.

Fishers Diskriminanzanalyse[1]

Skript zur Veranstaltung *Multivariate Verfahren*

U. Mortensen

.

30. 12. 2016

[1] Fisherdiscriminant.tex

Inhaltsverzeichnis

1 Fishers Ansatz

1.1 Diskriminanzfunktionen

Gegeben sei eine Menge von Objekten, Personen etc, und jedes Objekt bzw, jede Person soll einer von K möglichen Klassen oder Kategorien $\mathcal{C}_1, \ldots, \mathcal{C}_K$ zugeordnet werden. Dazu können die Werte von p Prädiktoren (Symptomen) verwendet werden. Das Ziel ist, auf der Basis der Symptome ein möglichst fehlerfreie Zurodnung zu den Klassen vorzunehmen. Für die Symptome liegen Messungen $X_1, \ldots, X_p$ vor. Die Idee ist, eine Skala Y zu spezifizieren, auf der die Fälle (Objekte, Personen, etc) so repräsentiert werden, dass die Y-Werte für eine gegebene Kategorie möglichst eng zusammen liegen, die Y-Werte für Fälle, die zu verschiedenen Kategorien gehören, so wiet es irgend geht separiert werden. Dazu wird eine Gewichtung der Prädiktoren vorgenommen:

$$Y = u_1 X_1 + \cdots + u_p X_p \tag{1.1}$$

Hierin sind sowohl die Y- wie die u_j-Werte ($j = 1, \ldots, u_p$) unbekannt und müssen aus den Daten geschätzt werden.

In einer Stichprobe werden sowohl die X_j-Werte wie jeweiligen Klassenzugehörigkeiten $\mathcal{C}_k$ erhoben. Die Messungen für Y können Nominalniveau haben, etwa $Y = k$, wenn der Fall $\omega \in \mathcal{C}_k$ liegt. Man hat dann n_k Objekte aus der Klasse $\mathcal{C}_k$. Fishers Idee war, für die Y-Werte eine Varianzzerlegung wie in der Varianzanalyse (ANOVA) vorzunehmen. Demnach kann man eine Quadratsumme Q_{ges} "gesamt" berechnen, sowie die Quadratsummen QS_{zw} "zwischen" und QS_{inn} "innerhalb" der Gruppen, die man für die verschienen KJlassen erhoben hat. QS_{zw} repräsentiert dann die Unterschiede zwischen den Gruppen, Klassen oder Kategorien, und QS_{inn} die Variation innerhalb der Gruppen. Die Klassifizierbarkeit der Fälle wird um so besser (fehlerfreier), je größer QS_{zw} im Vergleich zu QS_{inn}

Abbildung 1: Linear trennbare und linear nicht trennbare Konfigruationen

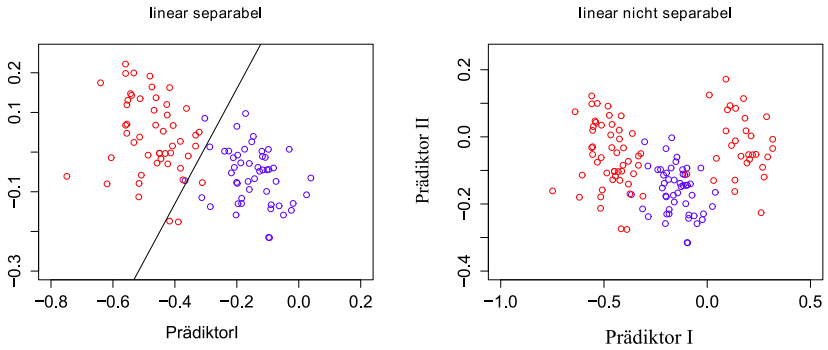
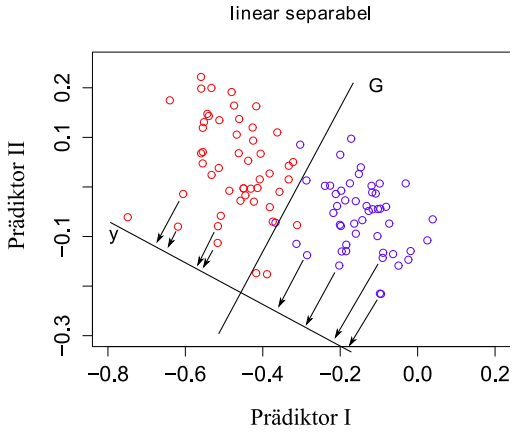


Abbildung 2: Klassifikation nach Fisher (1936) (I): Ω_1 blau, Ω_2 rot, eine mögliche Trennlinie T, eine Projektionsgerade Y



ist, und dieses Verhältnis hängt natürlich von den Gewichten u_j ab. Das Ziel ist also, $\mathbf{u} = (u_1, \ldots, u_p)'$ so zu bestimmen, dass

$$\lambda = \frac{QS_{zw}(\mathbf{u})}{QS_{inn}(\mathbf{u})} \tag{1.2}$$

als Funktion von $\mathbf{u}$ zu maximieren. Für einen neuen, nicht zur Stichprobe gehörenden Fall mit den Messwerten $X_1, \ldots, X_p$ kann dann der Y-Wert anhand der geschätzten u_j-Werte berechnet werden. Der Fall wird dann derjenigen Klasse zugeordnet, zu der der Y-Wert einen minimalen Abstand hat; damit ist gemeint, dass $Y - \bar{y}_k$ die kleinste Differenz sein soll, damit Y der Klasse $\mathcal{C}_k$ zugeordnet wird.

