- und MDS-Resultate sehr ähnlich sind. Man könnte vermuten, dass rein von den Distanzen zwischen den Fällen her gesehen die vier Kategorien nicht in separaten Clustern auftreten. Die Fishersche LDA legt nahe, dass eine Orientierung der Achsen existiert, in Bezug auf die die Klassen als separiert erscheinen. Ob man mit anderen als dem LDA-Verfahren zum gleichen Schluß kommt, ist eine andere Frage. Nicht auszuschließen ist auch, dass die euklidischen Distanzen nicht den Distanzen entsprechen, die den kategorienspezifischen Clustern entsprechen. Die Kernmethoden gestatten zumindest im Prinzip, derartige Möglichkeiten zu explorieren.

2 Entscheidungsregeln und Verteilungsannahmen

Gegeben sei ein Vektor $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ mit "Beobachtungen", d.h. Messungen, und die Aufgabe sei, das Objekt oder die Person, an dem bzw. an der diese Messungen gemacht wurden, einer der beiden Klassen Ω_1 oder Ω_2 zuzuordnen. Die Zuordnung von $\mathbf{x}$ zu einer Klasse werde mit $D(\mathbf{x})$ bezeichnet, wobei $D(\mathbf{x}) = D_1$, wenn eine Zuordnung von $\mathbf{x}$ zu Ω_1 erfolgt, und $D(\mathbf{x}) = D_2$, wenn die Zuordnung zu Ω_2 erfolgt. Man kann die Mengen bzw. Klassen Ω_k mit Teilmengen R_k des $\mathbb{R}^p$ identifizieren. Dann kann die Zuordnungsregel wie folgt angeschrieben werden:

$$D(\mathbf{x}) = \begin{cases} D_1, & \mathbf{x} \in R_1 \\ & \\ D_2, & \mathbf{x} \in R_2 \end{cases}, \quad R_1 \cup R_2 = R, \quad R_1 \cap R_2 = \emptyset, \tag{2.1}$$

d.h. wenn $\mathbf{x} \in R_1$ ist, soll das Objekt der Klasse Ω_1 zugeordnet werden, und wenn $\mathbf{x} \in R_2$ ist, soll das Objekt der Klasse Ω_2 zugeordnet werden. Die Aufgabe ist jetzt, R_1 und R_2 so zu bestimmen, dass diese den Mengen dem gewählten Entscheidungskriterium entsprechen. Es sei also eine Grundgesamtheit Ω gegeben, z.B. die Menge der psychiatrischen Patienten, oder die Menge aller Angestellten einer Firma, die Menge aller Studierenden des Faches Psychologie, etc. Ω sei zerlegbar in disjunkte g Teilmengen (Gruppen), d.h. es gelte

$$\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_g, \quad \Omega_i \cap \Omega_j = \emptyset, \quad i \neq j. \tag{2.2}$$

Für ein Element $\omega \in \Omega$ werden p Messungen $x_1, x_2, \ldots, x_p$ von p verschiedenen Variablen durchgeführt. Mit $\mathbf{x} = (x_1, \ldots, x_p)'$ werde der Vektor dieser p Messungen bezeichnet, und mit $\mathbf{x}(\omega)$ ist insbesondere der Vektor der Messungen für das Element ω gemeint. Die Aufgabe besteht nun darin, anhand von $\mathbf{x}(\omega)$ das Element ω einer Klasse oder Gruppe Ω_k, $1 \leq k \leq g$, zuzuordnen.

Kosten: Es seien K_{ij} die Kosten, die entstehen, wenn ein Objekt aus der Klasse C_i der Klasse C_j zugeordnet wird, wobei $i, j = 1, 2$. Im allgemeinen wird $K_{ij} \neq K_{ji}$ gelten, d.h. die Kosten, die bei einer Fehlklassifikation eines Objekts aus der Klasse C_i entstehen, müssen nicht gleich den Kosten sein, die bei einer Fehlklassifikation eines Objekts aus C_j entstehen. So sei etwa Ω_1 die Klasse der gesunden Personen, und Ω_2 die Klasse der an einer bestimmten Krankheit leidenden Personen. K_{12} sind die Kosten der fälschlichen Diagnose einer gesunden Person als "krank", und K_{21} sind die Kosten der falschen Diagnose einer kranken Person als "gesund". Handelt es sich z.B. bei der Krankheit um TBC und ist die Diagnose eine Röntgendiagnose im Rahmen einer Reihenuntersuchung, so ist sicherlich $K_{21} > K_{12}$; die fälschlich als krank betrachtete Person wird sich in Folgeuntersuchungen als gesund herausstellen, aber die fälschlich als gesund klassifizierte Person wird möglicherweise noch kränker, andere anstecken, etc.

Man kann nun die *erwarteten Kosten* einer Fehlklassifikation definieren:

$$\begin{aligned} E(K) &= K_{11}P(D_1|\Omega_1)p(\Omega_1) + K_{12}p(D_2|\Omega_1)p(\Omega_1) \\ &\quad + K_{21}P(D_1|\Omega_2)p(\Omega_2) + K_{22}P(D_2|\Omega_2)p(\Omega_2). \end{aligned} \tag{2.3}$$

K_{ij} wird hier also wie eine zufällige Veränderliche betrachtet, was auch korrekt ist, denn die Zuordnung einer Person zu einer Klasse ist ja in der Weise zufällig, wie die Messungen X mit einem zufälligen Fehler behaftet sind. $p(C_i)$, $i = 1, 2$ sind die *a priori* Wahrscheinlichkeiten für die Wahl eines Objektes oder einer Person aus C_i. Es ist

$$P(D_1|\Omega_1) = \int_{R_1} f_1(\mathbf{x})\, dX, \quad P(D_1|\Omega_2) = \int_{R_1} f_2(\mathbf{x})\, dx. \tag{2.4}$$

$$P(D_2|\Omega_1) = 1 - P(D_1|\Omega_2), \quad P(D_2|\Omega_2) = 1 - P(D_1|\Omega_2). \tag{2.5}$$

Diese Ausdrücke können in (2.3) eingesetzt werden; fasst man die korrespondierenden Ausdrücke zusammen, so ergibt sich

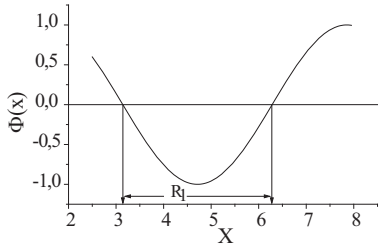
$$\begin{aligned} E(K) &= P(\Omega_1)K_{21} + (1 - P(\Omega_1))K_{22} \\ &+ \int_{R_1} ((1 - P(\Omega_1))(K_{12} - K_{22})f_2(x) - p(\Omega_1)(K_{21} - K_{11})f_1(x)\, dX \end{aligned} \tag{2.6}$$

Minimierung der erwarteten Kosten: Da die Kosten und die a priori Wahrscheinlichkeiten festliegen, sind die ersten beiden Terme auf der rechten Seite fest. Um $E(K)$ zu minimieren, muß der Bereich R_1 geeignet gewählt werden. Die Funktion, für die das Integral über R_1 in (2.6) berechnet werden soll, ist

$$\phi(\mathbf{x}) = ((1 - P(\Omega_1))(K_{12} - K_{22})f_2(\mathbf{x}) - p(\Omega_1)(K_{21} - K_{11})f_1(\mathbf{x}). \tag{2.7}$$

Ein möglicher Verlauf von $\phi(\mathbf{x})$ wird in Abb. 15 gezeigt. Das Integral der Funktion ϕ ist die Differenz der Teilintegrale über (i) den Bereich, in dem $\phi > 0$ ist, und (ii) den Bereich, in dem $\phi < 0$ ist. Das Integral wird dann minimal, wenn man

Abbildung 15: Zur Bestimmung des Integrationsbereichs



nur über den Bereich integriert, in dem $\phi < 0$ ist; dies ist der Bereich R_1. Ist ϕ wie in (3.12) definiert, so ist $\phi < 0$ wenn die Ungleichung

$$(1 - P(\Omega_1))(K_{12} - K_{22}) f_2(\mathbf{x}) < p(\Omega_1)(K_{21} - K_{11}) f_1(\mathbf{x}) \tag{2.8}$$

gilt. Für jeden x-Wert aus dem so definierten Bereich R_1 gilt demnach (2.8). Die Ungleichung läßt sich wie folgt umformen:

$$\frac{f_2(\mathbf{x})}{f_1(\mathbf{x})} < \frac{p(\Omega_1)}{1 - p(\Omega_1)} \frac{(K_{21} - K_{11})}{(K_{12} - K_{22})} \tag{2.9}$$

Der Quotient auf der linken Seite spielt in entscheidungstheoretischen Fragen eine zentrale Rolle:

Definition 2.1 *Es sei $f(x|\Omega_i)$ die Dichte von x unter der Bedingung Ω_i, d.h. die* Likelihood *von x unter der Bedingung Ω_i, $i = 1, 2$. Dann heißt*

$$\lambda(\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\Omega_2)}{f(\boldsymbol{x}|\Omega_1)} \tag{2.10}$$

der Likelihood-Quotient *für die Messungen $\underline{x}$.*

Die Entscheidung nach Maßgabe von (2.9) ist dann eine Entscheidung anhand des Likelihood-Quotienten: setzt man

$$\lambda_0 = \frac{p(\Omega_1)}{1 - P(\Omega_1)} \left(\frac{K_{21} - K_{11}}{K_{12} - K_{22}} \right), \tag{2.11}$$

so entscheidet man nach der Regel

$$\lambda(\mathbf{x}) < \lambda_0 \Rightarrow D(\mathbf{x}) = D_1, \quad \lambda(\mathbf{x}) \geq \lambda_0 \Rightarrow D(\mathbf{x}) = D_2, \tag{2.12}$$

wobei nach (2.1) D_1 die Entscheidung für Ω_1 und D_2 die Entscheidung für Ω_2 bdeutet, so wird man *im Durchschnitt* die Kosten minimieren.

Entscheidungsregeln: Die Beziehung (2.10) definiert bereits die allgemeine Entscheidungsregel: entscheide auf der Basis der Daten x für Ω_2, wenn der Likelihood-Quotient $\lambda(\mathbf{x})$ größer als λ_0 ist, und für Ω_1, wenn er kleiner als λ_0 ist. Der "kritische" Wert λ_0 wird (i) durch die a priori-Wahrscheinlichkeiten $p_1 = p(\Omega_1)$ und $p_2 = p(\Omega_2)$, und (ii) durch die Kosten K_{ij} bestimmt. Allerdings hat man das Problem, die Kosten explizit angeben zu müssen, damit die Regel (2.10) angewendet werden kann. Dies ist in vielen Fällen nicht möglich. Ein Ausweg aus dieser Lage ergibt sich, wenn man die Annahme

$$\left(\frac{K_{21} - K_{11}}{K_{12} - K_{22}}\right) = 1. \tag{2.13}$$

macht. Diesen Fall hat man insbesondere dann, wenn man $K_{ii} = 0$ und $K_{ij} = K_{ji}$ annehmen kann, wenn also korrekte Entscheidungen keine Kosten und Fehlentscheidungen gleiche Kosten verursachen. Diese Annahmen sind nicht in allen Fällen plausibel, aber sie sind gleichbedeutend mit dem Ansatz, die Kosten nicht explizit in Rechnung zu stellen. Jedenfalls ist, wenn (2.13) gilt, $\lambda_0 = p(\Omega_1)/p(\Omega_2)$. Gilt $\lambda(\mathbf{x}) > \lambda_0$, so folgt aus (2.10)

$$\lambda(\mathbf{x})\frac{p(\Omega_2)}{p(\Omega_1)} = \frac{f(\underline{\mathrm{x}}|\Omega_2)}{f(\mathbf{x}|\Omega_1)}\frac{p(\Omega_2)}{p(\Omega_1)} > 1; \tag{2.14}$$

eine analoge Aussage gilt für $\lambda(\mathbf{x}) \leq \lambda_0$. Aber $f(\mathbf{x}|\Omega_1)p(\Omega_1)$ und $f(\mathbf{x}|\Omega_2)p(\Omega_2)$ entsprechen nach dem Satz von Bayes den a posteriori-Wahrscheinlichkeiten $f(\Omega_1|\mathbf{x})$ und $f(\Omega_2|\mathbf{x})$, so daß (2.14) dem Quotienten

$$\frac{f(\Omega_2|\mathbf{x})}{f(\Omega_1|\mathbf{x})} = \frac{f(\mathbf{x}|\Omega_2)}{f(\underline{\mathrm{x}}|\Omega_1)}\frac{p(\Omega_2)}{p(\Omega_1)} \tag{2.15}$$

entspricht. Dieser Quotient führt zu den beiden folgenden Entscheidungsregeln:

1. **Maximum-a-priori-Regel:** Die a-priori-Wahrscheinlichkeiten $p(\Omega_i)$ seien bekannt. Man entscheide sich für Ω_2, wenn $p(\Omega_2|x) > p(\Omega_1|\mathbf{x})$, andernfalls für Ω_1.

   Die Regel läßt sich für g Alternativen verallgemeinern. Demnach hat man die Regel

   $$\text{Entscheide für } \Omega_k, \text{ wenn } p(\Omega_k|\mathbf{x}) = \max_{1\leq j\leq g} p(\Omega_j|\mathbf{x}). \tag{2.16}$$

   Die Regel heißt auch Bayes-Regel, da sie sich direkt aus dem Bayeschen Satz ergibt.

2. **Maximum-Likelihood (ML)-Regel:** Gelegentlich sind die a priori-Wahrscheinlichkeiten nicht bekannt; man kann dann den Fall gleicher a priori-Wahrscheinlichkeiten annehmen. Die a priori-Wahrscheinlichkeiten kürzen sich dann in (2.15) heraus und man erhält die Maximum-Likelihood (ML)-Regel

   $$\text{Entscheide für } \Omega_k, \text{ wenn } f(\mathbf{x}|\Omega_k) = \max_{1\leq j\leq g} f(\mathbf{x}|\Omega_j). \tag{2.17}$$

Diskriminanzfunktionen: Nach (2.14) und (2.16) entscheidet man h für Ω_2, wenn der Quotient $f(\vec{x}|\Omega_2)p(\Omega_2)/(f(\mathbf{x}|\Omega_1)p(\Omega_1) > 1$ ist, andernfalls entscheidet man für Ω_1, d.h. man entscheidet sich für Ω_2, wenn

$$f(\mathbf{x}|\Omega_2)p(\Omega_2) > f(\mathbf{x}|\Omega_1)p(\Omega_1)$$

ist, andernfalls für Ω_1. Nun ist der Logarithmus $\log(x)$ eine monotone Funktion von x: wächst x, so auch $\log(x)$, und fällt x, so auch $\log(x)$ (dies gilt für einen Logarithmus zu einer beliebigen Basis; hier wird immer der natürliche Logarithmus betrachtet). Die Entscheidungsregel kann also auch in der Form

$$\log f(x|\Omega_2) + \log p(\Omega_2) > \log f(x|\Omega_1) + \log p(\Omega_1) \tag{2.18}$$

geschrieben werden. Es wird die folgende Funktion eingeführt:

Definition 2.2 *Es sei*

$$d_k(\boldsymbol{x}) = \log f(\boldsymbol{x}|\Omega_k) + \log p(\Omega_k), \quad 1 \leq k \leq g. \tag{2.19}$$

d_k *heißt dann* Diskriminanzfunktion.

Für $g = 2$ hat man nur zwischen zwei Gruppen oder Klassen Ω_1 und Ω_2 zu entscheiden. Die Einführung der Diskriminanzfunktion erleichtert es, die Entscheidung zwischen einer größeren Zahl g von Klassen oder Gruppen zu diskutieren. Für $g > 2$ kann man paarweise den Likelihood-Quotienten betrachten und sich für dasjenige Ω_k entscheiden, das den größten Quotienten liefert. Dies entspricht der Regel

$$\text{Entscheide für } \Omega_k \text{ (d.h. } \mathbf{x} \in R_k\text{), wenn } d_k(\mathbf{x}) = \max_{1 \leq j \leq g} d_j(x). \tag{2.20}$$

Diese Regel enthält dann als Spezialfall die ML-Regel, wenn die a priori-Wahrscheinlichkeiten nicht berücksichtigt werden sollen bzw. wenn sie identisch sind.

Trennflächen: Will man zwischen den beiden Klassen Ω_j und Ω_k entscheiden, so wird man sich also für Ω_j entscheiden, wenn $d_j(\mathbf{x}) > d_k(\vec{x})$, und für d_k, wenn $d_j(\mathbf{x}) < d_k(\mathbf{x})$. Es sei $\vec{x}_0$ derart, daß

$$d_j(\mathbf{x}_0) = d_k(\mathbf{x}_0), \quad j \neq k. \tag{2.21}$$

$\mathbf{x}_0$ trennt dann die Bereiche von Datenvektoren x, für die man sich für Ω_j oder für Ω_k entscheidet. Die Menge der Vektoren x, die der Gleichung (2.21) genügt, bildet im allgemeinen Fall eine (Hyper-)Fläche im p-dimensionalen Raum, wenn p die Anzahl der Komponenten des Vektors $\mathbf{x}_0$ ist. Diese Flächen werden durch die Art der Dichten $f(x|\Omega)$ definiert, wobei man sich i.a. auf die multivariate Normalverteilung konzentriert, die in Abschnitt 2.1 eingeführt wird.

Im allgemeinen Fall ist $g > 2$; um sich für eine Klasse Ω_k zu entscheiden, muß man im Prinzip $\binom{g}{2} = g(g-1)/2$ Vergleiche durchführen. Andererseits wird durch die Bedingungen (2.21) der Raum in Teilräume R_k, $k = 1, 2, \ldots, g$ aufgeteilt; findet man $\mathbf{x} \in R_k$, so wird man sich für Ω_k entscheiden. Ist z.B. $g = 3$, so gibt es die Teilräume R_1, R_2 und R_3. Die R_k sind durch die Bedingungen

$$d_1(\mathbf{x}) = d_2(\mathbf{x}) \text{ und } d_1(\mathbf{x}) = d_3(\mathbf{x}) \tag{2.22}$$

$$d_2(\mathbf{x}) = d_1(\mathbf{x}) \text{ und } d_2(\mathbf{x}) = d_3(\mathbf{x}) \tag{2.23}$$

$$d_3(\mathbf{x}) = d_1(\mathbf{x}) \text{ und } d_3(\mathbf{x}) = d_2(\mathbf{x}) \tag{2.24}$$

definiert.

Fehlerraten: Alle hier betrachteten Entscheidungen sind probabilistisch und damit kann die Möglichkeit einer Fehlentscheidung nicht ausgeschlossen werden. Dementsprechend kann man die Fehlerrate bestimmen. Dazu sei T die Menge der Werte, die x überhaupt annehmen kann. Jede Entscheidungsregel definiert implizit einen Teilbereich T_k derart, dass man für Ω_k entscheidet, wenn $\mathbf{x} \in T_k$. Für den Fall, dass man nur zwischen zwei Möglichkeiten entscheiden muß, macht man also einen Fehler, wenn man für Ω_1 entscheidet, obwohl Ω_2 zutrifft, und umgekehrt. Die Wahrscheinlichkeit eines Fehlers ist dann durch

$$\epsilon = \int_{T_2} f(\mathbf{x}|\Omega_1)p(\Omega_1)d\mathbf{x} + \int_{T_1} f(\mathbf{x}|\Omega_2)p(\Omega_2)d\mathbf{x} \tag{2.25}$$

gegeben.

2.1 Klassifikation und die multivariate Normalverteilung

2.2 Multivariate Normalverteilung und Mahalanobis-Distanz

Es werde angenommen, dass $\mathbf{x}$ p-dimensional normalverteilt ist, d.h. man misst p "Symptome", die jeweils normalverteilt sind und die paarweise korreliert sein dürfen (nicht müssen). Die p-dimensionale Normalverteilung ist durch

$$f(\mathbf{x}|\Omega_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \vec{\mu}_k)'\Sigma_k^{-1}(\mathbf{x} - \vec{\mu}_k)\right) \tag{2.26}$$

definiert. $\vec{\mu}_k$ ist der Vektor der Erwartungs- (Mittel-)werte der Komponenten von $\mathbf{x}$ (also den gemessenen "Symptomen"), wenn Ω_k die Klasse ist, aus der ω kommt, und Σ_k ist die Matrix der Kovarianzen bzw. Varianzen (oder Korrelationen) zwischen den Komponenten von $\vec{x} \in \Omega_k$, und Σ_k^{-1} ist die zu Σ_k inverse Matrix; es wird vorausgesetzt, dass diese Inverse tatsächlich existiert, d.h. dass $|\Sigma^{-1}| \neq 0$ gilt[5].

[5] Mit $|\Sigma^{-1}|$ wird die *Determinante* von Σ^{-1} bezeichnet. Die Determinante einer Matrix ist eine reelle Zahl, die ungleich Null ist, wenn die $p \times p$-Matrix Σ den vollen Rang p hat. Determinanten werden im Folgenden nicht weiter benötigt, so dass keine weitere Definition dieser Größe gegeben wird.