Definition 1.1 *Die in (1.1) definierte Funktion Y heißt* Diskriminanzfunktion *oder* kanonische Variable*, und die in (1.2) definierte Größe λ heißt* Diskriminanzkriterium*.*

1.2 Die Varianzzerlegung

Für die k-te Gruppe, $k = 1, \cdots, K$, gebe es n_k Messungen der Variablen $X_1, \cdots, X_p$, d.h. es gebe n_k Personen oder Objekte Ω_{ik} in der k-ten Gruppe. Es sei X_{ikj} die Messung der Variablen X_j bei der i-ten Person oder dem i-ten Objekt in der k-ten Gruppe. Die Messungen X_{ikj} können in einer Matrix X zusammengefaßt werden, und die y_{ik}-Werte in einem Vektor Y:

$$Y = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_1 1} \\ \hline y_{12} \\ y_{22} \\ \vdots \\ y_{n_2 2} \\ \hline \vdots \\ \hline y_{1K} \\ y_{2K} \\ \vdots \\ y_{n_K K} \end{pmatrix}, \quad X = \begin{pmatrix} X_{111} & X_{112} & \cdots & X_{11p} \\ X_{211} & X_{212} & \cdots & X_{21p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_1 11} & X_{n_1 12} & \cdots & X_{n_1 1p} \\ \hline X_{121} & X_{122} & \cdots & X_{12p} \\ X_{221} & X_{222} & \cdots & X_{22p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_2 21} & X_{n_2 22} & \cdots & X_{n_2 2p} \\ \hline \vdots & \vdots & \cdots & \vdots \\ \hline X_{1K1} & X_{1K2} & \cdots & X_{1Kp} \\ X_{2K1} & X_{2K2} & \cdots & X_{2Kp} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_K K1} & X_{n_K K2} & \cdots & X_{n_K Kp} \end{pmatrix}, \tag{1.3}$$

Für die Y-Werte gelte

$$\begin{aligned} y_{ik} &= u_1 x_{ik1} + u_2 x_{ik2} + \cdots + u_p x_{ikp}, \ \ i = 1, \ldots, n_k && (1.4) \\ \bar{y}_k &= u_1 \bar{x}_{k1} + u_2 \bar{x}_{k2} + \cdots + u_p \bar{x}_{kp} && (1.5) \\ \bar{y} &= u_1 \bar{x}_1 + u_2 \bar{x}_2 + \cdots + u_p \bar{x}_p && (1.6) \end{aligned}$$

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 berechnet 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*

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

$$QS_{ges} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik}-\bar{y})^2, \quad QS_{inn} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik}-\bar{y}_k)^2, \quad QS_{zw} = \sum_{k=1}^{K} n_k(\bar{y}_k-\bar{y})^2. \tag{1.7}$$

Dann gilt

$$QS_{ges} = QS_{zw} + QS_{inn} \tag{1.8}$$

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 (1.2) umgeschrieben werden. Durch Einsetzen ergibt sich

$$QS_{inn} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(u_1(X_{k1i} - \bar{x}_{k1}) + \cdots + u_p(X_{kpi} - \bar{x}_{kp}))^2 \tag{1.9}$$

$$QS_{zw} = \sum_{k=1}^{K} n_k(u_1(\bar{x}_{k1} - \bar{x}_1) + \cdots + u_p(\bar{x}_{kp} - \bar{x}_p))^2 \tag{1.10}$$

$$QS_{ges} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(u_1(X_{ik1} - \bar{x}) + \cdots + u_p(X_{ikp} - \bar{x}))^2 \tag{1.11}$$

Man kann nun die rechten Seiten von (1.9) und (1.10) in den Ausdruck (1.2) 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 (1.2) und damit die Ausdrücke für QS_{zw} und QS_{inn} in Matrixform angeschrieben werden.

1.3 Die Matrixschreibweise für Quadrate von Summen

Ausdrücke wie der in (1.9) vorkommende Summand

$$(u_1(X_{k1i} - \bar{x}_{k1}) + \cdots + u_p(X_{kpi} - \bar{x}_{kp}))^2$$

können als quadratische Form dargestellt werden:

Beispiel 1.1 Zur Illustration werde

$$(u_1x_1 + u_2x_2)^2 = u_1^2x_1^2 + u_2^2x_2^2 + 2u_1u_2x_1x_2$$

betrachtet. Die Terme $x_1^2 = x_1x_1$, $x_2^2 = x_2x_2$ und x_1x_2 sind Elemente einer als ein dyadisches Produkt erzeugten Matrix, wenn man den Vektor $\mathbf{x} = (x_1, x_2)'$ definiert:

$$\mathbf{x}\mathbf{x}' = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} (x_1, x_2) = \begin{pmatrix} x_1^2 & x_1x_2 \\ x_2x_1 & x_2^2 \end{pmatrix}$$

und man rechnet leicht nach, dass dann mit $\mathbf{u} = (u_1, u_2)'$

$$\begin{aligned} (u_1x_1 + u_2x_2)^2 &= \mathbf{u}'(\mathbf{x}\mathbf{x}')\mathbf{u} \\ &= u_1^2x_1^2 + u_2^2x_2^2 + 2u_1u_2x_1x_2. \qquad \square \end{aligned}$$

Es sei also (vergl. (1.9)) $x_{kji} = X_{kji} - \bar{x}_{kj})$, $j = 1, \ldots, p$, und $\mathbf{x}_{ki} = (x_{k1i}, \ldots, x_{kpi})'$. Entsprechend Beispiel 1.1 hat man dann für QS_{inn}

$$QS_{inn} = \mathbf{u}' \left(\sum_k \sum_i \mathbf{x}_{ki}\mathbf{x}_{ki}' \right) \mathbf{u}. \tag{1.12}$$

Die dyadischen Produkte $\mathbf{x}_{ki}\mathbf{x}_{ki}'$ sind Matrizen, also ist die Summe ebenfalls eine Matrix $W = \sum_k \sum_i \mathbf{x}_{ki}\mathbf{x}_{ki}'$, so dass man

$$QS_{inn} = \mathbf{u}'W\mathbf{u} \tag{1.13}$$

erhält.

Weiter sei $d_{kj} = (\bar{x}_{kj} - \bar{x}_j))'$ und $\bar{\mathbf{x}}_k = (d_{k1}, \ldots, d_{kp})'$. Für QS_{zw} ergibt sich

$$QS_{zw} = \sum_{k=1}^{K} n_k \mathbf{u}'(\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k')\mathbf{u}) = \mathbf{u}'B\mathbf{u}, \tag{1.14}$$

mit $B = \sum_k n_k\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k'$. Damit erhält man für das Diskriminanzkriterium

$$\lambda(\mathbf{u}) = \frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'W\mathbf{u}}. \tag{1.15}$$

Die Aufgabe ist nun, $\mathbf{u}$ so zu bestimmen, dass $\lambda(\mathbf{u})$ ein Maximum wird.

Satz 1.2 *λ ist maximal, wenn*

$$W^{-1}B\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1, \tag{1.16}$$

d.h. wenn $\boldsymbol{u}$ der Eigenvektor von $W^{-1}B$ ist; $\lambda = \lambda_1$ ist der zugehörige Eigenwert.