Sind $\mathbf{x}$ und $\mathbf{y}$ zwei p-dimensionale Vektoren, so ist die Länge des Vektors $\mathbf{x}-\mathbf{y}$ oder des Vektors $\mathbf{y}-\mathbf{x}$ durch

$$d_{xy} = d_{yx} = \left(\sum_{j=1}^{p} (x_j - y_j)^2 \right)^{1/2}$$

gegeben, d.h. durch die Anwendung des Satzes von Pythagoras auf die Differenzen der korrespondierenden Komponenten von $\mathbf{x}$ und $\mathbf{y}$. $d_{xy} = d_{yx}$ heißt auch *Euklidische Distanz* zwischen den Endpunkten dieser beiden Vektoren. Die Euklidische Distanz ist einfach die Länge der kürzesten Verbindung zwischen den Punkten. Dieser Distanzbegriff ist aber ein Spezialfall: will man in einer Stadt von einem Punkt zu einem anderen gelangen, so wird man i.a. nicht die Luftlinie, eben die Euklidische Distanz zu bewältigen haben. Sind die Straßen, wie in Manhattan, in zwei Mengen jeweils parallel zueinander verlaufenden Straßen angeordnet, wobei die Straßen der einen Menge orthogonal zu denen der andere Menge verlaufen, so wird die zurückzulegende Strecke die Summe von jeweils orthogonalen Teilstrecken sein. Die Distanz ist dann definiert durch

$$d_{xy}^{M} = \sum_{j=1}^{p} |x_j - y_j|, \;\; p \geq 1.$$

Diese Definition ist wiederum ein Spezialfall einer klasse von Distanzen, auf die aber nicht weiter eingegangen werden muß, der Zweck dieser Betrachtungen ist nur, zu zeigen, dass der Distanzbegriff mehr als eine Spezifikation zuläßt. Für die Zwecke dieses Skriptums ist der in der folgenden Definition eingeführte Begriff der Mahalanobis-Distanz von Bedeutung:

Definition 2.3 *Die Größe*

$$\delta(\boldsymbol{x}, \vec{\mu}_k) = \sqrt{(\boldsymbol{x} - \vec{\mu}_k)' \Sigma_k^{-1} (\boldsymbol{x} - \vec{\mu}_k)} \tag{2.27}$$

heißt Mahalanobis-Distanz *(Mahalanobis (1936)) zwischen den durch die Endpunkte der Vektoren $\boldsymbol{x}$ und $\vec{\mu}_k$ definierten Punkten.*

Anmerkungen zum Begriff der Mahalanobis-Distanz:

1. Die Menge der $\mathbf{x}$, für die $\delta(\mathbf{x}, \vec{\mu}_k) =$ konstant ist, hat nach (2.26) die gleiche Dichte, d.h. $\delta(\mathbf{x}, \vec{\mu}_k) =$ konstant definiert einen geometrischen Ort gleicher Wahrscheinlichkeit[6].

[6] Diese Ausdrucksweise ist ein wenig lax, da Dichten ja keine Wahrscheinlichkeiten sind; streng genommen kann man nur von $f(\mathbf{x}|\Omega_k)d\mathbf{x}$, wobei das Differential $d\mathbf{x}$ ist, als einer Wahrscheinlichkeit reden.

Es sei $\vec{\xi}_k = \mathbf{x} - \vec{\mu}_k$ der Vektor der Differenzen $x_j - \mu_{kj}$, x_j die j-te Komponente von $\mathbf{x}$ und μ_{kj} die j-te Komponente von $\vec{\mu}_k$. Dann ist

$$\delta^2 = (\mathbf{x} - \vec{\mu}_k)' \Sigma_k^{-1} (\mathbf{x} - \vec{\mu}_k) = \vec{\xi}_k{}' \Sigma_k^{-1} \vec{\xi}_k \tag{2.28}$$

für festes δ eine *quadratische Form*: δ^2 ist stets eine positive reelle Zahl, d.h. $\delta^2 \geq 0$. Die Endpunkte der Vektoren $\vec{\xi}_k$, für die δ^2 einen bestimmten Wert hat, liegen auf der Oberfläche eines p-dimensionalen Ellipsoids; für $p = 2$ ist dies gerade eine Ellipse.

Die Länge des Vektors $\vec{\xi}_k$ ist gerade die Länge des Vektors, der vom Endpunkt des Vektors $\mathbf{x}$ zum Endpunkt des Vektors $\vec{\mu}_k$ zeigt, d.h. die Länge des Vektors $\vec{\xi}_k$ ist gerade die euklidische Distanz zwischen den Endpunkten von $\mathbf{x}$ und $\vec{\mu}_k$. Man betrachte nun die Menge der Vektoren ξ, für die $\vec{\xi}' \Sigma^{-1} \vec{\xi} = \delta^2$ eine Konstante ist. Der Anschaulichkeit wegen sei $p = 2$. Σ^{-1} ist eine symmetrische, positiv definite Matrix und definiert damit ein Menge von Ellipsoiden, d.h. im Fall $p = 2$ eine Menge von Ellipsen, und die Endpunkte der Vektoren $\vec{\xi}$ liegen auf der durch den Wert von δ^2 festgelegten speziellen Ellipse. Es seien insbesondere $\vec{\xi}_1$ und $\vec{\xi}_2$ die beiden Vektoren, die mit den beiden Hauptachsen dieser Ellipse zusammenfallen. Sie haben dann die gleiche Orientierung wie die beiden Eigenvektoren $\mathbf{y}_1$ und $\mathbf{y}_2$ von Σ^{-1}; sie unterscheiden sich von den Eigenvektoren nur insofern, als die Eigenvektoren üblicherweise die Länge 1 haben, aber diese Normierung ist nicht wesentlich. Es sei Y die Matrix der Eigenvektoren von Σ^{-1}; dann gilt $\Sigma^{-1} Y = Y\Lambda$, und wegen der Orthonormalität von Y folgt $\Sigma^{-1} = Y\Lambda Y'$. Die Inverse von Σ^{-1} ist Σ, so dass

$$\Sigma = (Y\Lambda Y')^{-1} = (Y')^{-1} \Lambda^{-1} Y^{-1} = Y\Lambda^{-1} Y', \tag{2.29}$$

denn es ist $Y^{-1} = Y'$. Die Matrizen Σ und Σ^{-1} haben also die gleichen Eigenvektoren, und die Eigenvektoren von Σ sind die Reziprokwerte der Eigenvektoren von Σ^{-1}. Insbesondere gilt dann

$$\Sigma^{-1} \mathbf{y}_1 = (1/\lambda_1)\mathbf{y}_1, \quad \Sigma^{-1} \mathbf{y}_2 = (1/\lambda_2)\mathbf{y}_2. \tag{2.30}$$

und wegen der Orthonormalität von $\mathbf{y}_1$ und $\mathbf{y}_2$ folgen die Beziehungen

$$\delta_1^2 = \mathbf{y}_1' \Sigma^{-1} \mathbf{y}_1 = 1/\lambda_1, \quad \delta_2^2 = \vec{y}_2' \Sigma^{-1} \mathbf{y}_2 = 1/\lambda_2, \tag{2.31}$$

d.h. die Quadrate der Mahalanobis-Distanzen für die Eigenwerte $\mathbf{y}_1$ und $\mathbf{y}_2$ sind gerade durch die Reziprokwerte der Eigenwerte von Σ gegeben, d.h. aber $\delta_1 = 1/\sqrt{\lambda_1}$ und $\delta_2 = 1/\sqrt{\lambda_2}$. Da λ_1 und λ_2 Eigenwerte von Σ sind, gilt $\sqrt{\lambda_1} \geq \sqrt{\lambda_2}$, und mithin $\delta_1 < \delta_2$. Nun sind die Längen der Hauptachsen der durch die Varianz-Kovarianz-Matrix Σ definierten Ellipsen stets proportional zu $1/\lambda_k$, $k = 1, 2$, vergl. das Skriptum Faktorenanalyse, Seite 28, d.h. die Länge der ersten Hauptachse ist gleich $a_1 = \sqrt{k_0/\lambda_1}$,

die der zweiten ist gleich $a_2 = \sqrt{k_0/\lambda_2}$. Die Mahalanobis-Distanzen für die Eigenvektoren sind gerade proportional zu den Längen der Hauptachsen. Die Konstante $k_0 = 1$ korrespondiert dann zur Länge der Eigenvektoren.

Obwohl also $\mathbf{y}_1$ und $\mathbf{y}_2$ die gleiche Länge haben, sind die zugehörigen Mahalanobis-Distanzen verschieden. Die Mahalanobis-Distanz zum Endpunkt von $\mathbf{y}_1$ ist kleiner als die zum Endpunkt von $\mathbf{y}_2$; es läßt sich zeigen, dass die Mahalanobis-Distanz für einen Punkt, der auf der durch $\mathbf{y}_1$ liegenden Geraden liegt und einen euklidischen Abstand von d_e vom Zentrum der Ellipse hat, minimal ist relativ zu den Mahalanobis-Distanzen für Punkte mit gleichem euklidischen Abstand d_e vom Zentrum dr Ellipse, aber mit einer von $\mathbf{y}_1$ abweichenden Orientierung. Die Mahalanobis-Distanz wird maximal, wenn der Punkt mit Abstand d_e auf der Geraden liegt, deren Richtung mit der von $\mathbf{y}_2$ zusammen fällt. Die Lage der Ellipse, die durch Σ beschrieben wird, wird ja durch die korrelative Beziehung zweier Variablen, etwa X_1 und X_2, bestimmt. Für einen gegebenen Wert x_1 von X_1 ist dann derjenige X_2-Wert am wahrscheinlichsten, für den $x_2 = b_1 x_1 + b_0$ gilt, wobei $b_1 = r(s_1/s_2)$ und b_0 die Regressionskoeffizienten sind. Der Punkt hat die euklidische Distanz $d_e = [(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2]^{1/2} = (\xi_1^2 + \xi_2^2)^{1/2}$. Punkte (ξ_1', ξ_2'), die auf einer Geraden liegen, deren Richtung sich von der der Regressionsgeraden unterscheidet, haben eine geringere Wahrscheinlichkeit, wenn sie die gleiche euklidische Distanz d_e von (μ_1, μ_2) haben. Die Wahrschlichkeit wird minimal (relativ zu d_e), wenn sie auf einer Geraden liegen, deren Orientierung orthogonal zu der der Regressionsgeraden ist.

2. Für den 2-dimensionalen Fall kann man sich eine Übersicht über die Abhängigkeit von δ von den Elementen von $\sum$ geben. Dazu sei

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}, \quad \sigma_{12} = \sigma_{21}. \tag{2.32}$$

Die zu Σ inverse Matrix Σ^{-1} ist dann durch

$$\Sigma^{-1} = \begin{pmatrix} \frac{\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, & -\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ -\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, & \frac{\sigma_1^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_1^2(1-r^2)} & -\frac{\sigma_2/\sigma_1}{1-r^2} \\ -\frac{\sigma_2/\sigma_1}{1-r^2} & \frac{1}{\sigma_2^2(1-r^2)} \end{pmatrix} \tag{2.33}$$

gegeben, wobei sich die einfachere rechte Matrix ergibt, wenn man von der Beziehung $r = \sigma_{12}/\sigma_1\sigma_2$ Gebrauch macht. Für die Mahalanobis-Distanz erhält man dann

$$\delta^2 = \vec{\xi}' \Sigma^{-1} \vec{\xi} = \frac{\xi_1^2\sigma_2^2 + \xi_2^2\sigma_1^2 - 2\xi_1\xi_2\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, \tag{2.34}$$

wobei wieder $\vec{\xi} = (\xi_1, \xi_2)' = ((x_1 - \mu_1), (x_2 - \mu_2))'$ gesetzt wurde; (2.34) definiert, wie oben schon angedeutet, für δ = konstant eine Ellipse, die

für $\sigma_{12} = 0$ achsenparallel wird. Für gegebenen Wert von σ_{12} liegt damit jeder Punkt auf einer vom Ellipsen haben die gleiche, durch σ_{12} defnierte Orientierung.

Man kann nun untersuchen, wie δ für gegebenen Vektor $\vec{\xi}$ von σ_{12} abhängt:

(a) $\sigma_{12} = 0$. In diesem Fall erhält man

$$\delta^2 = \frac{\xi_1^2\sigma_2^2 + \xi_2^2\sigma_1^2}{\sigma_1^2\sigma_2^2} = \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} = z_1^2 + z_2^2. \qquad (2.35)$$

Dies ist die Euklidische Distanz zwischen den Punkten (x_1, x_2) und (μ_1, μ_2), allerdings in mit $1/\sigma_1$ bzw. $1/\sigma_2$ skalierten Koordinaten. Der geometrische Ort aller Punkte (z_1, z_2), für die δ einen bestimmten Wert hat, ist demnach eine achsenparallele Ellipse.

(b) $\sigma_{12} \neq 0$. In diesem Fall hängt der Wert von δ einerseits von $\xi_1\sigma_2$ und $\xi_2\sigma_1$ ab, andererseits vom Produkt $\xi_1\xi_2\sigma_{12}$ ab. Wichtig dabei ist die Lage der Punkte (x_1, x_2) relativ zum Punkt (μ_1, μ_2). Um zu einer Veranschaulichung der Verhältnisse zu gelangen, faßt man $\xi_1 = x_1 - \mu_1$ und $\xi_2 = x_2 - \mu_2$ als Konstante auf, hält ebenfalls σ_1 und σ_2 konstant und variiert σ_{12} in (2.34), etwa im Intervall[7]

$$-\sigma_1\sigma_2 < \sigma_{12} < \sigma_1\sigma_2. \qquad (2.36)$$

Für $\sigma_{12} \to \sigma_1\sigma_2$ folgt $\delta \to \infty$, da dann der Nenner in (2.34) gegen Null strebt. Um das Verhalten von δ in Abhängigkeit von σ_{12} zu verdeutlichen, ist es illustrativer, ein kleineres Intervall zu betrachten, in dem $|\sigma_{12}| < \sigma_1\sigma_2$ gilt, etwa

$$-\min(\sigma_1, \sigma_2) \leq \sigma_{12} \leq \min(\sigma_1, \sigma_2). \qquad (2.37)$$

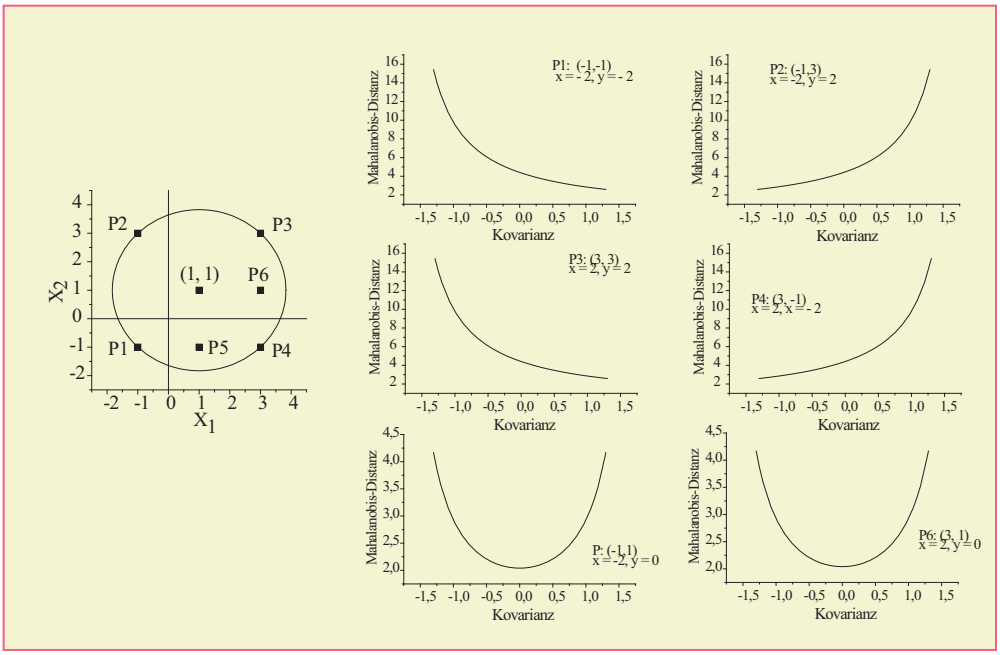
Die Inspektion von (2.34) zeigt, daß der Wert von δ als Funktion von σ_{12} vom Vorzeichen der Differenzen ξ_1 und ξ_2 abhängt; sind die Vorzeichen gleich, wird sich ein anderer Verlauf ergeben als wenn sie ungleich sind. In Abbildung 16 sind verschiedene Verläufe der Mahalanobis-Distanz in Abhängigkeit von der Kovarianz σ_{12} dargestellt worden. Die Form des Verlaufs und der Wertebereich von δ hängen von den Positionen des jeweiligen Punktepaares ab. Hier war einer der beiden Punkte, der Punkt mit den Koordinaten (1, 1), stets der Mittelpunkt

[7] Die Kovarianz ist durch $s_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y})/n$, definiert, die Varianzen durch $s_x^2 = \sum_i (x - \bar{x})^2$, $s_y^2 = \sum_i (y_i - \bar{y})^2$. Es sei $a_i = x_i - \bar{x}$, $b_i = y_i - \bar{y}$. Dann gilt die Schwarzsche Ungleichung

$$|\sum_i a_i b_i|^2 \leq \sum_i |a_i|^2 \sum_i |b_i|^2;$$

der Faktor $1/n$ kürzt sich heraus. Also folgt $|s_{xy}| \leq \sqrt{s_x^2 s_y^2} = s_x s_y$. Das Gleichheitszeichen 903gilt für den Spezialfall $\sigma_1 = \sigma_2$.

Abbildung 16: Mahalanobis-Distanzen zwischen $(1, 1)$ und verschiedenen Punkten für verschiedene Kovarianzen



eines Kreises und die Punkte P_1, P_2, P_3 und P_4 liegen auf dem Umfang des Kreises; sie haben deshalb alle den gleichen (euklidischen) Abstand von (1, 1). Kovariieren die x-Komponenten von x positiv, so müßten alle Punkte in der Nachbarschaft der Geraden durch P_1 und P_3 liegen, kovariieren sie negativ, so liegen sie in der Nachbarschaft der Geraden durch P_2 und P_4 liegen. Für die Punkte P_1 und P_3 ergibt sich demnach der gleiche Zusammenhang zwischen δ und σ_{12}: Für negative Werte von σ_{12} ist δ groß, d.h. $\sigma_{12} \to -\sigma_1\sigma_2$ folgt $\delta \to \infty$, und für $\sigma_{12} \to \sigma_1\sigma_2$ folgt $\delta \to 0$, d.h. je größer die Kovarianz, desto kleiner wird δ. Für die Punkte P_2 und P_4 ergibt sich der umgekehrte Zusammenhang: je größer der Wert der Kovarianz, desto größer wird auch die Mahalanobis-Distanz δ; für $\sigma_{12} \to \sigma_1\sigma_2$ folgt $\delta \to \infty$.

Die Punkte P_5 und P_6 liegen nicht auf den Geraden, die einer perfekten Kovarianz entsprechen, es ergeben sich die in Abb. 16 gezeigten U-förmigen Zusammenhänge. Die Mahalanobis-Distanzen sind hier minimal, wenn die Kovarianz gleich Null ist.

2.3 Ungleiche Varianz-Kovarianzmatrizen

Für die multivariate Normalverteilung sind die Diskriminanzfunktionen für die Maximum-a posteriori-Regel gemäß (2.19) durch

$$d_k(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \vec{\mu}_k)'\Sigma_k^{-1}(\mathbf{x} - \vec{\mu}_k) - \log|\Sigma_k^{-1}| - \log p(\Omega_k) \tag{2.38}$$

gegeben. Man sieht, dass d_k u.a. durch die in (2.27) eingeführte Mahalanobis-Distanz

$$\delta(\mathbf{x}, \vec{\mu}_k) = (\mathbf{x} - \vec{\mu}_k)'\Sigma_k^{-1}(\mathbf{x} - \vec{\mu}_k)$$

definiert ist, denn offenbar gilt

$$d_k(\mathbf{x}) = \frac{1}{2}\delta(\mathbf{x}, \vec{\mu}_k) - \log|\Sigma_k^{-1}| - \log p(\Omega_k). \tag{2.39}$$

Multipliziert man den Term $(\mathbf{x} - \vec{\mu}_k)'\Sigma_k^{-1}(\mathbf{x} - \vec{\mu}_k)$ aus, so erhält man

$$(\mathbf{x} - \vec{\mu}_k)'\Sigma_k^{-1}(\mathbf{x} - \vec{\mu}_k) = \mathbf{x}'\Sigma_k^{-1}\mathbf{x} - 2\vec{\mu}_k'\Sigma_k^{-1}\mathbf{x} + \vec{\mu}_k'\Sigma_k^{-1}\vec{\mu}_k,$$

so dass

$$d_k(\mathbf{x}) = \frac{1}{2}\mathbf{x}'\Sigma_k^{-1}\mathbf{x} - \vec{\mu}_k'\Sigma_k^{-1}\mathbf{x} + \frac{1}{2}\vec{\mu}_k{}'\Sigma_k^{-1}\vec{\mu}_k - \frac{1}{2}(\log|\Sigma_k^{-1}| - \log p(\Omega_k)), \tag{2.40}$$

oder

$$d_k(\mathbf{x}) = \mathbf{x}'A_k\mathbf{x} - B_k\mathbf{x} + C_{k0}, \tag{2.41}$$

mit $A_k = (1/2)\Sigma_k^{-1}$, $B_k = \vec{\mu}_k{}'\Sigma_k^{-1}$ und $C_{k0} = -\frac{1}{2}(\log|\Sigma_k^{-1}| - \log p(\Omega_k))$. A_k ist eine symmetrische Matrix, weshalb $\vec{x}'A_k\mathbf{x}$ eine *quadratische Form*[8] ist. Die Diskriminanzfunktion hängt also von den Quadraten der Komponenten von $\mathbf{x}$ ab.

Die Trennflächen sind Lösungen der Gleichungen $d_j(\mathbf{x}) - d_k(\mathbf{x}) = 0$. (2.41) liefert

$$\begin{aligned} 0 = d_j(\mathbf{x}) - d_k(\mathbf{x}) &= \mathbf{x}'A_j\mathbf{x} - B_j\mathbf{x} + C_{j0} - \mathbf{x}'A_k\mathbf{x} + B_k\mathbf{x} - C_{k0} \\ &= \mathbf{x}'(A_j - A_k)\mathbf{x} - (B_j - B_k)\mathbf{x} + C_{j0} - C_{k0} \end{aligned} \tag{2.42}$$

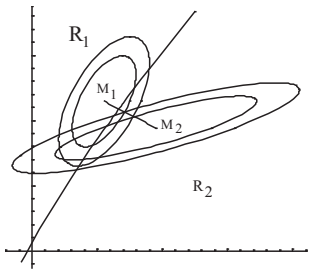
Die Lösungen dieser Gleichung sind im Spezialfall $p = 2$ Ellipsen, Hyperbeln oder Parabeln; für $p > 2$ ergeben sich die entsprechenden Flächen.