Beweis: W ist symmetrisch, also existiert die Spektralzerlegung $W = P\Lambda P'$, P die Matrix der Eigenvektoren von W und Λ die Diagonalmatrix der Eigenwerte von W. $\Lambda^{1/2}$ sei die Diagonalmatrix der Wurzeln aus den Eigenwerten; dann ist $W^{1/2} = P\Lambda^{1/2}$ die *Wurzel* aus W, denn $W^{1/2}W^{1/2} = P\Lambda^{1/2}\Lambda^{1/2}P' = P\Lambda P' = W$. $W^{1/2}\mathbf{u} = \mathbf{v}$ ist ein Vektor, und mit $\mathbf{u} = W^{-1/2}\mathbf{v}$ erhält man

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

Dies ist ein Rayleigh-Quotient und nach dem Satz von Courant-Fischer wird er maximal, wenn $\mathbf{v} = \mathbf{v}_1$ der erste Eigenvektor von $W^{-1/2}BW^{-1/2}$ ist, d.h.

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

Multplikation von links mit $W^{-1/2}$ liefert

$$W^{-1}BW^{-1/2}\mathbf{v}_1 = \lambda_1\mathbf{v}_1W^{-1/2},$$

und wegen $\mathbf{v}_1W^{-1/2} = \mathbf{u}_1$ folgt die Behauptung. □

$\mathbf{P}_1$ 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 (1.16) in der Form

$$W^{-1}BU = U\Lambda \tag{1.17}$$

schreiben.

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 Disktriminanzfunktionen $\boldsymbol{y}_j = X\boldsymbol{u}_j$, $j = 1, \ldots, r$ und es gilt*

$$\boldsymbol{y}_j'\boldsymbol{y}_k = 0, \quad j \neq k, \tag{1.18}$$

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

$$\mathbf{y}_j = X\mathbf{u}_j. \tag{1.19}$$

Es werde angenommen, dass X splatenzentriert ist, d.h. die Spalten von X seien durch $\mathbf{x}_j = \mathbf{X}_j - \bar{\mathbf{x}}_j$ definiert, wobei $\bar{\mathbf{x}}_j$ der Mittelwert der j-ten Prädiktorwerte sind. 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 $\Sigma = \frac{1}{m}X'X$ die Matrix der Kovarianzen zwischen den $\mathbf{x}_j$ ist. Nun wird Σ aber durch die Matrix W geschätzt (vergl. Gleichung (1.13), Seite 6), $\mathbf{u}_j$ 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

$$Y = XU, \quad U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1r} \\ u_{21} & u_{22} & \cdots & u_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pr} \end{pmatrix} \tag{1.20}$$

für die orthogonale Matrix Y der kanonischen Variablen.

Es ist wichtig, sich die Bedeutung dieser Gleichung klar zu machen. X ist eine $(n \times p)$-Matrix, U ist eine $(p \times r)$-Matrix, so dass Y eine $(n \times r)$-Matrix sein muß. Die r Spaltenvektoren $\mathbf{y}_1, \ldots, \mathbf{y}_r$ von Y sind Linearkombinationen der Spaltenvektoren $\mathbf{x}_1, \ldots, \mathbf{x}_p$ von X. Da die $\mathbf{y}_k$, $k = 1, \ldots, r$ orthogonal sind, definieren sie einen r-dimensionalen Teilraum des V_n.

Die *Zeilen*vektoren $\tilde{\mathbf{y}}_i$, $i = 1, \ldots, n$, von Y sind wiederum Linearkombinationen der *Zeilen*vektoren von U. Ist $\tilde{\mathbf{x}}_i'$ der i-te Zeilenvektor von X, so gilt

$$\tilde{\mathbf{y}}_i' = \tilde{\mathbf{x}}_i' U, \tag{1.21}$$

wobei $\tilde{\mathbf{y}}_i$ und $\tilde{\mathbf{x}}_i$ als *Spalten*vektoren aufgefasst werden, obwohl sie in den Matrizen Y und X als Zeilen stehen; diese Auffassung begründet die Schreibweise $\tilde{\mathbf{y}}_i'$ und $\tilde{\mathbf{x}}_i'$ in (1.21).

Der i-te Fall wird einerseits als Punkt mit den Koordinaten $x_{i1}, \ldots x_{ip}$ im Raum der Prädiktoren $\mathbf{x}_1, \ldots, \mathbf{x}_p$ repräsentiert und hat im Raum der Diskriminanzfunktionen die Koordinaten $y_{i1}, \ldots, y_{ir}$, – diese sind die Komponenten von $\tilde{\mathbf{y}}_i$. Die Punkte $\tilde{\mathbf{y}}_i$, die zu einer bestimmten Kategorie oder Klasse $\mathcal{C}_k$ gehören, bilden dann im Idealfall eine Punktwolke, in der nur die Fälle liegen, die zu $\mathcal{C}_k$ gehören. Im Normalfall werden sich die Punktwolken, die zu verschiedenen Klassen korrespondieren, überlappen, aber die Überlappung ist so gering wie es nur irgend möglich ist.

Beispiel 1.2 Fishers Irisdaten: Fishers (1936) eigenes Beispiel – die Klassifikation von Pflanzen – gehört zu den Standardbeispielen für die Anwendung der Linearen Diskriminanzanalyse. T 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).

Die Daten wurden mit dem Program lda re-analysiert. 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 (Diskriminanzfunktion) erklärt, die zweite dient mehr der Erhöhung der visuellen Deutlichkeit bei der Präsentation der Ergebnisse. Tabelle

Tabelle 1: Fishers (1936) 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

2 zeigt die Ergebnisse einer Kreuzvalidierung: die Tabelle ist ein Beispiel für eine *Konfsionsmatrix*, in der die tatsächliche Klassenzugehörigkeit mit der von der Diskriminanzanalyse vorhergesagten Klassenzugehörigkeit verglichen wird. Die Zahlen sind die Häufigkeiten, mit denen korrekte Entscheidungen bzw. Verwechslungen vorkommen. Die Anzahl der Verwechslungen ist bei diesenm Daten extrem gering; offenbar gelingt bei diesen Daten eine nahezu perfekte Klassifikation anhand der Prädiktoren. Abbildung 3 links zeigt die Ergebnisse der Diskriminan-

Tabelle 2: Kreuzvalidation: Konfusionen

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

zanalyse. Die erste kanonische Variable bzw. Diskriminanzfunktion trennt klar die verschiedenen Gruppen. Rechts wird das Ergebnis einer PCA gezeigt. Zur Erinnerung: die PCA sucht eine Achse, auf der die Fälle maximal separiert werden. In diesem Fall entspricht sie sehr genau der ersten Diskriminanzfunktion: die Variation der Fälle ist gering innerhalb der Gruppen relativ zu den Unterschieden zwischen den Gruppen. Das muß nicht notwendig der Fall sein; bei anderen Daten können PCA und Diskriminanzanalyse zu verschiedenen Ergebnissen führen. In

Abbildung 4 werden die Eigenwerte bzw. die Varianzanteile gezeigt, die zu den latenten Dimensionen korrespondieren. Man sieht, dass die Daten durch maximal zwei latente Dimensionen beschrieben werden; nimmt man das Kaiser-Kriterium ernst, so muß man sich sogar auf nur eine latente Variable beschränken.

Abbildung 3: Analyse der Fisherschen Irisdaten: LDA und PCA

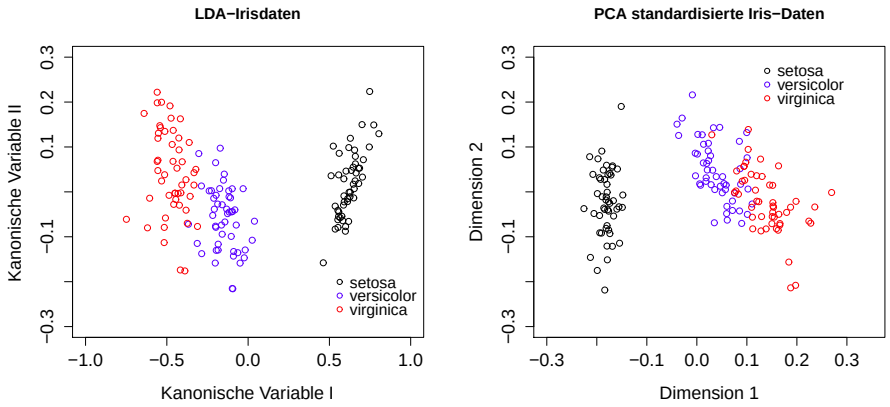
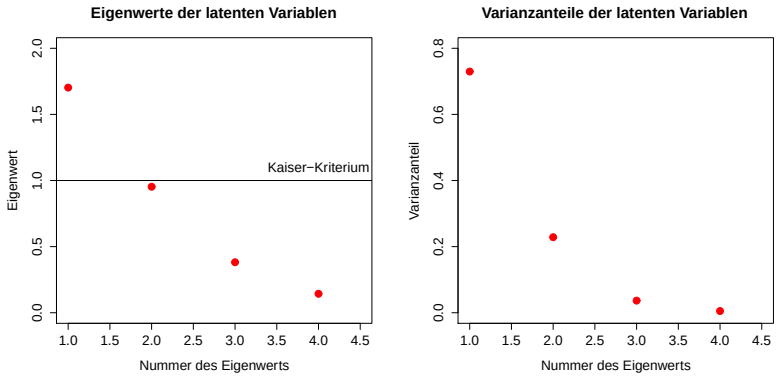
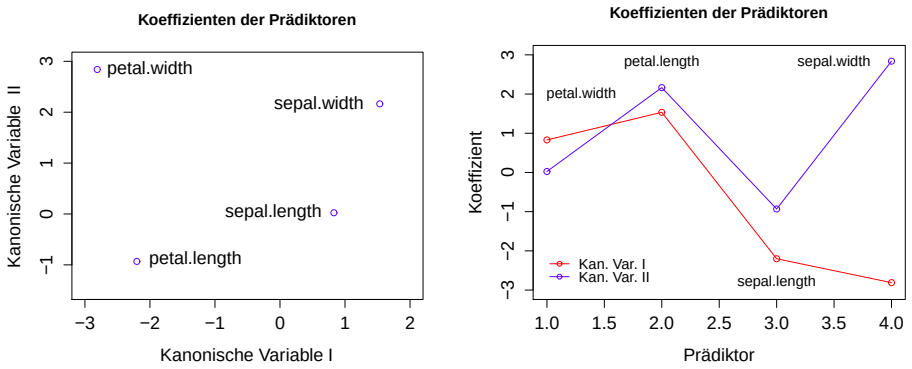


Abbildung 4: Analyse der Fisherschen Irisdaten: Eigenwerte und Varianzanteile der PCA



In der Tabelle 5 werden die Komponenten der Vektoren $\mathbf{u}_k$, $k = 1, 2$ gezeigt. Dem Ansatz (1.1) entsprechend haben die Komponenten u_j von $\mathbf{u}$ die Rolle von Regressionskoeffizienten, die das Gewicht des jeweiligen Prädiktors widerspiegeln. Bei einer solchen Interpretation ist allerdings Vorsicht geboten: wie bei der multiplen Regression gehen mögliche Abhängigkeiten zwischen den Prädiktoren in die Werte der u_j ein. Die Ergebnisse der Diskriminanzanalyse und der PCA legen nahe, dass die vier Prädiktoren durch maximal zwei latente Variablen erklärt werden. Dies steht einer einfachen Interpretation der u_j entgegen. □

Abbildung 5: Koeffizienten der Prädiktoren



1.4 Die Klassifikation von Beobachtungen

Die Gleichung (1.1) kann für einen gegebenen Vektor $\mathbf{u}$ in der Form $\mathbf{y} = X\mathbf{u}$ geschrieben werden. Nun gebe es allgemein $r \leq p$ Vektoren $\mathbf{u}_1, \ldots, \mathbf{u}_r$, die zu einer $(n \times r)$-Matrix U zusammengefasst worden seien. Dann ergibt sich die allgemeine Gleichung

$$Y = XU, \tag{1.22}$$

wobei $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_r]$ eine $(n \times r)$-Matrix ist. In (1.21) wurden bereits die Zeilenvektoren $\tilde{\mathbf{x}}_i$ von X und $\tilde{\mathbf{y}}_i = (y_{i1}, \ldots, y_{ir})'$ von Y zueinander in Beziehung gesetzt:

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

Die Komponenten von $\tilde{\mathbf{y}}_i$ sind die Koordinaten des i-ten Falls im Teilraum der Kanonischen Variablen. Im Allgemeinen sollte ein solcher Punkte innerhalb einer Punktwolke liegen, die zu einer bestimmten Kategorie gehört.