2.4 Gleiche Varianz-Kovarianzmatrizen

Es gelte nun $\Sigma_k = \Sigma$ für alle k. Der Spezialfall gleicher Kovarianzmatrizen ist für die Praxis von großer Bedeutung, da einerseits nicht für jedes Ω_k eine besondere

[8] Der Index k werde der Einfachheit halber fortgelassen. Für symmetrisches A ist $\mathbf{x}'A\mathbf{x} = \sum_k a_{ii}x_i^2 + 2\sum_{i\neq j} a_{ij}x_ix_j$; deshalb der Ausdruck quadratische Form.

Abbildung 17: Gaussverteilungen mit ungleichen Varianz-Kovarianz-Matrizen und nichtlinearer Trennung der Bereiche; M_1 und M_2 Mittelpunkte der Ellipsen.



Varianz-Kovarianzmatrix geschätzt werden muß, und andererseits sich einfachere Diskriminanzfunktionen ergeben. Das Quadrat der Mahlanobis-Distanz ist nun

$$\delta^2(\mathbf{x}, \vec{\mu}_k) = (\mathbf{x} - \vec{\mu}_k)'\Sigma^{-1}(\mathbf{x} - \vec{\mu}_k),$$

d.h. es ist $\Sigma_k = \Sigma$ für alle k, so dass nach (2.39)

$$d_k(\mathbf{x}) = \frac{1}{2}\delta(\mathbf{x}, \vec{\mu}_k) - \log|\Sigma^{-1}| - \log p(\Omega_k). \tag{2.43}$$

Aus (2.40) erhält man für $\Sigma_k = \Sigma$ sofort

$$d_k(\mathbf{x}) = \frac{1}{2}x'\Sigma^{-1}\mathbf{x} - \mu_k'\Sigma^{-1}\mathbf{x} + \frac{1}{2}\vec{\mu}_k{}'\Sigma^{-1}\vec{\mu}_k - \frac{1}{2}(\log|\Sigma^{-1}| - \log p(\Omega_k)). \tag{2.44}$$

Da Σ für alle k identisch ist, tragen die Terme $\vec{x}'\Sigma^{-1}\mathbf{x}$ und $\log|\Sigma^{-1}|$ nichts zur Diskriminierung bei und können bei der Definition der Diskriminanzfunktion weggelassen werden. Dementsprechend re-definiert man $d_k(\mathbf{x})$ und betrachtet die Funktion

$$d_k(\mathbf{x}) = -\vec{\mu}_k{}'\Sigma^{-1}\mathbf{x} + \frac{1}{2}\vec{\mu}_k{}'\Sigma^{-1}\vec{\mu}_k - \log p(\Omega_k)). \tag{2.45}$$

Flächen gleicher Distanz: Diese Flächen sind nach (2.21) durch die Gleichungen $d_j(\mathbf{x}) = d_k(\mathbf{x})$ definiert, d.h. es soll $d_j(\mathbf{x}) - d_k(\mathbf{x}) = 0$ gelten. Man findet

$$d_j(\mathbf{x}) - d_k(\mathbf{x}) = \vec{\mu}_k{}'\Sigma^{-1}\mathbf{x} - \vec{\mu}_j{}'\Sigma^{-1}\mathbf{x} + \frac{1}{2}\vec{\mu}_j{}'\Sigma^{-1}\vec{\mu}_j - \frac{1}{2}\vec{\mu}_k{}'\Sigma^{-1}\vec{\mu}_k - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right). \tag{2.46}$$

Diese Gleichung läßt sich zu

$$d_j(\mathbf{x}) - d_k(\mathbf{x}) = (\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}\mathbf{x} - \frac{1}{2}(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}(\vec{\mu}_k + \vec{\mu}_j) - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right) \tag{2.47}$$

vereinfachen. Für die die Trennflächen definierenden $\mathbf{x}$ muß demnach

$$(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}\mathbf{x}\frac{1}{2}(\vec{\mu}_j - \vec{\mu}_k)'\Sigma^{-1}(\vec{\mu}_j + \vec{\mu}_k) - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right) = 0,$$

d.h.

$$(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}\mathbf{x} = \frac{1}{2}(\vec{\mu}_j - \vec{\mu}_k)'\Sigma^{-1}(\vec{\mu}_j + \vec{\mu}_k) - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right) \tag{2.48}$$

gelten. Da aber $(\vec{\mu}_j - \vec{\mu}_k)'\Sigma^{-1}(\vec{\mu}_j + \vec{\mu}_k)$ ein Skalar ist, ist die rechte Seite eine Konstante. $(\vec{\mu}_j - \vec{\mu}_k)'$ ist ein (Zeilen-)Vektor, so dass das Produkt mit der Matrix Σ^{-1}, d.h. $(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}$, ebenfalls ein Zeilenvektor ist, also etwa

$$(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1} = \mathbf{b}'_{kj} = (b_{kj}^{(1)}, \ldots, b_{kj}^{(p)}). \tag{2.49}$$

Dann ist $(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}\mathbf{x}$ ein Skalarprodukt, und (2.48) kann in der Form

$$(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}\mathbf{x} = b_{kj}^{(1)}x_1 + b_2x_2 + \cdots + b_{kj}^{(p)}x_p = K_{jk} = \text{konstant} \tag{2.50}$$

geschrieben werden, wobei die Konstante K_{jk} durch die rechte Seite von (2.48), also

$$K_{jk} = \frac{1}{2}(\vec{\mu}_j - \vec{\mu}_k)'\Sigma^{-1}(\vec{\mu}_j + \vec{\mu}_k) - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right)$$

gegeben ist. Dies ist die Gleichung einer Hyperebene, und die verschiedenen Ω_k-Bereiche werden demnach durch Hyperebenen getrennt. Für den Fall $p = 3$ erhält man die Gleichung

$$b_{kj}^{(1)}x_1 + b_{kj}^{(2)}x_2 + b_{kj}^{(3)}x_3 = K_{jk}, \tag{2.51}$$

d.h. etwa

$$x_3 = K_{jk}/b_{kj}^{(3)} - (b_{kj}^{(1)}/b_{kj}^{(3)})x_1 - (b_{kj}^{(2)}/b_{kj}^{(3)})x_2, \tag{2.52}$$

also eine Ebene im Raum mit den Koordinaten x_1, x_2, x_3. Für $p = 2$ hat man

$$b_{kj}^{(1)}x_1 + b_{kj}^{(2)}x_2 = K_{jk}. \tag{2.53}$$

Für gegebenen K-Wert besteht zwischen x_1 und x_2 die Beziehung

$$x_2 = K_{jk}/b_{kj}^{(2)} - (b_{kj}^{(2)}/b_{kj}^{(1)})x_1, \tag{2.54}$$

also eine Gerade mit der Steigung $-(b_{kj}^{(2)}/b_{kj}^{(1)})$ und der additiven Konstanten $K_{jk}/b_{kj}^{(2)}$.

Position und Orientierung der Hyperebenen: Erweitert man in (2.48) den Term $\log P(\Omega_k)/P(\Omega_j)$ in (2.47) mit der Mahalanobis-Distanz $\delta(\vec{\mu}_j, \vec{\mu}_k) = (\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}(\vec{\mu}_k - \vec{\mu}_j)$ zwischen $\vec{\mu}_j$ und $\vec{\mu}_k$,

$$\log P(\Omega_k)/P(\Omega_j) = \log P(\Omega_k)/P(\Omega_j)\frac{(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}(\vec{\mu}_k - \vec{\mu}_j)}{(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}(\vec{\mu}_k - \vec{\mu}_j)}$$

und setzt man diesen Ausdruck in (2.46) für $\log P(\Omega_k)/P(\Omega_j)$ ein und zieht dann den Faktor $(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}$ heraus, erhält man

$$d_j(\mathbf{x}) - d_k(\mathbf{x}) = (\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}(\mathbf{x} - \frac{1}{2}(\vec{\mu}_j + \vec{\mu}_k) - \frac{\log(p(\Omega_k)/p(\Omega_j))}{\delta(\vec{\mu}_j, \vec{\mu}_k)}(\vec{\mu}_j - \vec{\mu}_k)).$$

Zur Vereinfachung kann man

$$\mathbf{x}_0 = \frac{1}{2}(\vec{\mu}_j + \vec{\mu}_k) - \frac{\log(p(\Omega_k)/p(\Omega_j))}{\delta(\vec{\mu}_j, \vec{\mu}_k)}(\vec{\mu}_j - \vec{\mu}_k)) \tag{2.55}$$

setzen; dieser Term ist von $\mathbf{x}$ unabhängig, so dass man für die Fläche, die die Bereiche für Ω_j und Ω_k trennt, vereinfacht

$$(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}(\mathbf{x} - \mathbf{x}_0) = 0 \tag{2.56}$$

schreiben kann.

Der Vektor $(\vec{\mu}_k - \vec{\mu}_j)$ entspricht der Geraden, die die Punkte $\vec{\mu}_j$ und $\vec{\mu}_k$ verbindet. (2.56) bedeutet dann, dass die Vektoren $\mathbf{x}$, die in der trennenden Hyperebenen liegen, nicht orthogonal zu $(\vec{\mu}_k - \vec{\mu}_j)$ sind. Orthogonal sind nur die Vektoren $(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1} = \mathbf{b}$ und $\mathbf{x} - \mathbf{x}_0$. Die Hyperebene $b_1x_1 + \cdots + b_px_p$ schneidet die Gerade $(\vec{\mu}_k - \vec{\mu}_j)$ auch nicht notwendig in deren Mittelpunkt.

Die Schätzung von μ_k und Σ_k: Im allgemeinen sind $\vec{\mu}_k$ und Σ_k nicht bekannt. Als Schätzungen $\hat{\vec{\mu}}_k$ und $\hat{\Sigma}_k$ nimmt man die empirischen Mittelwerte und Kovarianzen, d.h. man setzt

$$\hat{\vec{\mu}}_k = \vec{\bar{x}}_k, \quad \hat{\Sigma}_k = S_k, \tag{2.57}$$

wobei S_k die empirische Matrix der Varianzen und Kovarianzen ist.

Kann die Annahme gemacht werden, daß $\Sigma_k = \Sigma$ gilt, d.h. daß die Varianzen und Kovarianzen für alle Ω_k gleich groß sind, so berechnet man

$$S = \frac{1}{N-g}\sum_{s=1}^{g}\sum_{n=1}^{n_k}(x_{kn} - \bar{x}_k)(x_{kn} - \bar{x}_k)', \tag{2.58}$$

S_k bzw. S und $\bar{x}_k$ werden dann für die Größen Σ_k, Σ und μ_k eingesetzt ("plug-in"-Schätzungen).

2.5 Klassifikationen und Fehlklassifikationen

2.6 Klassifikation nach Fisher versus Klassifikation nach Gauss

Der Fishersche Ansatz setzt nicht die Annahme der multivariaten Normalverteilung voraus. Gleichwohl liefert Satz 1.4 eine Beziehung zwischen der Klassifikation nach Fisher einerseits und Gauss andererseits.

In (1.37) wird die Fishersche Kriteriumsgröße zur Mahalanobis-Distanz in Beziehung gesetzt. Klassifiziert man gemäß der Gaussverteilung, so ist nach (2.19)

$$d_k(\mathbf{x}) = \log f(\mathbf{x}|\Omega_k) + \log p(\Omega_k), \quad 1 \leq k \leq g.$$

die Diskriminanzfunktion, woei $\mathbf{x}$ der Vektor $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ der Beobachtungen ist. Für f wird die multivariate Gauss-Verteilung eingesetzt. Nach (2.44) erhält man ann

$$d_k(\mathbf{x}) = \frac{1}{2}\mathbf{x}'\Sigma^{-1}\mathbf{x} - \vec{\mu}_k'\Sigma^{-1}\mathbf{x} + \frac{1}{2}\vec{\mu}_k'\Sigma^{-1}\vec{\mu}_k - \frac{1}{2}(\log|\Sigma^{-1}| - \log p(\Omega_k)), \quad (2.59)$$

bzw. nach (2.45)

$$d_k(\mathbf{x}) = -\vec{\mu}_k'W^{-1}\mathbf{x} + \frac{1}{2}\vec{\mu}_k'W^{-1}\vec{\mu}_k - \log p(\Omega_k)),$$

da der Term $\mathbf{x}'W^{-1}\mathbf{x}$ für alle d_k identisch ist und daher für die Diskrimination keine Information liefert. Hier ist es sinnvoll, doch noch einmal den ursprünglichen Ausdruck (2.59) für d_k zu btrachten: subtrahiert man $\mathbf{x}'W^{-1}\mathbf{x}/2$ und addiert man $\log p(\Omega_k)$ auf beiden Seiten, so erhält man

$$d_k(\mathbf{x}) - \frac{1}{2}\mathbf{x}'W^{-1}\mathbf{x} + \log p(\Omega_k) = -\frac{1}{2}(\mathbf{x} - \vec{\mu}_k)'\Sigma^{-1}(\mathbf{x} - \vec{\mu}_k). \quad (2.60)$$

Man erhält man dann die Beziehung

$$-d_k(\mathbf{x}) + \frac{1}{2}\mathbf{x}'W^{-1}\mathbf{x} + \log p(\Omega_k) = \sum_{j=1}^{s}(y_j - \vec{\mu}_j(Y))^2 = \|\mathbf{y} - \vec{\mu}_k(y)\|^2, \quad (2.61)$$

wodurch die Beziehung zwischen dem Fisherschen Klassifikationsverfahren und dem Verfahren anhand der Gauss-Verteilung explizit gemacht wird.

3 Diskriminanzanalyse bei kategorialen Daten

3.1 Volles multinomiales Modell

Es seien p Merkmale $x_1, \ldots, x_p$ mit jeweils m_i Beobachtungen gegeben; es gibt dann

$$m = \prod_{i=1}^{p} m_i \quad (3.1)$$

mögliche Kombinationen von Merkmalsausprägungen. Die Verteilung der Häufigkeiten ist durch die Multinomialverteilung gegeben. Es sei $x = (x_1, \ldots, x_p)'$ ein Datenvektor, aufgrund dessen das beobachtete Objekt bzw. die Person einer bestimmten Klasse, etwa der k-ten, zugeordnet werden soll. Die Diskriminanzfunktion sei

$$d_k(x) = p(x|k)p(k), \tag{3.2}$$

wobei $p(x|k)$ die Wahrscheinlichkeit des Vektors unter der Bedingung der k-ten Kategorie sei, und $p(k)$ die a-priori-Wahrscheinlichkeit der k-ten Kategorie. Die Beobachtung wird dann der k-ten Klasse zugeordnet, wenn $d_k(x) = \max$.

Die Lernstichprobe bestehe aus einer $(p + 1)$-dimensionalen Kontingenztabelle. $\pi(x, k)$ seien die unbekannten Parameter der Multinomialverteilung. Die Maximum-Likelihood Schätzung für die $\pi(x, k)$ seien

$$\hat{\pi}(x, k) = \frac{n(x, k)}{N}, \tag{3.3}$$

und diese Schätzungen liefern die Diskriminanzfunktionen

$$\hat{d}_k(x) = \hat{\pi}(x, k), \tag{3.4}$$

d.h. es wird diejenige Klasse ausgewählt, die am häufigsten vorkommt. Sind die Stichprobenumfänge allerdings ungleich, so ergeben sich Probleme, da zu viele freie Parameter geschätzt werden müssen. So habe man z.B. 6 dichotome Variablen und 2 Klassen. Dann sind

$$k(\prod_{i=1}^{6} m_i - 1) = 2 \times 2 \times 2 \times 2 \times 2 - 1 = 126$$

freie Parameter zu schätzen!

3.2 Unabhängige binäre Variablen.

Die x_i mögen nur die Werte 1 oder 0 annehmen und stochastisch unabhängig sein. Dann ist

$$\pi_{i1} = p(x_i = 1|k), \quad \pi_{i2} = 1 - \pi_{i1} = p(x_i = 0|k).$$

Die Wahrscheinlichkeit, daß man die Beobachtungen $x_1, x_2, \ldots, x_p$ erhält, ist durch

$$p(x_1, \ldots, x_p|k) = \prod_{i=1}^{p} \pi_{ki}^{x_i}(1 - \pi_{ik}^{1-x_i}) \tag{3.5}$$

gegeben. Die Regel für die Zuordnung zur k-ten Klasse ist

$$\begin{aligned} d_k(x) &= \log p(x|k) + \log p(k) \\ &= \sum_{i^1}^{p} x_i \log \pi_{ik} + \sum_{i=1}^{p} (1 - x_i) \log(1 - \pi_{ik}) + \log p(k) \\ &= \sum_{i=1}^{p} \nu_i x_i + \nu_0, \end{aligned} \tag{3.6}$$

d.h. man erhält eine lineare Diskriminanzfunktion, mit

$$\nu_i = \log \frac{\pi_{ik}}{1 - \pi_{ik}}, \quad \nu_0 = \sum_{i=12}^{p} \log(1 - \pi_{ik}) + \log p(k).$$

Für die π_{ik} erhält man die Maximum-Likelihood Schätzer $\hat{\pi}_{ik} = n_i/N$, $n_i = n(x_i = 1)$, und $\hat{p}(k) = N_k/N$. Das Problem bei diesem Ansatz ist, daß die Unabhängigkeit der x_i i.a. nicht gegeben ist. So sind zum Beispiel Symptome im algemeinen korreliert. Dementsprechend muß man versuchen, das Problem der Abhängigkeiten irgendwie zu umgehen.

3.3 Log-lineare Modelle

Zur Illustration werde von drei dichotomen Merkmalen x_1, x_2, x_3 ausgegangen. Es gebe g Klassen; demnach werden g Stichproben gebildet, die jeweils eine 3-dimensionale Kontingenztabell liefern.

Das saturierte Modell für die k-te Klasse ist dann durch

$$\begin{aligned} \log n^{(k)}_{i_1 i_2 i_3} &= \mu^{(k)} + \mu^{(k)}_{1(i_1)} + \mu^{(k)}_{2(i_2)} + \mu^{(k)}_{3(i_3)} \\ &+ \mu^{(k)}_{12(i_1 i_2)} + \mu^{(k)}_{13(i_1 i_3)} + \mu^{(k)}_{23(i_2 i_3)} + \mu^{(k)}_{123(i_1 i_2 i_3)} \end{aligned} \tag{3.7}$$

gegeben. $n^{(k)}_{i_1 i_2 i_3}$ ist die zu erwartende Häufigkeit in der Zelle (i_1, i_2, i_3); ist n_k der Stichprobenumfang in der k-ten Stichprobe, so ist

$$n^{(k)}_{i_1 i_2 i_3} = p(x_1 = i_1 \cap x_2 = i_2 \cap x_3 = i_3) n_k.$$

(3.7) läßt sich durch Einführung von Dummy-Variablen als Regressionsmodell schreiben. Man erhält

$$\begin{aligned} \log n(x|k) &= \nu^{(k)} + \nu^{(k)}_1 x_1 + \nu^{(k)}_2 x_2 + \nu^{(k)}_3 x_3 \\ &+ \nu^{(k)}_{12} x_1 x_2 + \nu^{(k)}_{13} x_3 + \nu^{(k)}_{23} x_2 x_3 + \nu^{(k)}_{123} x_1 x_2 x_3, \end{aligned} \tag{3.8}$$

wobei $x_i = 0$ oder $x_i = 1$; alternativ kann auch $x_i = 1$ oder $x_i = -1$ gesetzt werden (Effektskalierung). Der Vergleich mit (3.7) liefert

$$\nu^{(k)} = \mu^{(k)}, \;\; \nu^{(k)}_1 = \mu^{(k)}_{1(1)}, \cdots, \nu^{k)}_{123} = \mu^{(k)}_{123(111)}.$$

Für die Bayes-Regel erhält man die logarithmierte Diskriminanzfunktion

$$d_k(x) = \log p(k) - \log n_k + \log n(x|k), \tag{3.9}$$

wobei $p(k)$ die a-priori-Wahrscheinlichkeit für die k-te Klasse ist.

Gleichung (3.8) entspricht dem vollen, d.h. saturiertem Modell. Ein interessanteres Modell erhält man, wenn man einige der Interaktionen weglassen kann. Im Extremfall läßt man alle Interaktionsterme weg; dann erhält man das Modell 1/2/3 der Unabhängigkeit der Variablen. Das beste Modell erhält man durch Durchführung einer log-linearen Analyse, d.h. man findet das sparsamste Modell und bestimmt damit die Diskriminanzfunktion, die die Zuordnung von Objekten bzw. Personen zu Klassen ermöglicht.

3.4 Logit-Modelle

Anhand der Daten $D = \mathbf{x}$ soll entschieden werden, ob ein Objekt der Klasse $\mathcal{C}_1$ oder der Klasse $\mathcal{C}_2$ zugeordnet werden soll. Nach dem Satz von Bayes hat man

$$P(\mathcal{C}_2|\mathbf{x}) = \frac{P(\mathcal{C}_2 \cap \mathbf{x})}{P(\mathbf{x})} = \frac{P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)}{P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2) + P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)} \tag{3.10}$$

Dividiert man rechts den Zähler und den Nenner durch $P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)$, so resultiert

$$P(\mathcal{C}_2|\mathbf{x}) = \frac{1}{1 + \frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)}} = \frac{1}{1 + e^{-\phi(\mathbf{x})}} \tag{3.11}$$

mit

$$\phi(\mathbf{x}) = \log\left(\frac{P(\mathbf{x}|\mathcal{C}_1)P(\mathcal{C}_1)}{P(\mathbf{x}|\mathcal{C}_2)P(\mathcal{C}_2)}\right) \tag{3.12}$$

Gleichung (3.11) repräsentiert das logistische Modell. Es muß allerdings noch festgelegt werden, wie die Funkion $\phi(\mathbf{x})$ aussieht. Eine einfache und direkte Möglichkeit ist, $\phi(\mathbf{x})$ in eine Reihe zu entwickeln:

$$\phi(\mathbf{x}) = a_0 + a_1x_1 + \cdots + a_px_p + a_{p+1}x_1x_2 + \cdots, \tag{3.13}$$

wobei der Ausdruck auf der rechten Seite natürlich nur endlich viele Terme enthält. Die a_j sind dann freie, aus den Daten zu schätzende Parameter. Alternativ dazu kann eine spezielle Verteilungs- bzw Wahrscheinlichkeitsfunktion angenommen werden. So sei

$$P(\mathbf{x}|\mathcal{C}_k) = \alpha \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_k)'\Sigma^{-1}(\mathbf{x} - \mu_k)\right), \quad k = 1, 2$$

wobei α eine Normierungskonstante derart, dass $\int P(\mathbf{x}|\mathcal{C}_k)d\mathbf{x} = 1$. Man rechnet leicht nach, dass diese Annahme eine Reihenentwicklung mit nur linearen Termen impliziert. Ebenso leicht rechnet man nach, dass allgemein

$$\log\frac{p(\mathcal{C}_2|\mathbf{x})}{1 - p(\mathcal{C}_2|\mathbf{x})} = a_0 + a_1x_1 + \cdots + a_px_p + a_{p+1}x_1x_2 + \cdots, \tag{3.14}$$

gilt. Man entscheidet sich für $\mathcal{C}_2$, wenn $P(\mathcal{C}_2)/(1 - P(\mathcal{C}_2)) > 1$, sonst für $\mathcal{C}_1$.