Die Komponentten von $\tilde{\mathbf{y}}_i$ sind die Koordinaten des i-ten Falls im Teilraum der Kanonischen Variablen. Diese Notation führt allerdings bei den folgenden Betrachtungen zu unübersichtlichen Ergebnissen, so dass andere Bezeichnungen eingeführt werden:

$$\mathbf{g}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} = \tilde{\mathbf{x}}_i', \quad \mathbf{h}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{ip} \end{pmatrix} = \tilde{\mathbf{y}}_i. \tag{1.23}$$

Allgemein hat man

$$Y' = U'X'. \tag{1.24}$$

Die Vektoren $\mathbf{h}_i$ sind also die Spaltenvektoren von Y', und die $\mathbf{g}_i$ sind die Spaltenvektoren von X'; für den i-ten Fall gilt demnach

$$\mathbf{h}_i = U'\mathbf{g}_i. \tag{1.25}$$

Nun sei ein neuer Fall mit den Messungen $\tilde{\mathbf{x}} = (x_1, \ldots, x_p)'$ gegeben. Man kann nun $\tilde{\mathbf{y}}' = \mathbf{x}'U$ berechnen und nachsehen, innerhalb welcher kategorienspezifischen Gruppe von Punkten $\mathbf{h}$ liegt; auf diese Weise könnte man eine Klassifikation von $\mathbf{g}$ vornehmen. Es ist aber von Vorteil, die Zuordnungsregel schärfer zu formulieren.

Es sei ω_i der i-te Fall. Um eine Entscheidungsregel zu formulieren werden die Vektormengen

$$\left.\begin{array}{l} G_k(\mathbf{g}) = \{\mathbf{g}_i | \omega_i \in \mathcal{C}_k\} \\ \\ G_k(\mathbf{h}) = \{\mathbf{h}_i | \omega_i \in \mathcal{C}_k\} \end{array}\right. , \quad k = 1, \ldots, K \tag{1.26}$$

Die mittleren Vektoren für $\mathcal{C}_k$ sind dann

$$\bar{\mathbf{g}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbf{g}_i,\ \mathbf{g}_i \in G_k(\mathbf{g}), \quad \overline{\mathbf{h}}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} \mathbf{h}_i,\ \mathbf{h}_i \in G_k(\mathbf{h}) \tag{1.27}$$

Die mittleren Vektoren $\bar{\mathbf{g}}_k$ und $\bar{\mathbf{h}}_k$ heißen auch *Zentroide* der jeweiligen Klasse $\mathcal{C}_k$.

Es sei nun $\mathbf{g} = (x_1, \ldots, x_p)'$ ein "neuer" Vektor, d.h. $\mathbf{g}$ repräsentiere einen neuen Fall, der einer der Kategorien $\mathcal{C}_k$ zugeordnet werden soll. Es sei $\mathbf{h} = U'\mathbf{g}$. Um die Zuordnung vorzunehmen, werden der Differenzenvektoren $\mathbf{h} - \overline{\mathbf{h}}_k$ für $k = 1, \ldots, K$ betrachtet. Man erinnere sich: $\|\mathbf{h} - \overline{\mathbf{h}}_k\|$ ist die Euklidische Distanz zwischen den Endpunkten der Vektoren $\mathbf{h}$ und $\bar{\mathbf{h}}_k$. Man wählt die

Entscheidungsregel:

$$\text{Entscheide für } G_k, \text{ wenn } \min_j \|\mathbf{h} - \overline{\mathbf{h}}_j\| = \|\mathbf{h} - \overline{\mathbf{h}}_k\|. \tag{1.28}$$

Man ordnet ω also derjenigen Klasse zu, zu deren Zentroid der Vektor $\mathbf{g}$ die kleinste Distanz hat.

Es ist

$$\mathbf{h} - \overline{\mathbf{h}}_k = U'\mathbf{g} - U'\bar{\mathbf{g}}_k = U'(\mathbf{g} - \bar{\mathbf{g}}_k),$$

so dass

$$\begin{aligned} \|\mathbf{h} - \overline{\mathbf{h}}_k\|^2 &= (\mathbf{g} - \overline{\mathbf{g}}_k)'UU'(\mathbf{g} - \overline{\mathbf{g}}_k) \\ &= (\mathbf{g} - \overline{\mathbf{g}}_k)'W^{-1}(\mathbf{g} - \overline{\mathbf{g}}_k) \end{aligned} \tag{1.29}$$

d.h. es gilt

$$UU' = W^{-1}. \tag{1.30}$$

Beweis: Nach (1.16) gilt $W^{-1}BU = U\Lambda$. Weiter ist $W^{-1} = W^{-1/2}W^{-1/2}$, so dass

$$W^{-1}B = W^{-1/2}W^{-1/2}BW^{-1/2}W^{1/2}U = U\Lambda.$$

Multiplikation von links mit $W^{1/2}$ liefert

$$W^{-1/2}BW^{-1/2}\underbrace{W^{1/2}U}_{P} = \underbrace{W^{1/2}U}_{P}\Lambda,$$

und P ist offenbar die Matrix der Eigenvektoren der symmetrischen Matrix $W^{-1/2}BW^{-1/2}$, so dass P orthonormal ist, d.h. $P'P = PP' = I$, und es folgt $U = W^{-1/2}P$ und damit $UU' = W^{-1/2}PP'W^{-1/2} = W^{-1}$. □

Fasst man (1.29) und (1.30) zusammen, so erhält man den

Satz 1.4 *Es gilt*

$$\|\boldsymbol{h} - \overline{\boldsymbol{h}}_k\|^2 = (\boldsymbol{h} - \overline{\boldsymbol{h}}_k)'W^{-1}(\boldsymbol{h} - \overline{\boldsymbol{h}}_k), \tag{1.31}$$

.

Kommentar: Das Bemerkenswerte an dieser Aussage ist, dass für eine Klassifikation gar nicht auf die Vektoren $\mathbf{h}$ zurückgegriffen werden muß, sondern anhand der Vektoren $\mathbf{h}$, die ja die Messungen der Prädiktorvariablen repräsentieren, entschieden werden kann. □

$\|\mathbf{h} - \mathbf{h}_k\|$ ist die euklidische Distanz zwischen den Endpunkten von $\mathbf{h}$ und $\overline{\mathbf{h}}_k$. Der Ausdruck rechts ist das Quadrat einer speziellen Distanz:

Definition 1.2 *Die Größe*

$$\delta(\boldsymbol{g}, \bar{\boldsymbol{g}}) = \sqrt{(\boldsymbol{g} - \bar{\boldsymbol{g}}_k)'W^{-1}(\boldsymbol{g} - \bar{\boldsymbol{g}}_k)} \tag{1.32}$$

heißt Mahalanobis-Distanz[2].