4 Beschränkungen und Erweiterungen der LDA

Fishers LDA leistet gute Dienste, auch wenn die Daten (Messungen der Prädiktorvariablen) nicht normalverteilt sind und wenn die Unterschiede zwischen den Varianzen in den Gruppen nicht allzu groß sind. Vor allem sollte die Anzahl m der "Fälle" – also der Personen oder Objekte, bei denen die Werte der Prädiktorvariablen erhoben werden – hinreichend größer als die Anzahl p der Prädiktorvariablen ist. Als Faustregel wird angenommen, dass $m \approx 3p$ oder besser noch $m \approx 4p$ sein sollte.

Probleme ergeben sich im Allgemeinen dann, wenn m klein im Vergleich zu p ist, insbesondere wenn $m < p$ gilt. Ebenfalls sollten die Korrelationen zwischen den Messungen der Prädiktorvariablen nicht zu hoch sein. Denn die Koeffizienten der Prädiktorvariablen ergeben sich ja als Komponenten von Eigenvektoren der Matrix $W^{-1}B$, und im Falle von hoch korrelierenden Prädoktoren impliziert W^{-1} instabile Schätzungen. Dieser Sachverhalt ist aus der Regressionsrechnung bekannt[9]. Unabhängig von diesen Faktoren kann die Voraussetzung der LDA, dass nämlich die Gruppen durch lineare Funktionen voneinander getrennt werden können (Geraden, Ebenen) nicht erfüllt sein. Man diese Punkte zusammenfassen:
1. Im Falle einer großen Zahl von hoch korrelierenden Prädiktoren ist die LDA zu flexibel, d.h. für die Trainingsstichprobe ergibt sich eine gute Anpassung des Modells, die Voraussagen für neue Fälle sind dagegen schlecht, man hat einen *overfit*,
2. Sind die Trennflächen nichtlinear, so ist die LDA zu rigide, d.h. man hat einen *underfit*.

Der Nachteil der Quadratischen Diskriminanzanalyse ist zunächst die große Anzahl von zu schätzenden Parametern, darüber hinaus ist nicht sicher, dass die Annahme der Gauß-Verteilung gerechtfertigt ist.

Es gibt verschiedene mögliche Auswege aus diesen problematiwchen Situationen: man kann durch Anwendung der PCA zu einer reduzierten Anzahl von unkorrelierten Prädiktoren gelangen, oder man versucht, redundante Prädiktoren aus der Menge der Prädiktoren herauszunehmen. Beim ersten Ansatz ist oft nicht klar, wieviele Dimensionen berücksichtigt werden sollen, beim zweiten Ansatz besteht die Schwierigkeit darin, "redundante" Prädiktoren zu identifizieren. Ein allgemeinerer Ausweg besteht darin, die Matrix W^{-1} durch eine regularisierte Version zu ersetzen. Das Prinzip besteht darin, W durch $W + \lambda\Omega$ zu ersetzen, wobei λ ein zu bestimmender Parameter ist und Ω eine ebenfalls noch zu bestimmende Matrix ist; Ω heißt auch Tychonoff-Matrix, zu Ehren des Erfinders dieses Ansatzes[10] Oft genügt es, für Ω eine Diagonalmatrix, insbesondere die Einheitsmatrix I zu wählen. Der Effekt der Addition von λI zu W besteht darin,

[9] http://www.uwe-mortensen.de/multregressaux.pdf
[10] Andrei Nikolajewitsch Tychonoff (1906–1993), russischer Mathematiker.

zu stabileren Schätzungen der Eigenwerte zu führen. Es läßt sich zeigen, dass $W + \lambda I$ dieselben Eigenvektoren wie W hat, aber die Eigenvektoren $\lambda_j + \lambda$. Da W symmetrisch ist, existiert die Zerlegung $W = P\Lambda P'$, P die Eigenvektoren von W, Λ die Diagonalmatrix der Eigenwerte von W. Dann ist $W^{-1} = P\Lambda^{-1}P'$, d.h.

$$W^{-1} = \sum_{k=1}^{r} \frac{\mathbf{P}_k \mathbf{P}_k'}{\lambda_k}, \quad (W + \lambda I)^{-1} = \sum_{k=1}^{r} \frac{\mathbf{P}_k \mathbf{P}_k'}{\lambda_k + \lambda}.$$

Hohe Korrelationen implizieren kleine Eigenwerte, was z.B. bei der Schätzung von Regressionskoeffizienten zu instabilen Schätzungen der Regressionsparameter führt. Dieser Effekt wird bei der Matrix $W+\lambda I$ reduziert, da λ zu den Eigenwerten addiert wird. Dies ist der Effekt der Regularisierung.

4.1 Anpassung parametrischer Funktionen

In vielen Fällen sind die Prädiktoren Elemente eines funktionalen Verlaufs; die Profile der OCT-Bilder von Gewebeproben (Abbildung 10, Seite 43) sind hierfür ein Beispiel; die Pixel sind die Prädiktoren und die Helligkeiten für die Pixel sind die Werte der Prädiktoren. Die Anzahl der PRädiktoren ist oft groß im Vergleich zur Anzahl der Fälle, – manchmal sogar größer als die Anzahl der Fälle. Die Werte der Prädiktoren mögen einer bestimmten, wenn auch unbekannten Funktion folgen und sind darüber hinaus von zufälligen "Störungen" überlagert. Benachbarte Prädiktorwerte können korreliert sein, – eine Autokorrelationsfunktion kann darüber Auskunft geben. Der Verlauf der Koeffizienten, also die Komponenten des Gewichtevektors $\mathbf{u}$, bildet deswegen nicht notwendig die Wichtigkeit der einzelnen Prädiktoren ab, sondern zeigt einen Verlauf, wie er in Abbildung 12, Seite 44, gezeigt wird: die Werte der Koeffizienten alternieren zwischen positiv und negativ. Es ergibt sich zwar ein guter Fit für die Trainingsstichprobe, aber die Voraussagen für neue Fälle können schlecht sein.

Ein mögliche Lösung des Problems schwer interpretierbarer Koeffizientenverläufe ist (i) der Fit von Funktionen, die den "wahren" Verlauf der Prädiktorwerte approximieren, und (ii) eine Penalisierung ("Bestrafung") für einen zu guten Fit der Approximation. Die Funktion, die den Daten unterliegt, ist normalerweise nicht bekannt, aber es ist bekannt, dass Funktionen im Allgemeinen durch Polynome beliebig genau approximiert werden können. Ein Polynom vom Grad d hat die Form

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_d x^d, \quad a \leq x \leq b \tag{4.1}$$

Das Problem bei der Repräsentation von Funktionen ist, dass man einerseits stets einen Grad d und zugehörige Koeffizienten $a_0, a_1, \ldots, a_d$ finden kann derart, dass $P(x)$ durch jeden Datenpunkt geht, aber natürlich zum Preis einer u. U. großen Zahl zu schätzender freier Parameter. Abgesehen davon werden auch Funktionswerte "erklärt", die nur Aspekte des Rauschens in den Daten, also ztufällige

Effekte darstellen. Die Aufgabe besteht dann darin, eine approximierende Funktion zu finden, die hinreichend "glatt" ist, von der man also annehmen kann, den funktionalen Zusammenhang zu reflektieren, ohne zufällige Komponenten als systematische Aspekte des Zusammenhanges zu überinterpretieren.

Ein oft verwendeter Ansatz besteht darin, *Splines* einzuführen. Dazu wird das Intervall $[a, b]$, auf dem der funkttionale Zusammenhang betrachtet wird, in Teilintervalle $[\xi_j, \xi_{j+1})$ aufgeteilt. Die Endpunkte ξ_j, $j = 1, \ldots, m$ heißen *Knoten*. Auf diesen Teilintervallen werden die Daten jeweils durch ein Polynom des Grades d approximiert, wobei d im Allgemein nicht größer als 3 ist. Etwas genauer werden Polynom-Splines wie folgt definiert (Fahrmeier et al, 2007, p. 295)

Definition 4.1 *Die Funktion* f: $[a, b] \to \mathbb{R}$ *heißt* Polynom-Spline *vom Grad* $l \geq 0$ *zu den Knoten* $a = \xi_1 < \cdots < \xi_m = b$, *wenn die Bedingungen*
1. $f(x)$ *ist* $(l-1)$*-mal stetig differenzierbar. Für* $l = 1$ *bedeutet dies, dass* f *stetig ist, und für* $l = 0$ *werden keine Glattheitsforderungen an die Approximation gestellt.*
2. f *ist auf den durch die Knoten definierten Intervallen* $[\xi_j, \xi_{j+1})$ *ein Polynom vom Grad* l.

Die Differenzierbarkeitsbedingung erlaubt es, die Glattheit der Approximation zu kontrollieren. Der Grad l definiert insgesamt die Glattheit der Approximation, die darüber hinaus durch die Anzahl der Knoten bestimmt wird. Je größer die Anzahl der Knoten, desto größer ist die Vielfalt von funktionalen Zusammenhängen, die approximiert werden können.

Die Approximation entspricht zunächst einer Regression:

$$f(x_i) = \gamma_1 + \gamma_2 x_i + \cdots + \gamma_{l+1} x_i^l + \gamma_{l+2}(x_i - \xi_2)_+^l + \cdots + \gamma_{l+m-1}(x_i - \xi_{m-l})_+^l + \varepsilon_i, \quad (4.2)$$

mit

$$(x - \xi_j)_+^l = \begin{cases} (x - \xi_j)^l, & x \geq \xi_j \\ 0, & \text{sonst} \end{cases}$$

Man sieht, dass sich der Koeffizient γ_k an jedem Knoten ändern kann, womit die globale Glattheitsforderung erfüllt wird.

Polynom-Splines sind nicht das einzige Mittel, um Funktionen zu approximieren, aber es genügt hier, sich auf Splines zu konzentrieren. Splines bilden einen Vektorraum, d.h. man kann sie in einer Weise kombinieren, wie man Vektoren in einem Vektorraum kombinieren kann. Wie die Vektoren in einem Vektorraum als Linearkombination von Basisvektoren dargestellt werden können, so kann man auch im Vektorraum der Splines Basisfunktionen oder Basis-Splines finden. Eine solche Basis ist

$$\begin{aligned} B_1(x) &= 1,\, B_2(x) = x,\, \ldots, B_{l+1}(x) = x^l, \\ B_{l+2}(x) &= (x - \xi_2)_+^l, \ldots, B_d(x) = x - \xi_{m-1})_+^l \end{aligned} \quad (4.3)$$

Die Funktionen $B_1, \ldots B_d$ heißen auch *Basisfunktionen*. Definiert man

$$Z = \begin{pmatrix} B_1(x_1) & B_2(x_1) & \cdots & B_d(x_1) \\ B_2(x_2) & B_2(x_2) & \cdots & B_d(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(x_n) & B_2(x_n) & \cdots & B_d(x_n) \end{pmatrix} \tag{4.4}$$

und ist $\vec{\gamma}$ der Vektor $(\gamma_1, \ldots, \gamma_d)'$ der Koeffizienten, so hat man

$$f(x) = Z\vec{\gamma} + \vec{\varepsilon} \tag{4.5}$$

Eine Schätzung für $\vec{\gamma}$ erhält man über die Methode der Kleinsten Quadrate als KQ-Schätzung

$$\hat{\gamma} = (Z'Z)^{-1}Z'\mathbf{x}. \tag{4.6}$$

Für die zu approximierende Funktion erhält man die Schätzung

$$\hat{f}(\mathbf{z}) = \mathbf{z}'\hat{\gamma}, \quad \mathbf{z} = (B_1(z), \ldots, B_d(z))'. \tag{4.7}$$

Penalisierung: Die Güte der Approximation hängt stark von der Anzahl der Knoten ab. Man hat zwei Möglichkeiten:
1. Man wählt den Wert von d adaptiv, d.h. man versucht durch verschiedene Annahmen über d eine optimale Anpasung zu finden, oder
2. Man regularisiert die Schätzung durch eine Penalisierung der Schätzung.

Bei der Penalisierung beginnt man mit einer großen Anzahl von Knoten (20 bis 40). Damit wird die Approximation auf jeden Fall hinreichend *flexibel*, um auch eine stark variierende Funktion annähern zu können. Dann führt man einen Straf- oder Penalisierungsterm ein, der eine zu große Variabilität der Schätzung bestraft. Dies führt zum *penalisierten KQ-Kriterium*:

$$\text{PKQ}(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=l+2}^{d} \gamma_j^2. \tag{4.8}$$

Die Summe $\sum_{j=l+2}^{d} \gamma_j^2$ bedeutet, dass insbesondere betragsmäßig große Koeffizienten penalisiert werden, wodurch "rauhe" Schätzungen der Funktion geglätted werden. Das Ausmaß der Penalisierung wird durch den Wert von *lambda* kontrolliert. Für einen kleinen Wert ($\lambda \to 0$) bekommt die Penalisierung nur ein kleines Gewicht, für einen großen Wert von λ wird stark geglätted. Die Glättung wird nicht mehr über die Anzahl und Position der Knoten, sondern über λ gesteuert.

Verwendet man B-Splines, so muß man etwas anders vorgehen. Es gibt keine Trennung in einen parametrischen Teil und den Abweichungen von diesem Teil. Die Aufgabe ist nun, ein Maß für die Glattheit der angepassten Funktion zu definieren. Ist f die Funktion, so gibt die erste Ableitung $f' = df/dx$ das Ausmaß der Veränderung von f im Punkt x an, und die zweite Ableitung $f'' = d^2f/dx^2$

die Veränderung der Veränderung, also das Ausmaß der Krümmung von f im Punkt x. Deswegen läßt sich ein Penalisierungsterm der Form

$$\lambda \int (f''(z))^2 dz \tag{4.9}$$

definieren[11].

4.2 Die Grenzen der LDA und mögliche Auswege

4.2.1 Die Grenzen der LDA

Die Fishersche LDA setzt voraus, dass $n > p$; eine Faustregel besagt, dass n 3- bis 4-mal sogroß wie p sein soll. In der Praxis ist es oft schwierig, für gegebene Anzahl p von Prädiktoren hinreichend große Stichproben zu erheben. So hat man bei den OCT-Daten nahezu 200 Prädiktoren, die überdies korreliert sein können, so dass sich die Frage stellt, was geschieht, wenn soganr $n < p$ ist. Da man oft relativ wenig über die Bedeutung einzelner Prädiktoren weiß, kann man nicht einfach Prädiktoren weglassen. Da die Eigenvektoren von $W^{-1}B$ bestimmt werden müssen, gibt sich schon immer dann ein Problem, wenn W^{-1} schlecht oder gar nicht definiert ist, – auch für $n > p$, aber klein im Vergelich zu p werden die Schätzungen der Kovarianz-Matrix instabil (= variabel) und für $n_k < p$, $1 \leq k \leq K$ folgt, dass nicht alle Parameter identifizierbar sind. Generell gilt

$$S_k = \sum_{j=1}^{p} \lambda_{ik} \mathbf{v}_{ik} \mathbf{v}'_{ik} \Rightarrow S_k^{-1} = \sum_{j} \frac{\mathbf{v}_{ik} \mathbf{v}'_{ik}}{\lambda_{ik}},$$

d.h. die Richtung Diskriminanzfunktion d_k wird stark durch kleine Eigenwerte beeinflußt, so dass es vermehrt zu Fehlklassifikationen kommt (Friedman (1989)). Die Schätzungen von W implizieren im instabilen Fall einen Bias für die Eigenwerte, – sie werden zu klein geschätzt. Diese Effekte werden um so größer, je kleiner die Stichprobe ist. Für $n_k < p$ werden die kleinsten $p - n_k + 1$ Eigenwerte als gleich Null geschätzt, so dass die korrespondierenden Eigenvektoren beliebig werden. Der Haupteffekt ist:
1. Die Wichtigkeit des Teilraums mit kleiner Varianz wird überschätzt.
2. Die Empfindlichkeit für Ausreisser steigt.

Ein Vorteil der Fisherschen LDA besteht darin, dass sie sich im Falle inhomogenen Varianzen als robust erweist. Das Gleiche gilt für den Fall nichtnormalverteilter Daten.

Der Punkt 1. führt auf Methoden der Stabiliserung; dies sind (i) Regularisierung, und (ii) Shrinkage. Dazu müssen die Schätzungen der Eigenwerte verbessert werden. .

[11] R-Paket 'mgcv'.

4.2.2 Regularisierte DA

Ein Schätzproblem ist "poorly-posed", wenn $p \approx n$, und "ill-posed", wenn $n < p$, n Anzahl der Beobachtungen. Regularisierung bedeutet, dass man auf eine noch anzugebende Weise die Varianz der Schätzungen reduziert, so dass diese von den stichprobengebundenen Werte weggeführt werden, – allerdings zum Preis eines Bias in der Schätzung. Damit wird ein durch einen "degree-of-belief"-Parameter kontrollierter trade-off erzeugt. Der Bias ist dabei im Allgemeinen klein im Vergleich zur Reduktion der Varianz der Parameterschätzung.

Friedman (1989) hat hierzu einen ersten Ansatz geleifert. Die Quadratische DA ist ill-posed, wenn $n_k < p$ für irgendein k, und poorly-posed, wenn n_k nur wenig größer als p ist. *Eine* Möglichkeit besteht darin, die $\hat{S}_k$ durch die pooled Schätzung $\hat{S}$ zu ersetzen (vergl. Friedman (1989)). Dazu sei

$$\hat{\Sigma}_k(\lambda) = S_k(\lambda)/W_k(\lambda), \tag{4.10}$$

mit

$$S_k(\lambda) = (1-\lambda)S_k + \lambda S, \quad W_k(\lambda) = (1-\lambda)W_k + \lambda W, l \;\; 0 \leq \lambda \leq 1. \tag{4.11}$$

Die Fishersche LDA ersdcheint als regularisierte Form der Quadratischen DA; Für $\lambda = 0$ erhält man die QDA, für $\lambda = 1$ die LDA. Für Werte dazwischen erhält man eine Regularisierung, die nicht so ausgeprägt wie die LDA ist.

Wie Hastie, Buja und Tibshirani (1995) (p. 75) ausführten, lieferte Friedmann zwar einen wichtigen Ansatz, der aber gleichwohl sein Grenzen hat: Der mit der Ridge-Regression verwandte Ansatz Friedmans impliziert einen Bias gegen den Gesamtmittelwert, der die räumliche Struktur der Prädiktoren ('index domain' bei Hastie et al) ignoriert. Die Koeffizienten ("Gewichte") der Prädiktoren werden dadurch schwer interpretierbar. Die Autoren schlagen vor, den Verlauf dieser Koeffizienten als Funktion ihrer Position über Spline-Funktionen zu glätten, wodurch lokale Kontraste penalisiert werden.

4.3 Flexible, Penalisierte und Mixture Diskriminanzanalyse

Dieser Ansatz geht auf Hastie, Tibshirani und Buja (1994) zurück. Die Idee ist, die Diskriminanzanalyse (LDA) auf eine verallgemeinerte, nichtparametrische lineare Regression zurückzuführen, wodurch sich u.a. Möglichkeit ergibt, die Einschränkung auf lineare Entscheidungsgrenzen zu überwinden. Eine weitere Einschränkung der Fisherschen LDA ist die Idee, dass jede Klasse durch einen Prototyp repräsentiert wird; dieser Prototyp ist das Klassencentroid. Bei den OCT-Profilen ist der Prototyp das jeweils mittlere Profil (dies ist ebenfalls ein Centroid). Es ist aber möglich, dass eine Klasse durch verschiedene Prototypen repräsentiert werden kann. Darüber hinaus kann es zu viele, korrelierte Prädiktoren geben. Hierfür können die OCT-Profile ebenfalls als ein Beispiel dienen:

die Helligkeitswerte benachbarter Pixel können durchaus korreliert sein, was den "jagged" Verlauf der Koeffizienten für die Prädiktoren erklären würde. Hier kann die Regularisierung bzw Penalisierung helfen.

Typen von Analysen: Der erste Ansatz, die Einschränkungen der LDA zu überwinden, besteht darin, die LDA als Regressionsproblem aufzufassen und dabei von verallgemeinerten, nichtparametrischen Versionen der linearen Regression auszugehen. Dazu werden die Prädiktoren über *Basisentwicklungen* erweitert (analog zu den Support Vector Machines (SVM)).

Der zweite Ansatz zielt auf den Fall zu vieler Prädiktoren, wie sie bei der Diskretisierung von Bildern oder kontinuierlicher Signale entstehen. Die OCT-Profile sind ein solcher Fall. Hat das volle, 2-dimensionale Bild 190×190 Pixel, so hat man bereits 36100 Prädiktoren, und die Reduktion auf ein geeignet gewähltes Profil bedeutet immer noch 190 Prädiktoren. In diesem Fall läßt sich eine Lösung finden, indem man die entsprechenden Koeffizienten penalisiert und damit glätted. Dieser Ansatz führt zur *Penalisierten Diskriminanzanalyse* (PDA) als Spezialfall der *Flexiblen Diskriminanzanlyse* (FDA). Die PDA kann wiederum als regularisiertes Regressionsmodell im Rahmen der FDA betrachtet werden.