Mahalanobis[3] sprach in Bezug auf $\delta(\tilde{\mathbf{x}}, \bar{\tilde{\mathbf{x}}})$ von einer *generalisierten Distanz*. Für $W = I$ die Einheitsmatrix wird die Mahalanobis-Distanz zu einer euklidischen Distanz. Für $W \neq I$ definiert sie eine quadratische Form und für fixe Distanz δ ein Ellipsoid, dessen Hauptachsen durch die Eigenvektkoren von W bzw. W^{-1} bestimmt sind (die Matrizen W^{-1} und W haben dieselben Eigenvektoren, nur die Eigenwerte sind reziprok zueinander). Die Längen der Hauptachsen sind durch die Wurzeln aus den entsprechenden Eigenwerten gegeben. Die multivariate Normalverteilung ist durch eine Mahalanobis-Distanz definiert, worauf später noch eingegangen wird. Die Gültigkeit von (1.31) ist aber nicht an die Annahme einer solchen Verteilung gekoppelt; diese Beziehung ist rein algebraisch.

[2]nach Prasanta Chandra Mahalanobis (1893 - 1972), indischer Physiker und Statistiker.

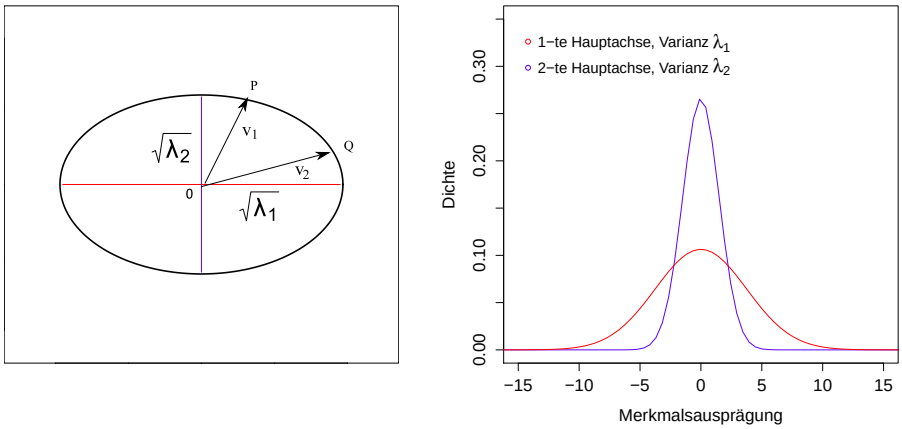
[3]P. C. Mahalanobis: On the generalised distance in statistics. In: Proceedings of the National Institute of Science of India. Vol. 2, Nr. 1, 1936, S. 49–55

Zur Erläuterung der Mahalanobis-Distanz. Gegeben sei eine Ellipse

$$\frac{y_1^2}{a^2} + \frac{y_2^2}{b^2} = 1 \tag{1.33}$$

und es sei $a = \sqrt{\lambda_1}$, $b = \sqrt{\lambda_2}$, wobei λ_1 und λ_2 die Eigenwerte einer symmetrischen Matrix C seien. Ist C keine Diagonalmatrix, so beschreibt (1.33) die zu C korrespondierende achsenparallele Ellipse, die aus der zu C korrespondierenden Ellipse durch Rotation hervorgeht. Die Länge der ersten Hauptachse ist dann gleich $\sqrt{\lambda_1}$, und die der zweiten Hauptachse ist $\sqrt{\lambda_2}$. Gegeben seien zwei Vekto-

Abbildung 6: Zur Mahalanobis-Distanz (zentrierte Daten)



ren $\mathbf{v}_1 = (v_{11}, v_{21})'$ und $\mathbf{v}_2 = (v_{12}.v_{22})'$. Bei $\mathbf{v}_1$ ist die erste Komponente v_{11} klein im Vergleich zu v_{21}, bei $\mathbf{v}_2$ ist es umgekehrt. Für $\lambda_1 > \lambda_2$ ist die erste Hauptachse länger als die zweite, d.h. die Variation der Größen hinsichtlich des Merkmals, das von der ersten Hauptachse repräsentiert wird, ist größer als die Variation des Merkmals, das von der zweiten Hauptachse repräsentiert wird. Je mehr die Länge eines Vektors wie $\mathbf{v}_1$ durch die erste Komponente betimmt wird, desto länger kann der Vektor sein im Vergleich zu einem Vektor, dessen Länge mehr durch die zweite Komponente bestimmt wird, wie etwa $\mathbf{v}_2$. Man kann sagen, dass der Raum nicht mehr, wie bei der euklidischen Distanz, isotrop ist: Bewegungen in Richtung der zweiten Hauptachse sind eingeschränkter als Bewegungen in Richtung der ersten Hauptachse.

Man kann davon ausgehen, dass die Komponenten x_{ij} der Vektoren $\mathbf{v}$ zentrierte Mssungen sind. Komponenten x_{11} oder x_{12} gleich Null bedeuten dann, dass die ursprünglche Messung mit dem jeweiligen Mittelwert der Messungen übereinstimmt. So sei etwa $x_{12} = 0$. Der Vektor $\mathbf{v}_2$ kann dann nur noch in Bezug auf as zweite, von der zweiten Hauütachse repräsntierte Merkmal variiaen, d.h. x_{22}

kann noch variieren. Wenn die Varianz der Messwerte in dieser Richtung klein ist (im Vergleich zur Varianz des ersten Merkmals), so sind eben nur kleinere Vektorlängen in dieser Richtung möglich. Der Mangel an Isotropie reflektiert dann eben nur die unterschiedlichen Varianzen.

Bisher war noch nicht die Rede von Wahrscheinlichkeiten. Es ist möglich, dass die Merkmalsausprägungen hinsichtlich der beiden Merkmale jeweils gleichwahrscheinlich sind, oder irgendeiner anderen Verteilung folgen. Insbesondere können die beiden Verteilungen Gauß-verteilt sein. Die Messungen x_{1j}, x_{2j} für einen Fall ω_j sind dann 2-dimensional Gauß-verteilt, und bei p Messungen sind sie p-dimensional Gauß-verteilt. Die Dichtefunktion ist dann durch die p-dimensionale Gauß-Verteilung

$$f(\mathbf{x}) = A \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{1.34}$$

gegeben. Hier ist Σ die Varianz-Kovarianz-Matrix für die Komponenten der Vektoren $\mathbf{x} = (x_1, \ldots, x_p)'$, $\boldsymbol{\mu}$ ist der Vektor der Mittelwerte der Komponenten, und A ist eine Normierungskonstante. Die multivariate Gauß-Verteilung ist also durch die Mahalanobis-Distanz definiert, und dies bedeutet, dass die Ellipsoide – im 2-dimensionalen Fall die Ellipsen – Orte gleicher Dichte sind.