Der dritte Ansatz besteht darin, jede Klasse durch eine Mischung von zwei oder mehr Gaußschen Verteilungen mit verschiedenen Centroiden anzusehen. Dies führt zur *Mixture Discriminant Analysis* (MDA).

4.3.1 Flexible Diskriminanzanalyse

Der Regressionsansatz ist im Wesentlichen der des *Optimal Scoring*. Die Messungen (Reaktionen etc) mögen jeweils in eine von K Klassen fallen, $\mathcal{G} = (1, \ldots, K)$. Weiter sei $\theta : \mathcal{G} \to \mathbb{R}$ eine Abbildung, die jeder KJlasse einen Score zuordnet derart, dass die Klassen anhand einer linearen Regression anhand einer Messung optimal vorhergesagt werden können:

$$Q(\mathbf{b}, \theta) = \min_{\mathbf{b},\theta} \sum_{i=1}^{N} (\theta(g_i) - \mathbf{x}_i' \mathbf{b})^2, \tag{4.12}$$

wobei g_i die "Reaktionen" für den i-ten Fall sind und $\mathbf{x}_i$ die Prädiktorwerte für den i-ten Fall enthält. $\theta(g_i)$ ist der bestimmende optimale Score, und $\mathbf{b}$ ist ein zu bestimmender Vektor von Gewichten. Allgemein kann man bis zu $L \leq K - 1$ Sätze von Scores θ_ℓ finden, $l = 1, \ldots, L$, und dazu korrespondierende lineare Abbildungen $\mathbf{x}'\mathbf{b}_\ell$, so dass

$$\text{ASR} = \frac{1}{N} \sum_{\ell=1}^{L} \left[\sum_{i=1}^{N} (\theta_l(g_i) - \mathbf{x}_i' \mathbf{b}_\ell)^2 \right] \tag{4.13}$$

ASR steht für Average Squared Residuals. Die $\theta_1, \ldots, \theta_L$ sind paarweise orthogo-

Abbildung 18: Flexible DA - Scores

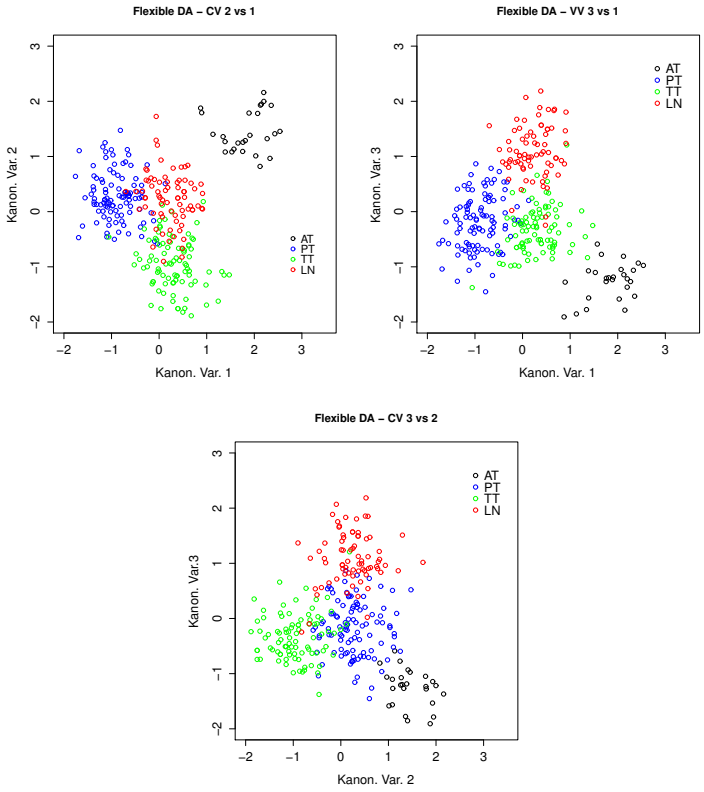
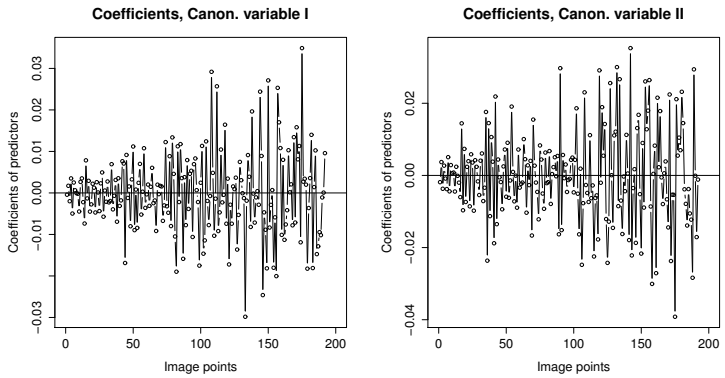


Abbildung 19: Flexible DA - Koeffizienten



nal und normalisiert. Es läßt sich zeigen, dass die Fisherschen Diskriminanzfunktionen bis auf eine Konstante mit den Regressionsgewichten $\mathbf{b}_\ell$ identisch sind (Hastie et al (1995), und die Mahalanobis-Distanz einer neuen Beobachtung $\mathbf{x}$ zum k-ten Klassencentroid $\hat{\mu}_k$ ist durch

$$\delta_J(\mathbf{x}, \hat{\mu}_k) = \sum_{\ell=1}^{K-1} w_\ell (\hat{\eta}_\ell(\mathbf{x}) - \bar{\eta}_\ell^k)^2 + D(\mathbf{x}), \quad \eta_\ell(\mathbf{x}) = \mathbf{x}'\mathbf{b}_\ell \tag{4.14}$$

gegeben, wobei $\bar{\eta}_\ell^k$ der Mittelwert der $\hat{\eta}_\ell(\mathbf{x}_i)$ für die k-te Klasse ist. w_ℓ sind die Gewichte der Koordinaten:

$$w_\ell = \frac{1}{r_\ell^2(1 - r_\ell^2)}. \tag{4.15}$$

Die $\eta_\ell(\mathbf{x})$ sind zunächst durch $\eta_\ell(\mathbf{x}) = \mathbf{x}'\mathbf{b}_\ell$ gegeben, können aber durch *flexible* alternative Größen definiert werden, wodurch die Verallgemeinerung der LDA erreicht wird, zB durch Spline-Funktionen. Man erhält den allgemeinen Ansatz

4.3.2 Penalisierte Diskriminanzanalyse

$$\mathrm{ASR}(\{\theta_\ell, \eta_\ell\}_{\ell=1}^L) = \frac{1}{N} \sum_{\ell=1}^{L} \left[\sum_{i=1}^{N} (\theta_\ell(g_i) - \eta_\ell(\mathbf{x}_i))^2 + \lambda J(\eta_\ell) \right] \tag{4.16}$$

(Hastie, Tibshirani, Friedman (2011), p. 441). J ist ein Regularisierungsterm, der durch Glättungs-Splines, additive Splines etc definiert werden kann. Eine ausführlichere Darstellung wird in Witten & Tibshirani (2011) gegeben.

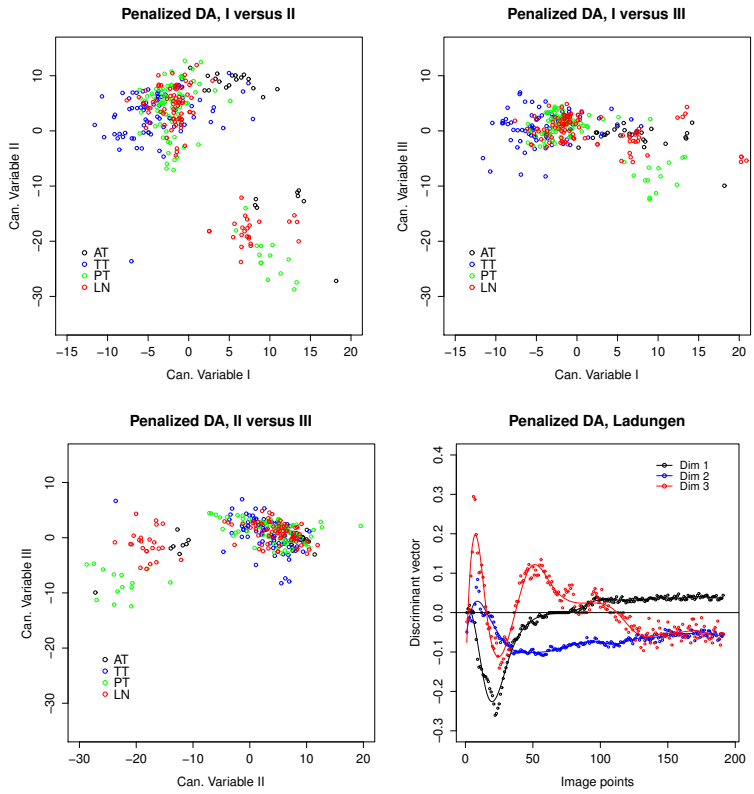
4.3.3 Mixture Diskriminanzanalysis

(Hastie & Tishirani (1996)) Es sei X_k die $(n_k \times p)$-Matrix der Messwerte auf den p Prädiktorvariablen bei den n_k Fällen der k-ten Gruppe, Klasse oder Kategorie. Es sei weiter $\mathbf{x}_k$ der p-dimensionale Vektor der Mittelwerte (gemittelt über die n_k Fälle) für die p Prädiktoren. $\mathbf{x}_k$ ist das *Centroid* für die k-te Gruppe. Das Centroid definiert einen *PRototyp* für die k-te Gruppe. Jede Gruppe (Klassse, Kategorie) kann durch ein Centroid beschrieben werden. Die Klassifikation eines Falles kann nach Maßgabe der Distanz zu den verschiedenen Centroiden vorgenommen werden, – es muß nur eine geeignete Metrik (DEfinition einer Distanz) gewählt werden. In der Fishershen LDA ist dies die durch die Mahalanobis-Distanz definierte Metrik.

Die Mixture DA ist wie die Mischung einre Wahrscheinlichkeitsfunktion definiert: sind $f_j(x)$, $j = 1, \ldots, n$ Wahrscheinlichkeitsdichten, so ist

$$f = a_1 f_1 + a_2 f_2 + \cdots + a_n f_n, \quad \sum_{j=1}^{n} a_j = 1$$

Abbildung 20: Penalized DA



eine Mischung der Dichten f_j. In Bezug auf die Diskriminanzanalyse ist die Mischung durch

$$P(\mathbf{x}|\mathcal{C}_k) = \sum_{r=1}^{R_k} \pi_{kr}\phi(\mathbf{x}; \mu_{kr}, \Sigma), \quad \sum_r \pi_{kr} = 1 \tag{4.17}$$

definiert. Die π_{kr} heißten auch *Mischungsanteile* (mixing proportions. Es gibt R_k Prototypen für die k-te Gruppe, d.h. es ist möglich, für jede Gruppe oder Klasse Untergruppen zu definieren. Σ ist eine Kovarianzmatrix, die für alle Gruppen geltgen soll. Für die *a posteriori*- Wahrscheinlichkeit für die Klasse $\mathcal{C}_k$ hat man dann

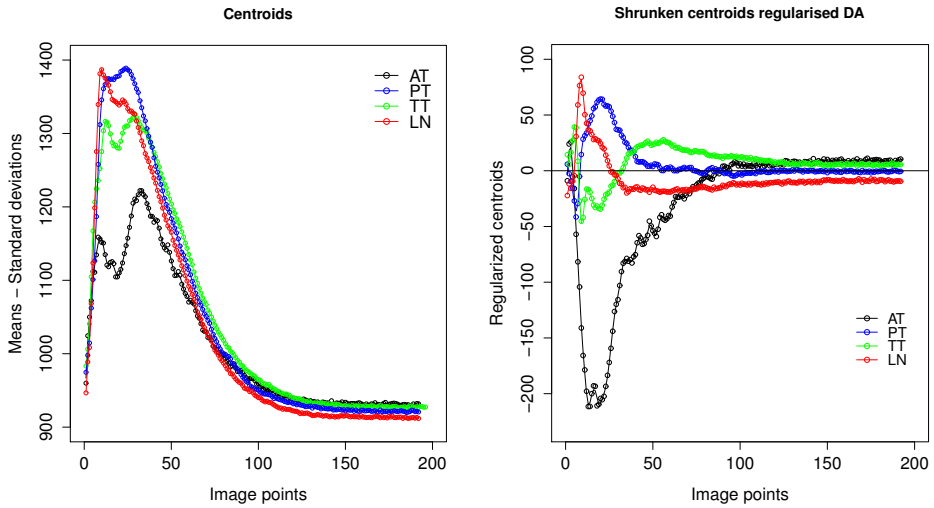
$$P(\mathcal{C}_k|\mathbf{x}) = \frac{\sum_{r=1}^{R_k} \pi_{kr}\phi(\mathbf{x}; \mu_{kr}, \Sigma)\Pi_k}{\sum_{\ell=1}^{K}\sum_{r=1}^{R_k} \pi_{\ell r}\phi(\mathbf{x}; \mu_{\ell r}, \Sigma)\Pi_\ell} \tag{4.18}$$

Π_k, Π_ℓ sind a priori-Wahrscheinlichkeiten für die Gruppen. Freie Parameter werden mit der Maximum-Likelihood-Methode geschätzt, – das ist eine aufwändige, numerische Angelegenheit (vergl. Hastie, Tibshirani, Friedman (2009). Das Verfahren ist in R mdamda verfügbar.

4.3.4 Shrunken Centroids Regularized DA (SCRDA)

Guo, Y. Hastie, T., Tibshirani, R. (2005)

Abbildung 21: Centroide (mittlere Profile) und Shrunken centroids (SCRDA-Schätzungen) der Koeffizienten ("Gewichte") der Prädiktoren



4.4 Partial Least Squares DA

Partial Least Squares (PLS) ist zunächst ein Regressionsverfahren, das der Kanonischen Korrelation ähnlich ist. Für m "Fälle" (Objekte, Personen, etc) liegen zwei Datensätze vor: eine $(m \times p)$-Matrix X und eine $(m \times q)$-Matrix Y. Die Aufgabe ist, die Y-Daten anhand der X-Daten vorherzusagen. Wie üblich sind Korrelationen zwischen den X-Variablen, ebenso wie zwischen den Y-Variablen, ein Problem. Bei der Kanonischen Korrelation werden sowohl für X wie auch für Y orthogonale Teilbasen der m-dimensionalen Vektoren bestimmt derart, dass die jeweils ersten Basisvektoren (korrespondieren zum größten Eigenwert) maximal miteinander korrelieren, ebenso die dazu orthogonalen Basisvektoren korrespondierend zum zweitgrößten Eigenwert, etc. Bei der PLS-Methode müssen die Basisvektoren nicht orthogonal sein und statt der Korrelationen können Kovarianzen verwendet werden.

4.4.1 Partial Least Squares (PLS)

Die PLS-Methode geht auf den norwegischen Statistiker Hermann Wold, der es in den 60-er Jahren als eine Art Verallgemeinerung der Multiplen Regression

entwickelte, wobei mehr als eine abhängige Variable vorhergesagt werden soll. Das Ziel war, das Problem der Kollinearitäten zwischen Prädiktoren zu überwinden. Dabei ergab sich eine Möglichkeit, den Fall von weniger Fällen als Prädiktoren zu behandeln. Das Verfahren wurde zunächst hauptsächlich in der Chemometrie angewandt. Mittlerweile gibt es eine ausgedehnte Literatur zu diesem Verfahren, zumal es eine Vielzahl von unterschiedlichen Versionen des Verfahrens gibt. In diesem Skript soll auf die Anwendung des Verfahrens auf diskriminanzanalytische Probleme fokussiert werden, so dass die folgende Darstellung sehr kurz gehalten ist.

Gegeben seien zwei Matrizen X (Prädiktoren, unabhängige Variablen) und Y (Kriteriumsvariablen, abhängige Variablen), wobei X eine (n, p)- und Y eine (n, q)-Matrix ist. Die verschiedenen unabhängigen Variablen entsprechen den Spalten von X, dementsprechend stehen die Spalten von Y für verschiedene abhängige Variablen. X und Y werden als zentriert angenommen, dh die Spalten haben jeweils den Mittelwert 0.

PLS wurde ursprünglich als Regressionsverfahren konzipiert: gesucht ist die optimale lineare Vorhersage von Y durch X. Andererseits dient PLS ebenfalls als Diskriminationsverfahren (Barker & Rayens (2003) liefern eine Reihe von Beispielen). Ist Y ein Vektor, so liegt ein Fall der multiplen Regression vor. Ein Problem sind jeweils mögliche Kollinearitäten in X (Korrelationen zwischen den unabhängigen Variablen, oder $n < p$, wenn also weniger gemessene Objekte als unabhängige Variablen gegeben sind). Die PCA-Regression, bei der die unabhängigen Variablen durch eine Auswahl der Hauptachsen (principal components) für X ersetzt werden, ist eine Möglichkeit, die Problematik von Kollinearitäten zu überwinden. Problematisch ist allerdings die Auswahl der Score-Vektoren (vergl. Joliffe (1986), Kap. 8), andererseits erklären die Score-Vektoren nur die Prädiktoren in X, nicht aber die Strukturen in Y.

Im Gegensatz dazu soll PLS eine simultane Zerlegung von X und Y liefern. In der bei der PLS üblichen Notation soll

$$X = TP' + E \tag{4.19}$$
$$Y = UQ' + F \tag{4.20}$$

gelten (Rosipal & Krämer (2006)). Sind $\mathbf{t}_j$ Spaltenvektoren von T und $\mathbf{u}_j$ die Spaltenvektoren von U, so soll

$$cov(\mathbf{t}_1, \mathbf{u}_1) \geq cov(\mathbf{t}_2, \mathbf{u}_2) \geq \cdots \geq cov(\mathbf{t}_r, \mathbf{u}_r), \quad r \leq \min(p, q) \tag{4.21}$$

sein, oder, in der Notation von Höskuldsson (1988) (Vorsicht: Definition von $\mathbf{r}$, $\mathbf{s}$ fehlt!),

$$|cov(\mathbf{t}, \mathbf{u})|^2 = |cov(X\mathbf{w}, Y\mathbf{c}|^2 = \max_{|r|=|s|=1} |cov(X\mathbf{r}, Y\mathbf{s})|^2, \tag{4.22}$$

und $cov(\mathbf{t}, \mathbf{u}) = \mathbf{t}'\mathbf{u}/n$ ist die Stichprobenkovarianz zwischen den Score-Vektoren $\mathbf{t}$ und $\mathbf{u}$. Das Problem ist die Abschätzung der latenten Variablen und deren

Anzahl. Die Lösung wird üblicherweise iterativ durch *nonlinear iterative partial least squares* (zB NIPALS-Algorithmus) gefunden. Die Approximationsfehler(matrizen) sind E und F. T und U enthalten *latente Variable* ("Scores"), während P und Q "Gewichte" enthalten. Der Ansatz erinnert an die Faktorenanalyse (FA) im Unterschied zur Hauptachsentransformation als Approximation der FA, bei der Fehlermatrizen nicht explizit betrachtet werden. PLS wurde ursprünglich als Iterationsverfahren zur Bestimmung der Matrizen T, P, U und Q definiert (Wold, 1966). Sind $\mathbf{t}_j$ Spaltenvektoren von T und $\mathbf{u}_j$ die Spaltenvektoren von U, so soll

$$cov(\mathbf{t}_1, \mathbf{u}_1) \geq cov(\mathbf{t}_2, \mathbf{u}_2) \geq \cdots \geq cov(\mathbf{t}_r, \mathbf{u}_r), \quad r \leq \min(p, q) \tag{4.23}$$

sein, oder, in der Notation von Höskuldsson (1988) (Vorsicht: Definition von $\mathbf{r}$, $\mathbf{s}$ fehlt!),

$$|cov(\mathbf{t}, \mathbf{u})|^2 = |cov(X\mathbf{w}, Y\mathbf{c}|^2 = \max_{|r|=|s|=1} |cov(X\mathbf{r}, Y\mathbf{s})|^2, \tag{4.24}$$

und $cov(\mathbf{t}, \mathbf{u}) = \mathbf{t}'\mathbf{u}/n$ ist die Stichprobenkovarianz zwischen den Score-Vektoren $\mathbf{t}$ und $\mathbf{u}$. Das Problem ist die Abschätzung der latenten Variablen und deren Anzahl. Die Lösung wird üblicherweise iterativ durch *nonlinear iterative partial least squares* (zB NIPALS-Algorithmus) gefunden:

$$\begin{array}{ll} 1)\, \mathbf{w} = X'\mathbf{u}/(\mathbf{u}'\mathbf{u}) & 4)\, \mathbf{c} = Y'\mathbf{t}/(\mathbf{t}'\mathbf{t}) \\ 2)\, \|\mathbf{w}\| \to 1 & 5)\, \|\mathbf{c}\| \to 1 \\ 3)\, \mathbf{t} = X\mathbf{w} & 6)\, \mathbf{u} = Y\mathbf{c} \end{array} \tag{4.25}$$

Höskuldson (1988) zeigt, dass der Vektor $\mathbf{w}$ der erste Eigenvektor von

$$X'YY'X\mathbf{w} = \lambda\mathbf{w} \tag{4.26}$$

ist, und

$$\mathbf{t} = X\mathbf{w}, \quad \mathbf{u} = Y\mathbf{c}. \tag{4.27}$$

Gleichwohl, PLS ist "ein iterativer Prozess" (Rosipal & Krämer (2006), p. 36). Nahc der Bestimmung von $\mathbf{t}$ und $\mathbf{u}$ werden die Matrizen X und Y deflationiert und gemäß (4.19) und (4.20) werden $\mathbf{p}$ und $\mathbf{q}$ berechnet:

$$\mathbf{p} = X'\mathbf{t}/(\mathbf{t}'\mathbf{t}), \quad \mathbf{q} = Y'\mathbf{u}/(\mathbf{u}'\mathbf{u}) \tag{4.28}$$

PLS als PLS1 ist definiert für den Fall, dass Y nur aus einer Spalte besteht, der Fall, dass Y eine Matrix ist, definiert PLS2. Dabei ist die Beziehung zwischen X und Y *asymmetrisch*: hier entsteht ein erster Unterschied zur Kanonischen Korrelation (CCA), die man ja wegen (4.19) und (4.20) mit PLS assoziieren kann. Zwischen den in (4.19) und (4.20) eingeführten Matrizen T und U soll die Beziehung

$$U = TD + H \tag{4.29}$$

bestehen, wobei D eine $(p \times p)$-Diagonalmatrix und H eine Matrix mit Residuen ist. Satt über die Gleichungen (4.19) und (4.20) kann man deshalb PLS auch gemäß

$$Y = TQ' + F \tag{4.30}$$
$$X = TP' + E \tag{4.31}$$

einführen; Q ist eine $(q \times r)$-Matrix von Koeffizienten, und P ist eine $(p \times r)$-Matrix von Koeffizienten. E ($(m \times p)$-Matrix) und F ($(m \times q)$-Matrix) enthalten zufällige Fehler (Boulesteix & Strimmer (2006)). T heißt 'Score'-Matrix und ist eine Matrix, deren Spaltenvektoren als latente Vektoren aufgefasst werden können. T ist als lineare Transformation von X darstellbar:

$$T = XW. \tag{4.32}$$

W wird interativ bestimmt (s. unten). Hat man T gewonnen, so ergibt sich die Matrix Q als Kleinste-Quadrate-Lösung

$$\hat{Q}' = (T'T)^{-1}T'Y \tag{4.33}$$

Man betrachtet das Modell

$$Y = XWQ' + F = BX + F \tag{4.34}$$

mit

$$B = WQ' = W(T'T)^{-1}T'Y$$

und für die Schätzung $\hat{Y}$ erhält man

$$\hat{Y} = T(T'T)^{-1}T'Y. \tag{4.35}$$

Die Matrizen B, W, T, P und Q werden iterativ und simultan geschätzt; dieser Sachverhalt erklärt den Ausdruck 'Partial Least Squares'. P und Q heißen auch X-Ladungen bzw Y-Ladungen. Diese Bezeichnungen werden angemerkt, weil sie in den PLS-Programmen (in R) auftreten.

In der "normalen" multiplen Regression, einschließlich der PCA-Regression, werden die Parameter zur Vorhersage von Y nur aus den Prädiktorwerten (X) berechnet. PLS unterscheidet sich von diesem Ansatz dadurch, dass bei der Parameterschätzung die Y-Werte mit einbezogen werden, wie aus (4.33) und (4.35) hervorgeht. Wenn Voraussagen gemacht werden sollen, funktioniert PSL deshalb besser als die PCA. Man erinnere sich: CCA setzt vorauss (vergl. (1.71)), dass die Inversen R_{xx}^{-1} und R_{yy}^{-1} müssen, damit eine CCA gerechnet werden kann. PLS hält dagegen einen Ausweg für den Fall, dass diese Inversen nicht existieren parat.

Die Beziehung zwischen PLS und CCA läßt sich auch auf andere Weise charakterisieren (Indahl, Liland, Næs (2009)). Die Matrizen X und Y seien zentriert.

Dann repräsentiert das Produkt $X'Y$ die Kovarianzen zwischen X und Y. Gesucht werden die Vektorten $\mathbf{u}$ undf $\mathbf{v}$ derart, dass

$$f_1(\mathbf{u}, \mathbf{v}) = \mathbf{u}'X'Y\mathbf{v} = \max. \tag{4.36}$$

$W = X'Y$ ist eine $(p \times q)$-Matrix. Es sei $Q\Lambda^{1/2}P'$ die Singularwertzerlegung von W. Q ist die $(p \times r)$-Matrix der Eigenvektoren von WW', P ist die $(q \times r)$-Matrix der Eigenvektoren von $W'W$ und Λ die die $(r \times r)$-Diagonalmatrix von Eigenwerten ungleich Null von WW' bzw. $W'W$. Die Eigenwerte seien der Größe nach geordnet. Dann gilt

$$\max(\mathbf{u}'W\mathbf{v}) = \mathbf{q}_1'W\mathbf{p}_1 = \lambda_1^{1/2} \tag{4.37}$$

Der zweitgrößte Wert ist durch

$$\mathbf{q}_2'W\mathbf{p}_2 = \lambda_2^{1/2} \tag{4.38}$$

gegeben, etc., und

$$f_1^{(r_0)} = Q_{r_0}\Lambda_{r_0}^{1/2}P_{r_0}, \quad r_0 < r \tag{4.39}$$

ist die r_0-Approximation von f_1.

Beziehung zwischen PLS und CCA: Es ist

$$\frac{1}{m}(\mathbf{u}'W\mathbf{v}) = Kov(X\mathbf{u}, Y\mathbf{v}). \tag{4.40}$$

Beschränkt man die PLS-Lösung auf orgthogonale Vektoren so liefert die Singularwertzerlegung die Vektoren $\mathbf{u}$ und $\mathbf{v}$, für die die Kovarianz von $X\mathbf{u}$ und $Y\mathbf{v}$ maximal wird. CCA geht von einem ähnlichen Ansatz aus: es sollen Vektoren $\mathbf{u}$, $\mathbf{v}$ gefunden werden, für die

$$\text{corr}(X\mathbf{u}, Y\mathbf{v}) = \frac{\mathbf{u}'X'Y\mathbf{v}}{\sqrt{\mathbf{u}'X'X\mathbf{u}}\sqrt{\mathbf{v}'Y'Y\mathbf{v}}} \tag{4.41}$$

maximal wird. Das Maximum wird erreicht, wenn man für $\mathbf{u}$ und $\mathbf{v}$ die Eigenvektoren aus der Singularwertzerlegung von $(X'X)^{-1/2}X'Y(Y'Y)^{-1/2}$ wählt, die zum maximalen Eigenwert korrespondieren.

Im Falle von Multikollinearität können sich Probleme für die Interpretation ergeben, desgleichen, wenn es (bei geringen Fallzahlen) zu einem overfitting kommt. In diesem Fall können Penalisierungstechniken eine verbesserte Analyse bedeuten (Waaijenborg & Zwinderman (2007)), worauf hier aber nicht im Detail eingegangen werden kann.

4.4.2 PLS und Diskrimination

Die Anwendung von PLS auf diskriminanzanalytische Fragen wurde von Barker & Rayens (2003) diskutiert. Es soll die Funktion

$$f_2(\mathbf{u}, \mathbf{v}) = \mathbf{u}'X'Y(Y'Y)^{-1/2}\mathbf{v} = \mathbf{u}'W_\Delta\mathbf{v} \tag{4.42}$$

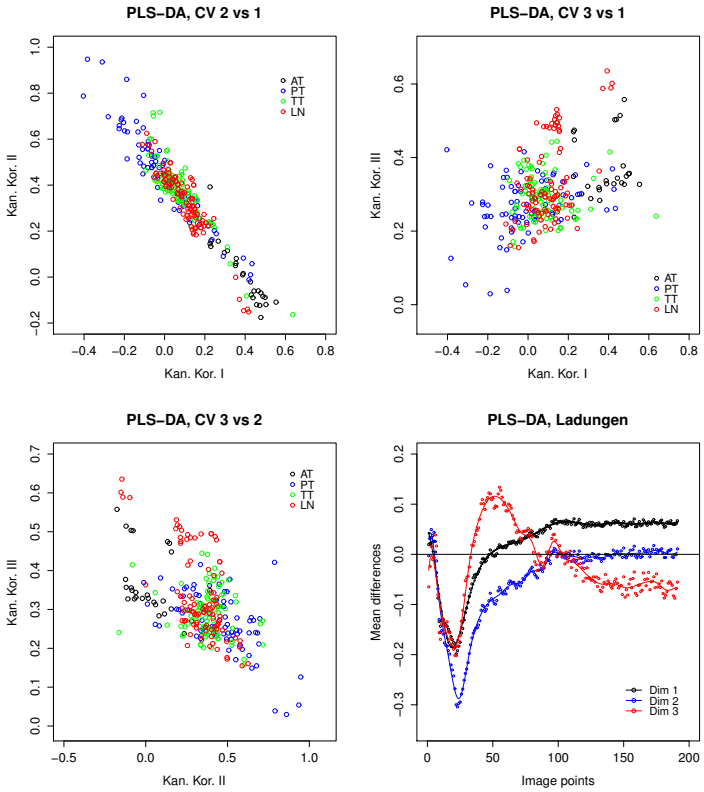
maximiert werden.

PLS kann als penalisierte Kanonische Korrelation gesehen werden, wobei eine PCA des X-Raums und eine PCA des Y-Raums als Penalisierung betrachtet werden können. Ist nun das Ziel eine Diskriminierung der Y-Werte und ist Y eine Dummy-Kodierung von Variablen, so ist die Y-Penalisierung nicht sinnvoll. Man erhält dann

$$(\text{corr}(\mathbf{a}'X, \mathbf{b}'Y))^2 Var(\mathbf{a}'X) = \frac{(Kov(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}))^2}{Var(\mathbf{b}'\mathbf{y})} \tag{4.43}$$

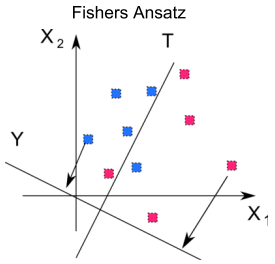
Die linke Seite definiert eine *objective function*, d.h. eine Funktion, deren Wert maximiert werden soll. Sie wird maximal, wenn $\mathbf{a}$ und $\mathbf{b}$ der Reihe nach die Die Eigenvektoren von $\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx}$ annehmen, also für die Paare $\{\mathbf{a}_{k+1}\}$ mit $\mathbf{b}_{k+1} = \Sigma_y^{-1}\Sigma_{yx}\mathbf{a}_{k+1}$. Weitere Details findet man in Barker & Rayens (2003).

Abbildung 22: PLS-DA

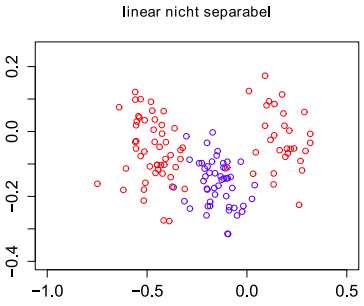


4.5 Support Vector Machines (SVMs)

4.5.1 Der Ansatz



In Fishers Ansatz werden zwei Populationen – hier ein wenig optimistisch, weil so gut wie überlappungsfrei vorgestellt – durch eine Ebene T (hier: Gerade) getrennt. Eine neue Beobachtung wird je nach ihrer Position relativ zu dieser Geraden der einen oder der anderen Population zugeordnet. Die Gerade wird bestimmt, indem die Projektionen der Punkte auf eine zu T orthogonale Gerade Y betrachtet werden: die Projektionen der verschiedenen Gruppen auf Y sollen soweit nur irgend möglich separiert werden. Dieses Kriterium legt die Orientierung von Y und damit von T fest. In die Bestimmung von Y gehen alle Punkte ein.



Die Support Vector Machines (SVMs) gehen zunächst ebenfalls von einer linearen Trennbarkeit der Gruppen aus, bestimmen aber die trennenden Hyperebenen nicht durch Einbeziehung aller Punkte, sondern nur durch Punkte, die auf einer der beiden zur eigentlichen Trennebene parallelen Ebenen ("Margins") liegen; diese Punkte sind die *support vectors*. Darüber hinaus können nichtlineare Probleme gelöst werden, indem die die Punkte definierenden Vektoren $\mathbf{x}$ nichtlinear transformiert werden, $\mathbf{x} \mapsto \phi(\mathbf{x})$, womit üblicherweise eine Erhöhung der Dimensionalität verbunden ist. In dem durch die ϕ definierten Raum werden dann wieder linere Trennebenen bestimmt. Diese Trennebenen entsprechen im ursprünglichen $\mathbf{x}$-Raum nichtlinearen Trennebenen. Dadurch wird es möglich, Fälle wie der Abbildung 1 (nebenstehend noch einmal präsentiert) zu behandeln. Das Prinzip einer SVM wird zunächst für den linear trennbaren Fall erläutert werden, der nichtlineare Fall ergibt sich dann aus diesem Ansatz. Es wird nicht nur eine trennende Hyperebene – eine Gerade im 2-dimensionalen Fall – gesucht, sondern eine Art Einbettung der Geraden zwischen zwei zu ihr parallelen Geraden, in der der einfache Fall von nur zwei Kategorien betrachtet wird. Diese zwei Parallelen definieren den *Margin*, also den "Rand" für G. Für einen gegebenen Datenpunkt $\mathbf{x}_i$ wird eine Funktion f gesucht derart, dass

$$y_i = f(\mathbf{x}_i) = \begin{cases} \geq +1, & \mathbf{x}_i \in \mathcal{C} \\ \leq -1, & \mathbf{x}_i \in \mathcal{C}^c \end{cases} \tag{4.44}$$

$$\mathbf{x} \mapsto f_{\mathbf{w},b}(\mathbf{x}) = \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b) \tag{4.45}$$

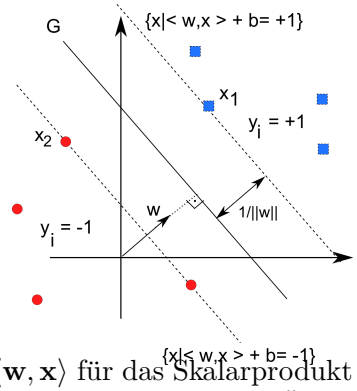
gilt; sgn steht für *signum*, soll heißen für die Vorzeichen + bzw. - ; man kann auch die Faktoren $+1$ bzw. -1 damit bezeichnen. Die Gerade G und die zu ihr parallelen Geraden werden nur durch Punkte bestimmt, die auf den parallelen Geraden liegen, wie $\mathbf{x}_1$ und $\mathbf{x}_2$. Die Frage ist nun, wie diese Geraden bestimmt werden können.

Eine Trennebene wird stets durch eine lineare Gleichung

$$w_1x_1 + w_2x_2 + \cdots + w_px_p + b = 0$$

beschrieben. Eine Geradengleichung ergibt sich für $p = 2$. Es sei $\mathbf{w} = (w_1, \ldots, w_p)'$; die Ebenen- bzw Geradengleichung für G ist dann durch

$$\mathbf{w}'\mathbf{x} + b = \langle \mathbf{w}, \mathbf{x} \rangle + b = 0 \tag{4.46}$$



gegeben, wobei b der Abstand der Ebene vom Nullpunkt des Koordinatensystems ist; die Schreibweise $\langle \mathbf{w}, \mathbf{x} \rangle$ für das Skalarprodukt ist in Texten zu SVMs üblich und wird deshalb hier übernommen, um den Übergang zur Originalliteratur zu erleichtern.

Der Vektor $\mathbf{w}$ ist orthogonal zur Orientierung der Ebene. Um dies zu sehen, betrachte man zwei Punkte $\mathbf{X}_a$ und $\mathbf{x}_b$ in der Ebene (auf der Geraden); hier wird von der Vektorschreibweise für die Punkte Gebrauch gemacht, weil man ja die Punkte als Endpunkte von Vektoren mit gemeinsamem Anfangspunkt im Ursprung des Koordinatensystems betrachten kann. Die Differenz $\mathbf{x}_a - \mathbf{x}_b$ ist dann ein Vektor, der in der Ebene liegt. Es muß dann $\langle \mathbf{w}, \mathbf{x}_a \rangle + b = \langle \mathbf{w}, \mathbf{x}_b \rangle + b$ gelten, woraus

$$\langle \mathbf{w}, \mathbf{x}_a \rangle - \langle \mathbf{w}, \mathbf{x}_b \rangle = \langle \mathbf{w}, \mathbf{x}_a - \mathbf{x}_b \rangle = 0$$

folgt, und dies bedeutet eben, dass $\mathbf{w}$ orthogonal zur Ebene $\langle \mathbf{w}, \mathbf{x} \rangle + b = 0$ ist. Die Parallellen zu G werden durch die Forderung $\min_i \langle \mathbf{w}, \mathbf{x}_i \rangle + b = \pm 1$ definiert, d.h. durch

$$\min_i |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = 1 \tag{4.47}$$

definiert. Damit wird gesagt, dass der Punkt mit dem kleinsten Abstand zu G den Abstand $1/\|\mathbf{w}\|$ hat. Dass diese Aussage gilt, ist leicht zu sehen. Man betrachte zwei Punkte $\mathbf{x}_1$ und $\mathbf{x}_2$ mit $\langle \mathbf{w}, \mathbf{x}_1 \rangle + b = +1$ und $\langle \mathbf{w}, \mathbf{x}_2 \rangle + b = -1$, d.h. $|\langle \mathbf{w}, \mathbf{x} \rangle + b| = 1$. Subtrahiert man die zweite Gleichung von der ersten, so erhält man

$$\langle \mathbf{w}, (\mathbf{x}_1 - \mathbf{x}_2) \rangle = 2.$$

Normiert man den Vektor $\mathbf{w}$, d.h. geht man von $\mathbf{w}$ zu $\mathbf{w}/\|\mathbf{w}\|$ über, so erhält man

$$\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, \mathbf{x} \rangle = \frac{2}{\|\mathbf{w}\|}. \tag{4.48}$$

Dies bedeutet, dass der Abstand eines Margins zu G gerade gleich $1/\|\mathbf{w}\|$ ist.

Die Gerade G separiert optimal, wenn der Margin-Abstand maximal ist. Dies ist der Fall, wenn $1/\|\mathbf{w}\|$ maximal, d.h. wenn $\|\mathbf{w}\|$ minimal ist. $\mathbf{w}$ kann also duch Minimalisierung von $\|\mathbf{w}\|$ bestimmt werden, wobei allerdings Nebenmbedingungen erfüllt sein müssen: es soll ja $|\langle \mathbf{w}, \mathbf{x}\rangle + b| \geq 1$ gelten. Diese Forderung kann man in der Form $y_i(\langle \mathbf{w}, \mathbf{x}\rangle + b) \geq 1$ schreiben: ist $\langle \mathbf{w}, \mathbf{x}\rangle + b > 0$, so ist auch $y_i > 0$ und ist $\langle \mathbf{w}, \mathbf{x}\rangle + b < 0$, so ist auch $y_i < 0$ udn das Produkt der beiden Größen ist positiv. Damit besteht die Aufgabe der Bestimmung von $\mathbf{w}$ also darin, $\mathbf{w}$ gemäß

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{H}, b\in\mathbb{R}} \tau(\mathbf{w}) &= \frac{1}{2}\|\mathbf{w}\|^2 && (4.49) \\ y_i(\langle \mathbf{w}, \mathbf{x}\rangle + b) &\geq 1, \text{ für alle } i && (4.50) \end{aligned}$$

zu bestimmen; der Faktor $1/2$ in (4.49) hat den Sinn, den Faktor 2 loszuwerden, der beim Differenzieren von $\tau(\mathbf{w})$ entsteht. Die Aufgabe wird gelöst, indem man die den Gleichungen (4.49) und (4.50) enstsprechende Lagrange-Funktion

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w}\|^2 = \sum_{i=1}^{m} \alpha_i (y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1) \tag{4.51}$$

betrachtet, wobei $\boldsymbol{\alpha}(\alpha_1, \ldots, \alpha_m)'$ der Vektor der Lagrange-Multiplikatoren ist mit $\alpha_i \geq 0$. $L(\mathbf{w}, b, \boldsymbol{\alpha})$ muß bezüglich der α_i maximalisiert und bezüglich $\mathbf{w}$ minimalisiert werden. Dazu muß

$$\frac{\partial L}{\partial b} L(\mathbf{w}, b, \boldsymbol{\alpha}) = 0, \quad \frac{\partial L}{\partial \mathbf{w}} = 0 \tag{4.52}$$

gelten, woraus

$$\sum_{i=1}^{m} \alpha_i y_i = 0, \quad \mathbf{w} = \sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \tag{4.53}$$

folgt (Schölkopf und Smola (2002), p. 197). Der gesuchte Vektor $\mathbf{w}$ ergibt sich demnach als Linearkombination der Vektoren $\mathbf{x}_i$, also aus den Daten der Trainingsstichprobe. Es läßt sich zeigen, dass nur diejenigen $\alpha_i \neq 0$ den Bedingungen von (4.49) und (4.50) genügen, für die

$$\alpha_i (y_i(\langle \mathbf{w}, \mathbf{x}_i\rangle + b) - 1) = 0 \tag{4.54}$$

gilt. Diejenigen $\mathbf{x}_i$ mit $\alpha_i > 0$ sind die *Support Vectors* (SVs)[12], denn (4.54) bedeutet ja gerade, dass diese $\mathbf{x}_i$ exakt auf dem Rand (Margin), also auf den Parallelen zu G liegen. Alle übrigen $\mathbf{x}_j$ sind irrelevant, da für sie $\alpha_i = 0$. Die Bedingungen (4.50) sind automatisch erfüllt.

Um das eigentliche Optimierungsproblem zu lösen, müssen die Bedingungen von (4.53) in die Lagrange-Funktion (4.51) eingesetzt werden. Man erhält

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} W(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle \tag{4.55}$$

[12]Man sollte vielleicht besser von Stützvektoren reden, um das unselige Denglisch zu vermeiden. Andererseits ist der englische Ausdruck zum Standardausdruck geworden, so dass er hier beibehalten wird

unter den Nebenbedingungen

$$\alpha_i \geq 0, \quad \sum_{i=1}^{m} \alpha_i y_i = 0. \tag{4.56}$$

Setzt man nun die Definition von $\mathbf{w}$ in die Entscheidungsfunktion (4.45) ein, so erhält man für f den Ausdruck

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{m} \alpha_i y_i \langle \mathbf{x}, \mathbf{x}_i \rangle + b\right) \tag{4.57}$$

Damit wird deutlich, dass $f(\mathbf{x})$ u.a. durch die Skalarprodukte $\langle \mathbf{x}_i, \mathbf{x} \rangle$ definiert wird.