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 3 angegeben werden. Für die durchschnittliche Varianz-Kovarianz-Matrix S und

Tabelle 3: 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

$$S = \begin{pmatrix} 100 & 30 & 32 \\ 30 & 100 & 44 \\ 32 & 44 & 100 \end{pmatrix}, \quad S^{-1} = \begin{pmatrix} .0115 & -.0023 & -.0027 \\ -.0023 & .0129 & -.0049 \\ -.0027 & -.0049 & .0130 \end{pmatrix} \tag{1.35}$$

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 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

$$d_1 = (108 - 114, 112 - 111, 101 - 110) S^{-1} \begin{pmatrix} 108 \text{ - } 114 \\ 112 \text{ - } 111 \\ 101 \text{ - } 110 \end{pmatrix} \tag{1.36}$$

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. □

1.5 Eigenschaften der Schätzung

Die Klassifikation neuer Fälle kann nach (1.31) 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

$$\|\mathbf{y} - \bar{\mathbf{y}}_j\| = \min_k \|\mathbf{y} - \bar{\mathbf{y}}_k\|. \tag{1.37}$$

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 von W ist und Λ die Diagonalmatrix der Eigenwerte von W. Es ist dann

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

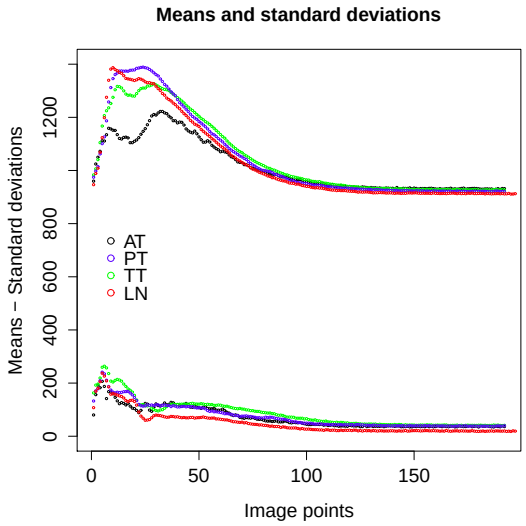
denn P is orthonormal. Ist $\mathbf{p}_k$ der k-te Eigenvektor von W (k-te Spalte von P), so gilt

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

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 bei (i) korrelierenden Prädiktoren und (ii) 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*.

Abbildung 7: Mittlere Helligkeiten (Profile) und Standardabweichungen von OCT-Bildern (Schilddrüsengewebe)

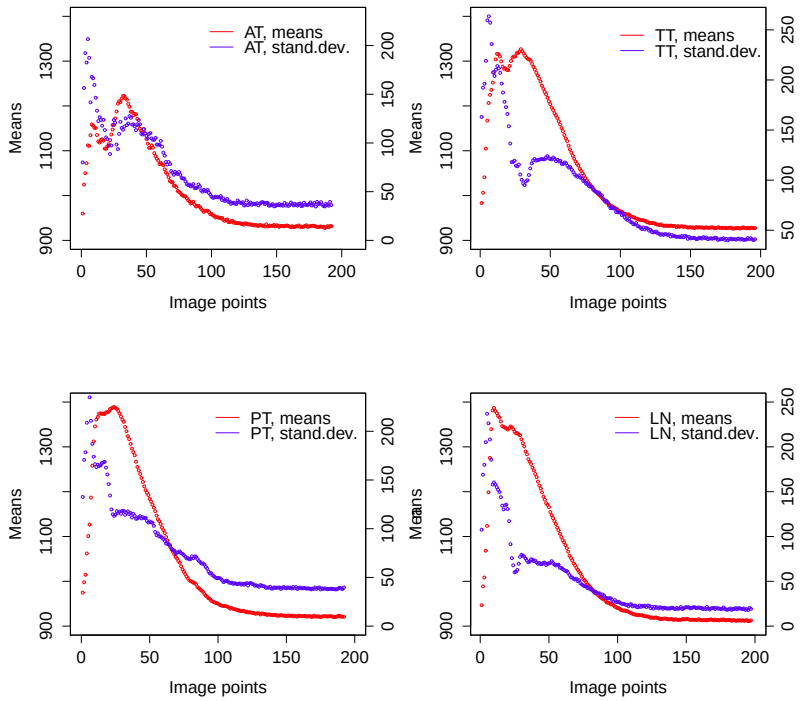


Beispiel 1.4 Klassifikation von Schilddrüsengeweben: In der Medizin müssen vielfach Gewebeproben klassifiziert werden. Für diese Aufgabe können OCT-Bilder (OCT – Optical Coherence Tomography) hilfreich sein. Bei dem hier betrachteten Gewebeproben handelt es sich um Schilddrüsengewebe. Die OCT-Bilder 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 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

Abbildung 8: Mittlere Helligkeiten (Profile) und Standardabweichungen (rechte Skala) von OCT-Bildern (Schilddrüsengewebe)



Annahme der multivariaten Normalverteilung mit homogenen Varianzen bei diesen Daten keinen Sinn macht.

Die Jackknife-Kreuzvalidierung lieferte die folgende Konfusiontabelle: Die Be-

Tabelle 4: 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 %

rechnung 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.