Dieser Sachverhalt ermöglicht den Übergang zu nichtlinearen Trennfunktionen, denn $\langle \mathbf{x}_i, \mathbf{x} \rangle$ ist der Spezialfall einer allgemeinen *Kernfunktion* $k(\mathbf{x}_i, \mathbf{x})$. Der Begriff der Kernfunktion soll zunächst allgemein erläutert werden.

4.5.2 Kernfunktionen

Der Begriff der Kernfunktion wurde im Zusammenhang mit der Diskussion von Lösungen von linearen Integralgleichungen eingeführt. Dies sind Gleichungen der Form

$$f(s) = \phi(s) - \lambda \int k(s,t)\phi(t)dt, \tag{4.58}$$

deren Lösung eine Funktion ϕ sein soll. $k(s,t)$ ist eine Kernfunktion, und f ist vorgegeben. für $f \equiv 0$ entsteht eine homogene Integralgleichung. Auf Integralgleichungen muß im Folgenden nicht eingegangen werden, es genügt der Hinweis, dass Lösungen über Reihenentwicklungen der Kernfunktion möglich sind (vergl. z.B. Courant & Hilbert (1968)). Allerdings sind die Reihenentwicklungen für die Anwendung von Kernfunktionen auf Klassifikationsfragen von Bedeutung.

Da $k(s,t)$ eine Funktion ist, legt die Idee der Reihenentwicklung von $k(s,t)$ nahe, dass k als Summe bzw. Linearkombination von Funktionen dargestellt werden soll. Tatsächlich hat man es mit einer Verallgemeinerung des aus der linearen Algebra bekannten Begriffs der Linearkombination zu tun, wo etwa ein n-dimensionaler Vektor $\mathbf{x}$ in der Form

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + \cdots + x_n\mathbf{e}_n$$

dargestellt werden kann, wobei die $\mathbf{e}_j$ Einheitsvektoren sind, deren Komponenten alle gleich Null sind bis auf die j-te, die gleich 1 ist. Die $\mathbf{e}_j$ sind paarweise orthogonal. d.h. es gilt

$$\langle \mathbf{e}_j, \mathbf{e}_k \rangle = \delta_{jk} = \begin{cases} 1, & j = k \\ 0, & j \neq k \end{cases}$$

woraus sofort

$$x_j = \langle \mathbf{x}, \mathbf{e}_j \rangle \tag{4.59}$$

folgt. δ_{jk} heißt Kronecker-delta, nach dem Mathematiker Leopold Kronecker (1823 – 1891), der diese Schreibweise als erster einführte.

Nun seien f und g zwei auf einem Intervall (a, b) definierte, integrierbare Funktionen. Für sie läßt sich das Skalarprodukt

$$\langle f, g \rangle = \int_a^b f \bar{g} dx \tag{4.60}$$

definieren, wobei $\bar{g}$ die zu g konjugiert komplexe Funktion gemeint ist, falls g komplex ist; ist g reell, so ist $\bar{g} = g$. Analog zu den orthogonalen und normierten (orthonormalen) Vektoren $\mathbf{e}_j$ lassen sich Funktionen ϕ_j, ϕ_k finden derart, dass

$$\langle \phi_j(x), \phi_k(x) \rangle = \delta_{jk} \tag{4.61}$$

gilt, wobei das Skalarprodukt wie in (4.60) als Integral definiert ist und $(a, b) = (-\infty, \infty)$ sein kann. Ist f wieder eine auf einem Intervall (a, b) definierte Funktion, so kann es gelingen, f als Linearkombination von Funktionen eines Orthonormalsystems darzustellen:

$$f(x) = \sum_{j=1}^{\infty} a_j \phi_j(x), \quad \sum_{j=1}^{\infty} |a_j|^2 < \infty \tag{4.62}$$

Es soll also möglich sein, dass die Summe rechts unendlich viele Terme enthält. Die Summe soll für alle $x \in (a, b)$ konvergieren, d.h. die Reihe rechts soll gleichmäßig konvergent sein. Analog zu (4.59) gilt

$$a_j = \langle f(x), \phi_j(x) \rangle \tag{4.63}$$

Für Anwendungen im Bereich der Mustererkennung und -klassifikation interessiert man sich für solche Orthonormalsysteme, für die eine Kernfunktion existiert derart, dass

$$k(s, t) = \sum_{j=1}^{\infty} \phi_j(s) \bar{\phi}_j(t) \tag{4.64}$$

gilt, wobei $\bar{\phi}_j$ die zu ϕ_j konjugiert komplexe Funktion ist; sind die Funktionen ϕ_j reell, so gilt wieder $\bar{\phi}_j = \phi_j$. Es wird natürlich unterstellt, dass die Reihe rechts konvergiert. Nicht alle Orthonormalsystem konvergieren in diesem Sinne: ein Beispiel sind die Funktionen $\phi_j = (\sin(jx))/\sqrt{\pi}$, die zwar orthonomal sind, für die aber die rechte Seite, insbesondere

$$k(\pi/2, \pi/2) = \frac{1}{\pi} \sum_{j=1}^{\infty} \sin^2 j\pi/2$$

von (4.64) divergiert.

Reproduktion: Es sei f in eine gleichmäßig konvergente Reihe der Form (4.62) darstellbar. Dann gilt

$$f(u) = \langle f(x), k(x, \bar{u})\rangle = \int_a^b f(x)\overline{k(x, \bar{u})}dx. \tag{4.65}$$

$\bar{u}$ etc bedeutet wieder die zur jeweiligen Größe konjugiert komplexe Größe.

Beweis: Wegen der vorausgesetzten Orthonormalität der ϕ_j gilt

$$\begin{aligned}\langle f(x), k(x, \bar{u})\rangle &= \left\langle \sum_{j=1}^{\infty} a_j\phi_j(x), \sum_{j=1}^{\infty} \phi_j(x)\overline{\phi_j(x)}\rangle \right\rangle \\ &= \sum_{j,k}\langle a_j\phi_j(x), \phi_k(x)\overline{\phi_k(u)}\rangle \\ &= \sum_{j=1}^{\infty} a_j\phi_j(u) = f(u)\end{aligned}$$

□

Die Gleichung (4.65) definiert die *reproduzierende* Eigenschaft einer Kernfunktion.

Innere Produkträume:

Hilbert-Räume: Ein Hilbert-Raum ist ein Vektorraum, in dem ein Skalarprodukt erklärt ist. Sind die Vektoren Funktionen, so ist ein Hilbert-Raum ein Raum von Funktionen über einem Intervall derart, dass Linearkombinationen von Elementen des Raums wieder Elemente des Raums sind und für die das Integral (4.60) existiert, – für die also paarweise die korrespondierenden Skalarprodukte existieren. Funktionen, für die eine Darstellung der Form (4.62) existiert, bilden einen Hilbert-Raum (um dies zu sehen, muß man nur nachweisen, dass Linearkombinationen derart definierter Funktionen wieder in der Form (4.62) darstellbar sind und dass das jeweilige Skalarprodukt existiert). Ein Hilbert-Raum heißt *separierbar*, wenn die ϕ_j abzählbar sind. Ist umgekehrt ein Hilbert-Raum $\mathcal{H}$ separierbar, so kann man folgern, dass jedes Orthonormalsystem von $\mathcal{H}$ aus höchsten abzählbar vielen Elementen besteht (vergl. Meschkowski (1962), p. 17).


David Hilbert, (1862 - 1943)

Hilbert-Räume, deren Elemente die Bedingung (4.65) erfüllen, heißen *reproduzierende Kern-Hilberträume* (reproducing kernel Hilbert spaces (RKHS).

Merkmalsräume: $\mathbf{x}$ sei ein p-dimensionaler Vektor von Prädiktorvariablen. Die Komponenten definieren Werte von Merkmalen, und insofern definiert $\mathbf{x}$ einen Punkt in einem Merkmalsraum χ, wenn der Anfangspunkt von $\mathbf{x}$ in den Ursprung des Koordinatensystems gelegt wird. Es ist aber möglich, dass eine gute

Klassifikation von Mustern von Funktionen der Komponenten ϕ von $\mathbf{x}$ abhängt. Die Kernfunktion $k(\mathbf{x}, \mathbf{x}_j)$ wird nun über die $\phi(\mathbf{x}), \phi(\mathbf{x}_j)$ defininiert.

Beispiel 4.1 Es sei $\mathbf{x} = (x_1, x_2)'$ und

$$\phi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2).$$

$\phi(\mathbf{x})$ ist also ein Vektor, dessen Komponenten nichtlineare Funktionen der Komponenten von $\mathbf{x}$ sind, und darüber hinaus ist die Anzahl der Komponenten von $\phi(\mathbf{x})$ größer als die von $\mathbf{x}$. Das zu klassifizierende Muster wird demnach in einem Raum von höherer Dimension als p repräsentiert, dessen Koordinaten den Komponenten von $\phi(\mathbf{x})$ entsprechen. Weiter sei $k(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle$; dann hat man

$$\begin{aligned} \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle &= \langle (x_{11}^2, x_{21}^2, \sqrt{2}x_{11}x_{21}), (x_{21}^2, x_{22}^2, \sqrt{2}x_{12}x_{22}) \rangle \\ &= x_{11}^2x_{12}^2 + x_{21}^2x_{22}^2 + 2x_{11}x_{12}x_{21}x_{22} \\ &= (x_{11}x_{21} + x_{21}x_{22})^2 = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle^2. \end{aligned} \tag{4.66}$$

Das heißt,

$$k(\mathbf{x}_1, \mathbf{x}_2) = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle^2 \tag{4.67}$$

ist eine Kernfunktion. Das Bemerkenswerte am Ergebnis (4.66) ist, dass die nichtlineare Transformation $\phi(\mathbf{x})$ gar nicht explizit durchgeführt werden muß, – es genügt, das Skalarprodukt $\langle \mathbf{x}_1, \mathbf{x}_2 \rangle$ zu quadrieren. □

Kernel-Trick: Der Raum, in dem die Vektoren $\mathbf{x}$ liegen, wird auch Input-Raum (input space) genannt. Die Abbildung $\phi(\mathbf{x})$ bildet $\mathbf{x}$ in einen höherdimensionalen Merkmalsraum (feature space) ab. Innerhalb des Merkmalsraums werden lineare Trennebenen bestimmt. Die Rückprojektion in den Input-Raum bedeutet, dass in diesem Raum die Trennflächen nichtlinear sind. Die Transformation $\mathbf{x} \to \phi(\mathbf{x})$ muß allerdings gar nicht explizit durchgeführt werden, da nur mit der Kernfunktion k operiert werden muß. Dieser Sachverhalt definiert den *Kernel-Trick*.

Die Hyperebenen, die bestimmt werden, genügen also der Gleichung

$$\langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b = 0, \tag{4.68}$$

und die Entscheidungsfunktion ist

$$f(\mathbf{x}) = \operatorname{sgn}(\langle \mathbf{w}, \phi(\mathbf{x}) \rangle + b). \tag{4.69}$$

$\mathbf{w}$ ist durch

$$\mathbf{w} = \sum_i \alpha_i \phi(\mathbf{x}_i). \tag{4.70}$$

Der Kernel-Trick impliziert, dass die Entscheidungsfunktion am Ende nur vom Skalarprodukt zwischen den Mustern abhängt. Insbesondere für die "soft-margin-Klassifikation hat man

$$L(\mathbf{w}, \xi) = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m}\xi_i \tag{4.71}$$

$$y_i(\langle\phi(\mathbf{x}_i), \mathbf{w}\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, m \tag{4.72}$$

wobei (4.72) die Bedingung für (4.71) definiert und $\geq 1 - \xi_i$ die Soft-Margin-Bedingung definiert. m ist die Anzahl der Trainingsmuster, und $y_i = \pm 1$. Der Parameter C in (4.71) definiert die "Kosten". Weiter kann gezeigt werden, dass

$$\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \phi(\mathbf{x}_i). \tag{4.73}$$

Die $\alpha_i \neq 0$ definieren die Support Vectors, und dies ist der Fall, wenn $(\mathbf{x}_i, y_i)$ den Bedingungen (4.72) entspricht. Die α_i werden bestimmt, in dem das QP-Problem[13]

$$\max_{\alpha} W(\boldsymbol{\alpha}) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j k(\mathbf{x}, \mathbf{x}_j) \tag{4.74}$$

$$0 \leq \alpha_i \leq \frac{C}{m}, \ \sum_{i=1}^{m}\alpha_i y_i = 0, \quad i = 1, \ldots, m \tag{4.75}$$

C definiert, wie oben bereits angemerkt, den Strafterm für Fehlklassifikationen. Ein hoher Wert von C impliziert hohe Kosten für Fehlklassifikationen und erzwingt deshalb eine komplexere Vorhersagefunktion, die die Anzahl der Fehler so klein wie jeweils möglich macht. Man spricht deshalb auch von *C-SVMs*.

Eine Alternative zu C-SVMs sind *ν-SVMs*. ν definiert einen oberen Wert für den Trainingsfehler und einen unteren Wert für den Anteil der Support Vectors.

Typen von Kernfunktionen Kernfunktionen bilden eine Klasse, die durch die Mercer-Bedingungen charakterisiert sind.

Definition 4.2 *Mercer-Bedingungen hier*

Üblich sind die folgenden Typen:

Kernfunktion	Formel	Parameter
linear	$\mathbf{x}'\mathbf{y}$	keine
Polynomial	$\gamma(\mathbf{x}'\mathbf{y} + b)^p$	γ, b, p
Radial Basis Function (RBF)	$\exp(-\gamma\|\mathbf{x} - \mathbf{y}\|^2)$	γ
Sigmoid	$\tanh(\gamma\mathbf{x}'\mathbf{y} + b)$	γ, b

(4.76)

[13] QP = Quadratisches Programmieren; $\mathbf{c}$ sei ein n-dimensionaler Vektor, $\mathbf{b}$ sei ein m-dimensionaler Vektor, A sei eine $(m \times n)$-Matrix, Q sei eine symmetrische $(n \times n)$-Matrix. Das QP-Problem besteht darin, den Ausruck $\frac{1}{2}\mathbf{x}'Q\mathbf{x} + \mathbf{c}'\mathbf{x}$ unter der Nebenbedingung $A\mathbf{x} \leq \mathbf{b}$ zu minimieren, wobei $A\mathbf{x} \leq \mathbf{b}$ bedeutet, dass jede Komponente von $A\mathbf{x}$ kleiner oder höchstens gleich der korrespondierenden Komponente von $\mathbf{b}$ sein soll.

Die Typen Polynomial, RBF und Sigmoid sind nichtlinear. Welche dieser Kernfunktion für einen gegebenen Datensatz die richtige ist, muß man durch Probieren herausfinden.

Die Ergebnisse einer SVM-Analyse lassen sich nicht so leicht visuell repräsentieren wie die einer Fisherschen LDA. Bei letzerer wird ja eine Reduktion auf wenige latente Variablen vorgenommen, die dann als Koordinaten für die Fälle verwendet werden. Bei der SVM-Analyse wird dagegen oft eine Transformation in einen höherdimensionalen Raum vorgenommen, sogar in einen unendlichdimensionalen Raum. Die Güte eines SVM-Modells läßt sich zunächst an der Konfusionsmatrix ablesen: je weniger Konfusionen bei der Klassifikation vorkommen, desto besser ist ein Modell. Eine zweite Größe ist die Anzahl der benötigten Support Vektoren. Je höher die Anzahl der SVen, desto eher muß man annehmen, dass die Daten viel Rauschen enthalten. Die Anzahl der benötigten SVen wird vom Programm, etwa svm im Paket e1071 (R) für jede der Kategorien ausgegeben, dazu die Koeffizienten α_i. Weiter wird angegeben, welche der Fälle als SV bestimmt wurden. Hat man K Kategorien, so gibt es insgesamt $K(K-1)/2$ mögliche Trennungen. Eine Trennebene wird aber durch zwei dazu parallele Margins charakterisiert, so dass es zwei mögliche SVs geben kann, einen für den "linken" und eine für den "rechten" Margin. Die Tabelle 17 zeigt die möglichen Kombinationen. Für die Kombinationen (AT, AT), (TT, TT), (PT, PT) und (LN, LN) wird hier einfach X angegeben, da hier keine Trennung existiert. Die Tabelle kann um eine Zeile verkleinert werden, wenn man die X entfernt. Dazu werden die Elemente (TT, AT), (PT, AT) und (LN, AT) der ersten Spalte um eine Zeile nach oben verschoben, in der zweiten Spalte werden die Elemente (PT, TT) und (LN, TT) um eine Zeile nach oben verschoben, und in der dritten Spalte wird das Element (LN, PT) um eine Zeile nach oben verschoben. "Stürzt" man dann die Tabelle wie eine Matrix, so erhält man die Tabelle 18. Offenbar entsteht auf diese Weise stets eine Tabelle mit $K-1$ Spalten, und die Anzahl der Zeilen wird durch die Anzahl der benötigten SVen bestimmt, – hier wird nur der jeweilige Klassenvergleich angegeben. Die Spalten sind mit AT, PT und TT bezeichnet worden, weil diese Kategorien in der zugehörigen Spalte in jeder Kombination vorkommen; eine analoge Aússage gilt für die Zeilen. Die Tabelle enthält nur die Bezeichung der Paare, nicht die einzelnen SVen. Abbildung 23 zeigt die Koeffizienten. Die Legende bezieht sich auf die Spalten der Tabelle 18. Die ersten 17 SVen, kommen

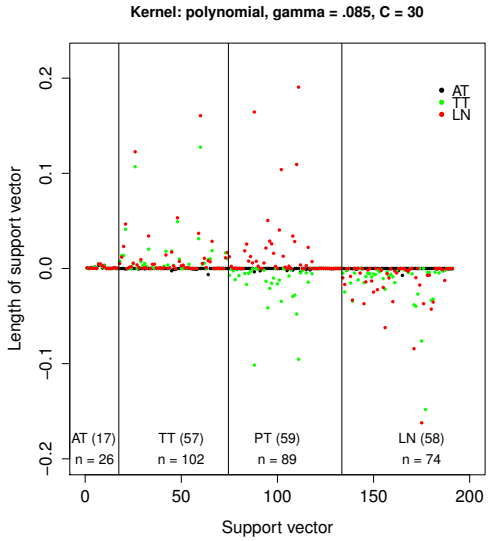
Tabelle 17: Mögliche Trennungen bei den Schilddrüsendaten

i/j	AT	TT	PT	LN
AT	X	AT, TT	AT, PT	AT, LN
TT	TT, AT	X	TT, PT	TT, LN
PT	PT, AT	PT, TT	X	PT, LN
LN	LN, AT	LN, TT	LN, PT	X

Tabelle 18: Ausgabetabelle für die Koeffizienten

i/j	AT	TT	LN
AT	TT, AT	AT, TT	LN, AT
TT	AT, TT	PT, TT	LN, TT
PT	AT, PT	TT, PT	LN, PT
LN	AT, LN	TT, LN	PT, LN

Abbildung 23: Koeffizienten der Schildrüsendaten. AT(17) etc.: number of support vectors in group, $n = 26$ etc.: number of cases in group

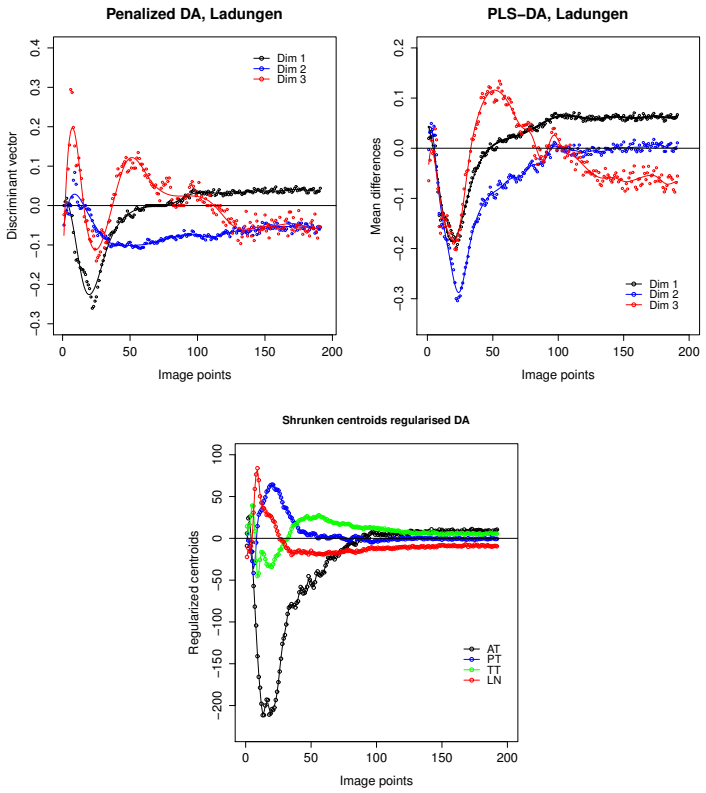


aus der AT-Gruppe, die zweiten 57 SVen gehören zur TT-Gruppe, die dritten 59 SVen gehören zur PT-Gruppe, und die letzten 58 gehören zur LN-Gruppe. Welche Fälle einer Gruppe einen SV bilden kann der Graphik nicht entnommen werden. Festzuhalten ist zunächst dass sehr viele SVen benötigt werden: insgesamt 191 von 296 Fällen insgesamt. Dies scheint die Minimalzahl der SVen zu sein, die für die Daten notwendig sind: sowohl für den linearen Kern wie auch für den radialsymmetrischen Kern erhält man deutlich größere Anzahlen für die SVen. Die Parameterwerde werden in der Graphik angegeben werden. Dabei kann statt $C = 30$ auch $C = 300$ oder $C = 3000$ etc gewählt werden: die Wahl von C bestimmt anscheinend nur den Bereich der Werte der α_i, nicht aber deren Größe relativ zueinander. Die Konfusionstabelle enthält keinerlei Fehlklassifikationen.

4.6 Zusammenfassung

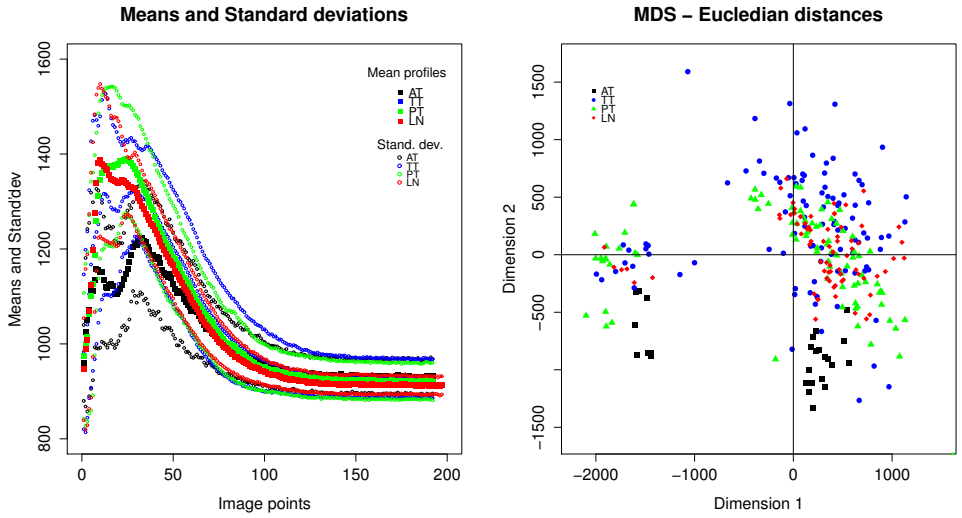
Die Fishersche LDA sowie die Flexible LDA liefern Gewichte der Pixel, deren Verlauf nahelegt, dass die gute Diskriminanzleistung auf einem overfitting beruht. Die Penalized DA, die Shrunken centroid sowie die PLS-Analyse versprechen hier Abhilfe, und die Abschätzungen der Gewichte sollen deshalb in Abbildung 24 noch einmal zusammengefasst werden. Die Abbildungen für die Penalized DA

Abbildung 24: Ladungen bzw. Gewichte der Prädiktoren



einerseits und die PLS-DA andererseits haben miteinander gemein, dass sie sich auf latente Dimensionen beziehen. Es gibt maximal drei latente Dimensionen, und jedes Pixel hat einen "Koeffizienten" oder eine "Ladung" auf ihnen. Die Verläufe für die Dimenensionen 1 und 3 sind einander ziemlich ähnlich, nur die Verläufe der Dimension unterscheiden sich. Die Frage scheint zu sein, welche Darstellung denn nun die "richtige" ist, d.h. welche die Gewichtung angibt, die eine optimale Klassifikation erlaubt. Offenbar ist diese Frage so nicht zu beantworten, – m,an muß die Frage nach der Gewichtung bzw. der Bedeutung der Pixel in Abhängigkeit von der verwendetetn Methode stellen.

Abbildung 25: Schilddrüsendaten: Mittlere Profile ± Standardabweichungen, MDS



In diesem Sinne sind auch die Shrunken Centroids zu interpretieren. Die individuellen Profile werden ja nach Maßgabe ihrer Ähnlichkeit zu den shrunken centroids klassifiziert (Diskriminanzfunktionen wie bei der üblichen LDA gibt es nicht!). Es wird allerdings klar, dass AT uhc hier eine gewisse Sonderrolle zu spielen scheint: das shrunken centroid für AT weicht deutlich von den anderen ab, so dass AT als deutlich von den anderen Gewebearten getrennt erscheint. Allen Methoden gemein, dass es die ersten 50 bis 100 Pixel sind, die über die Zugehörigkeit zu einer der Klassen entscheiden.

Um die schlechte Klassifikation der Schilddrüsendaten zu erläutern, werden in Abbildung 25 noch einmal die Mittleren Profile (große Symbole) plus/minus der jeweiligen Standardabweichungen vorgestellt. Es zeigt sich, dass sich die $(\mu \pm \sigma)$-Kurven für TT, PT und LN stark überlappen, d.h. dass individuelle Profile sich schwer einer Klasse zuordnen lassen. Zusätzlich werden noch einmal die Ergebnisse der MDS mit euklidischen Distanzen präsentiert. Offenbhar gibt es zwei Cluster, – aber sie sind nicht kategorienspezifisch. Dass sowohl Fishers LDA sowie die SVM-Methode zu guten Klassifikationen kommen, scheint daran zu liegen, dass diese Methoden zufällige Variationen für die Klassifikation ausnützen, di8e SVM-Methode durch Transformation in höherdimensionale Räume, in denen dann eine lineare Klassifikation gefunden werden kann.

5 Anhang

5.1 Der Satz von Courant-Fischer

Es gilt der

Satz 5.1 *(Satz von Courant-Fischer) Es sei A eine symmetrische Matric mit Eigenwerten $\lambda_1 \geq \cdots \geq \lambda_n$ und Eigenvektoren $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$. Der Rayleigh-Quotient λ nimmt den maximalen Wert*

$$\max_{\boldsymbol{x}\neq 0} \frac{\boldsymbol{x}'A\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \lambda_1 = \frac{\boldsymbol{p}_1'A\boldsymbol{p}_1}{\boldsymbol{p}_1'\boldsymbol{p}_1} \tag{5.1}$$

und den minimalen Wert

$$\min_{\boldsymbol{x}\neq 0} \frac{\boldsymbol{x}'A\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \lambda_n = \frac{\boldsymbol{p}_n'A\boldsymbol{p}_n}{\boldsymbol{p}_n'\boldsymbol{p}_n} \tag{5.2}$$

an.

Beweis: Da A symmetrisch existiert die Darstellung $A = P\Lambda P'$, P die orthonormalen Eigenvektoren $\mathbf{p}_1, \ldots, \mathbf{p}_n$ und Λ die Diagonalmatrix der Eigenwerte $\lambda_1, \ldots, \lambda_n$ von A. Dann gilt

$$\lambda = \frac{\mathbf{x}'P\Lambda P\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \frac{\mathbf{x}'P\Lambda P\mathbf{x}}{\mathbf{x}'PP'\mathbf{x}},$$

denn $PP' = I$ die Einheitsmatrix. $P'\mathbf{x}$ ist ein Vektor, und es werde $\mathbf{v} = P'\mathbf{x}$ gesetzt, so dass

$$\lambda = \frac{\mathbf{v}'\Lambda\mathbf{v}}{\mathbf{v}'\mathbf{v}} = \frac{\sum_j \lambda_j \mathbf{v}_j^2}{\sum_j v_j^2}. \tag{5.3}$$

Hierin kann man die λ_j durch $\lambda_{\mathbf{x}} = \lambda_1$ ersetzen; man hat dann

$$\lambda = \frac{\sum_j \lambda_j v_j^2}{\sum_j v_j^2} \leq \frac{\sum_j \lambda_1 v_j^2}{\sum_j v_j^2} = \frac{\lambda_1 \sum_j v_j^2}{\sum_j v_j^2} = \lambda_1.$$

Der Wert von λ ist also höchstens gleich dem größten Eigenwert λ_1 von A, – gleich, welchen Vektor $\mathbf{x}$ man wählt. Inbesondere kann man also den zu λ_1 korrespondierenden Eigenvektor $\mathbf{p}_1$ wählen. Man erhält dann

$$\lambda = \frac{\mathbf{p}_1'P\Lambda P'\mathbf{p}_1}{\mathbf{p}_1'\mathbf{p}_1} = \lambda_1,$$

denn die Eigenvektoren in P sind orthonormal, so dass $\mathbf{p}_1'P = (1, 0, \ldots, 0)'$, denn $\mathbf{p}_1'\mathbf{p}_j = 0$ für $j \neq 1$ und $\mathbf{p}_1'\mathbf{p}_j = 1$ für $j = 1$. λ nimmt also den maximal möglichen Wert λ_1 an genau dann, wenn $\mathbf{x} = \mathbf{p}_1$.

Ersetzt man in (5.3) die λ_j durch den kleinsten Eigenwert λ_n, so wird man auf analoge Weise auf (5.2) geführt. □

5.2 Die Wurzel einer Matrix

Ist $0 \leq a \in \mathbb{R}$, so ist die Wurzel $b = a^{1/2} = \sqrt{a}$ diejenige Zahl, für die $b^2 = a$ ist. In analoger Weise kann 245man die Wurzel $\mathbf{A}^{1/2}$ einer positiv definiten, quadratischen Matrix $\mathbf{A}$ erklären: $\mathbf{A}^{1/2}$ ist diejeinige Matrix, für die $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$ gilt.

Da $\mathbf{A}$ als positiv definit vorausgesetzt wird, sind alle Eigenwerte von $\mathbf{A}$ positiv. Ist λ_k ein Eigenwert von $\mathbf{A}$ und $\mathbf{p}_k$ der zu λ_k korrespondierende normierte Eigenvektor, so gilt $\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}'$, wobei $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_n]$ die Matrix der Eigenvektoren von $\mathbf{A}$ ist; es gilt $\mathbf{P}'\mathbf{P} = \mathbf{I}$ die Einheitsmatrix. Λ ist die Diagonalmatrix der igenwerte. Dann gilt

$$\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}' = \sum_{k=1}^{n} \lambda_k \mathbf{p}_k \mathbf{p}_k'. \tag{5.4}$$

Es ist

$$\mathbf{A}^{-1} = (\mathbf{P}\Lambda\mathbf{P}')^{-1} = (\mathbf{P}')^{-1}\Lambda^{-1}\mathbf{P}^{-1}, \tag{5.5}$$

und da $\mathbf{P}^{-1} = \mathbf{P}'$ und $(\mathbf{P}')^{-1} = \mathbf{P}$, hat man

$$\mathbf{A}^{-1} = \mathbf{P}\Lambda^{-1}\mathbf{P}'. \tag{5.6}$$

Es sei $\Lambda^{-1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. Man definiert nun

$$\mathbf{A}^{1/2} = \sum_{k=1}^{n} \sqrt{\lambda_k}\mathbf{p}_k\mathbf{p}' = \mathbf{P}\Lambda^{-1/2}\mathbf{P}'. \tag{5.7}$$

Die Matrix $A^{1/2}$ hat die zu $\sqrt{a}$, $a \in \mathbb{R}$, analogen Eigenschaften:

$$(\mathbf{A}^{1/2})' = \mathbf{A}^{1/2}, \quad (\mathbf{A}^{1/2} \text{ ist symmetrisch}) \tag{5.8}$$

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A} \tag{5.9}$$

$$(\mathbf{A}^{1/2})^{-1} = \sum_{k=1}^{n} \frac{1}{\sqrt{\lambda_k}}\mathbf{p}_k\mathbf{p}_k' = \mathbf{P}\Lambda^{-1/2}\mathbf{P}' \tag{5.10}$$

$$\mathbf{A}^{1/2}\mathbf{A}^{-1/2} = \mathbf{I}, \quad \mathbf{A}^{-1/2}\mathbf{A}^{-1/2} = \mathbf{A}^{-1}. \tag{5.11}$$

5.3 Cauchy-Schwarzsche Ungleichung

Es seien $\mathbf{a} = (a_1, a_2, \ldots, a_n)'$ und $\mathbf{b} = (b_1, b_2, \ldots, b_n)'$ irgend zwei n-dimensionale Vektoren. Dann gilt

$$(\mathbf{a}'\mathbf{b}) \leq (\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b}), \tag{5.12}$$

und der Spezialfall $\mathbf{a}'\mathbf{b} = (\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b})$ gilt genau dann, wenn $\mathbf{b} = c\mathbf{a}$ mit $c \in \mathbb{R}$.

Beweis: Für $\mathbf{a} = 0$ oder $\mathbf{b} = 0$ ist die Aussage trivial. Es seien also $\mathbf{a} \neq 0$ und $\mathbf{b} \neq 0$. Es sei $x \in \mathbb{R}$ ein beliebiger Skalar (d.h. reelle Zahl), und $\mathbf{y}$ sei der Vektor

$\mathbf{y} = \mathbf{a} - x\mathbf{b} \neq 0$, so dass $\mathbf{y}$ eine von Null verschiedene Länge hat. Also gilt

$$\begin{aligned} 0 < |\mathbf{a} - x\mathbf{b}|^2 = (\mathbf{a} - x\mathbf{b})'(\mathbf{a} - x\mathbf{b}) &= \mathbf{a}'\mathbf{a} - x\mathbf{b}'\mathbf{a} + \mathbf{a}'x\mathbf{b} + x\mathbf{b}'x\mathbf{b} \\ &= \mathbf{a}'\mathbf{a} - 2x(x\mathbf{b}'\mathbf{a}) + x^2(\mathbf{b}'x\mathbf{b}). \end{aligned} \quad (5.13)$$

$|\mathbf{a} - x\mathbf{b}|^2$ ist offenbar eine quadratische Funktion von x. Addiert subtrahiert man nun $(\mathbf{a}'\mathbf{b})^2/\mathbf{b}'\mathbf{b}$, so erhält man

$$\begin{aligned} 0 &< \mathbf{a}'\mathbf{a} - 2x(x\mathbf{b}'\mathbf{a}) + x^2(\mathbf{b}'x\mathbf{b}) + (\mathbf{a}'\mathbf{b})^2/\mathbf{b}'\mathbf{b} - (\mathbf{a}'\mathbf{b})^2/\mathbf{b}'\mathbf{b} \\ &= \mathbf{a}'\mathbf{a} - \frac{\mathbf{a}'\mathbf{b}}{\mathbf{b}'\mathbf{b}} + \frac{(\mathbf{a}'\mathbf{b})^2}{\mathbf{b}'\mathbf{b}} - 2x(\mathbf{a}'\mathbf{b}) + x^2(\mathbf{b}'\mathbf{b}) \\ &= \mathbf{a}'\mathbf{a} - \frac{(\mathbf{a}'\mathbf{b})^2}{\mathbf{b}'\mathbf{b}} + (\mathbf{b}'\mathbf{b})\left(x - \frac{\mathbf{a}'\mathbf{b}}{\mathbf{b}'\mathbf{b}}\right)^2. \end{aligned}$$

Der rechte Term verschwindet, wenn $x = \mathbf{a}'\mathbf{a}/\mathbf{b}'\mathbf{b}$, mithin folgt

$$0 < \mathbf{a}'\mathbf{a} - (\mathbf{a}'\mathbf{b})^2/\mathbf{b}'\mathbf{b},$$

so dass

$$(\mathbf{a}'\mathbf{b})^2 < (\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b}),$$

wenn $\mathbf{b} \neq x\mathbf{a}$. □

5.4 Verallgemeinerte Cauchy-Schwarzsche Ungleichung

Es seien $\mathbf{a}$ und $\mathbf{b}$ irgend zwei n-dimensionale Vektoren, und $\mathbf{A}$ sei eine positivdefinite $(n \times n)$-Matrix, d.h. alle Eigenwerte λ_k, $1 \leq k \leq n$ sind größer als Null. Dann gilt

$$(\mathbf{a}'\mathbf{x})' \leq (\mathbf{a}'\mathbf{B}\mathbf{a})(\mathbf{b}'\mathbf{B}^{-1}\mathbf{b}), \quad (5.14)$$

und $(\mathbf{a}'\mathbf{b})' = (\mathbf{a}'\mathbf{A}\mathbf{a})(\mathbf{b}'\mathbf{A}^{-1}\mathbf{b})$ gilt genau dann, wenn $\mathbf{a} = c\mathbf{A}^{-1}\mathbf{b}$ oder $\mathbf{b} = c\mathbf{A}\mathbf{a}$, für ein $c \in \mathbb{R}$.

Beweis: Es sei $\mathbf{B}^{1/2} = \sum_{k=1}^{n} \sqrt{\lambda_k}\mathbf{p}_k\mathbf{p}_k'$, λ_k und $\mathbf{p}_k$ die Eigenwerte und Eigenvektoren von B. Dann ist

$$\mathbf{B}^{-1/2} = \sum_{k=1}^{n} \frac{1}{\sqrt{\lambda_k}}\mathbf{p}_k\mathbf{p}_k'.$$

Dann ist

$$\mathbf{a}'\mathbf{b} = \mathbf{a}'\mathbf{I}\mathbf{b} = \mathbf{a}'B^{1/2}B^{-1/2}\mathbf{b} = (\mathbf{B}^{1/2}\mathbf{a})'(\mathbf{B}^{-1/2}\mathbf{b}).$$

Wendet man jetzt die Cauchy-Schwarzsche Ungleichung auf die Vektoren $\mathbf{B}^{1/2}\mathbf{a}$ und $\mathbf{B}^{-1/2}\mathbf{a}$ an, so folgt die Behauptung. □

Es sei nun $\mathbf{B}$ eine positiv definite $(n \times n)$-Matrix und $\mathbf{a}$ sei ein gegebener n-dimensionaler Vektor. $\mathbf{x}$ sei ein beliebiger n-dimensionaler Vektor. Dann gilt

$$\max_{\mathbf{x} \neq 0} \frac{\mathbf{x}'\mathbf{a}}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}, \quad (5.15)$$

und das Maximum wird angneommen für $\mathbf{x} = c\mathbf{B}^{-1}\mathbf{a}$, für beliebige reelle Konstante c.

Beweis: Nach der verallgemeinerten Cauchy-Schwarzschen Ungleichung gilt

$$(\mathbf{xa})^2 \leq (\mathbf{xBx})(\mathbf{a'B}^{-1}\mathbf{a}).$$

Es ist $\mathbf{x'Bx} > 0$, denn $\mathbf{B}$ ist positiv-definit und $\mathbf{x} \neq 0$. 412Nimmt man den Vektor $\mathbf{x}$, für den das Maximum erreicht wird, erhält man die obere Grenze

$$\frac{(\mathbf{x'a})^2}{\mathbf{x'Bx}} \leq \mathbf{a'B}^{-1}\mathbf{a}.$$

Für $\mathbf{x} = c\mathbf{B}^{-1}\mathbf{a}$ folgt (5.15). □

Literatur

[1] Barker, M., Rayens, W. (2003) Partial least squares for discrimination. *Journal of Chemometrics*, 17, 166 – 173

[2] Bartlett, M.S. (1938) Further aspects of the theory of multiple regression. *Proc. Camb. Philos. Soc.*, 34, 33 – 40

[3] Boulesteix, A.L., Strimmer, K. (2006) Partial Least Squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8 (1), 32 – 44

[4] Courant, R., Hilbert, D.: Methoden der mathematischen Physik I. Springer-Verlag, Berlin, Heidelberg, New York 1968

[5] Delac, K., Grgic, M., Grgic, S. (2005) A comparative Study of PCA, ICA and LDA. *International Journal of Imaging Systems and Technology*, 15 (5), 252-260 Key: citeulike:9070164

[6] Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*}, 7, 179 – 188

[7] Frank, I., Friedman, J.A., (1993) A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109 – 148

[8] Friedman, J. H. (1989) Regularized discriminant analysis. *Journal of the American Statistical Association*, 84, N0. 405 (Theory and Methods), 165 – 175

[9] Guo, Y. Hastie, T., Tibshirani, R. (2005) Regularized discriminant analysis and its application to mircroarerays. *Biostatistics*, 1 (1), 1 – 18

[10] Hastie, T., Buja, A., Tibshirani, R. (1995) Penalized Discriminant Analysis. *The Annals of Statistics*, 33 (1), 73 – 102

[11] Hastie, T., Tibshirani, R., Buja, A. (1994) Flexible Discriminant Analysis by optimal scoring. *Journal of the American Statistical Association*, 89, No. 428, 1255 – 1270

[12] Hastie, T., Tibshirani, R. (1996) Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society B* 58 (1), 155 – 176

[13] Höskuldson, A. (1988) PLS Regression Methods. *Journal of Chemometrics*, 2, 211 – 228

[14] Hastie, T., Tibshirani, R., Friedman, J. (2009) The elements of statistical learning – Data mining, Inference, and Prediction. Springer-Verlag 2009

[15] Indahl, U.G., Liland, K.H., Næs, T. (2009) Canonical partial least squares – a unified PLS approach to classification and regression problems. *Journal of Chemometrics*, 23, 495 – 504

[16] Mahalanobis, P.C. (1936) On the generalized distance in statistics. Proc. Nat. Inst. Sci. Calcutta, 12, 49-55

[17] Martinez, A.M., Kak, A.C. (2001) PCA versus LDA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), 228–233

[18] Meschkowski, H.: Hilbertsche Räume mit Kernfunktionen. Springer-Verlag Berlin 1962

[19] Ripley, B.D.: Pattern recognition and neural networks. Cambridge 2009

[20] Rosipal, R., Krämer, N. (2006) Overview and recent advances in Partial LEast Squares. In: Saunnders et al (Eds): *SLFFS, LNCS 3940*, 34 – 51 (Subspace, Latent Structure and Feature Selection (workshop), – http://dblp1.uni-trier.de/db/conf/slsfs/slsfs2005.html)

[21] Sun, L., Ji, S. Yu, S., Ye, J. (2009) On the equivalence betwenn canonical correlation analysis and orthonormalized partial least squares. *Proceedings of the Twenty-First International Joiunt Conference on Artificial Intelligence* (IJCAI - 09) 1230 – 1235

[22] Tibshirani, R., Hastie, T., Narasinham, B., Chu, G. (2003) Class prediction by nearest shrubnken centroids, with application to DNA microarrays. *Statistical Science*, 18 (1), 104–117

[23] Waaijenborg, S., Zwinderman, A.H. (2007) Penalized canonical correlation analysis to quantify the association between gene expression anbd DNA markers. *BMC Proceedings*, I(Suppl I), 122

[24] Witten, D., Tibshirani. R. (2011) Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society B*, 73 (5), 753 – 772

[25] Wold, H. (1966) Nonlinear estimation by interative least squares procedures. In: David. J. (Ed) *Research Papers in Statististics: Festschrift for J. Neyman* London 1966

Index