□

Abbildung 9: LDA-Ergebnisse

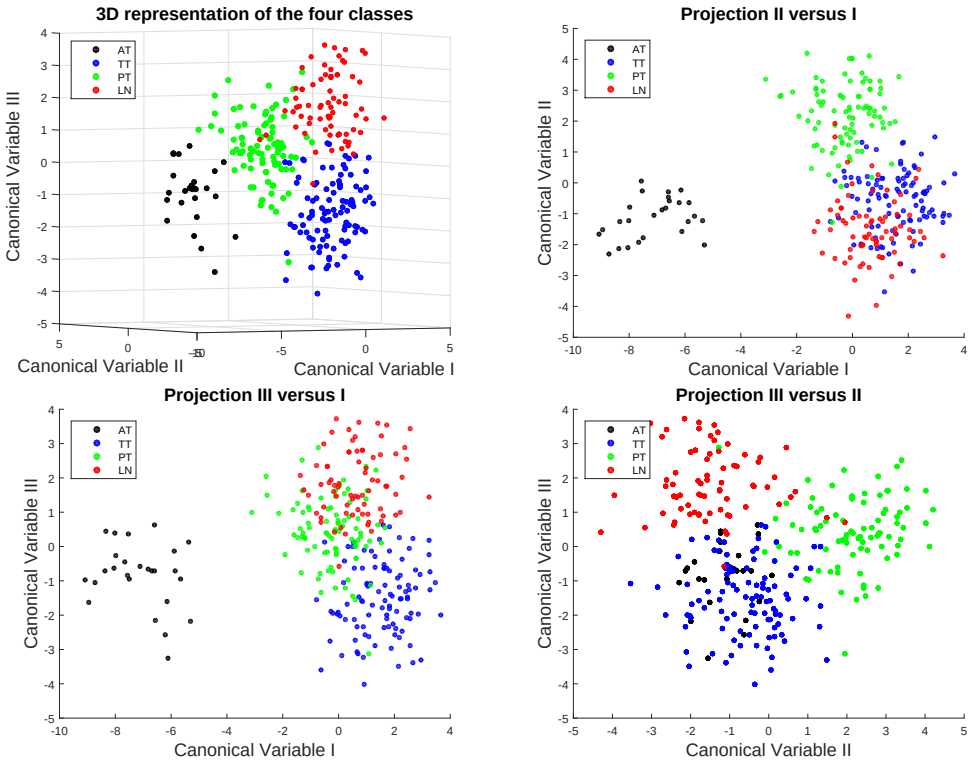


Abbildung 10: LDA – Koeffizienten der Prädiktoren (Pixel)

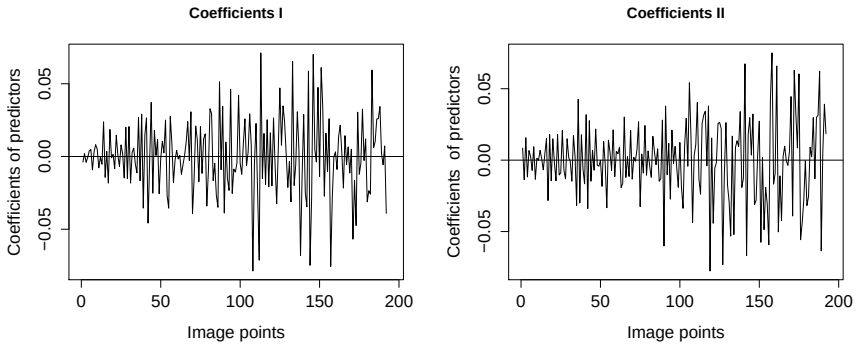
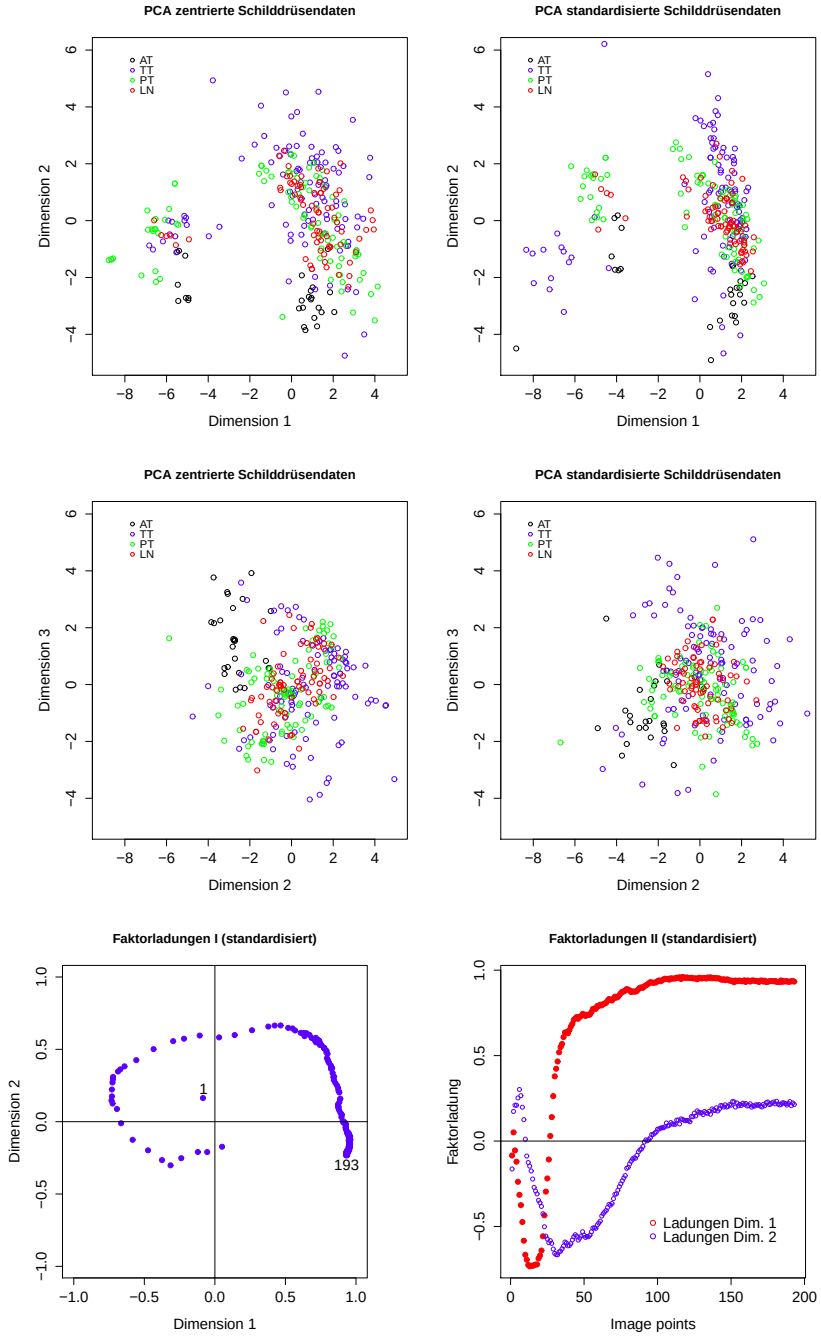


Abbildung 11: Schilddrüsendaten: PCA-Ergebnisse



1.6 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 χ^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

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K, \tag{1.38}$$

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 (1.31) 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

$$H_0: \quad \mu_1 = \mu_2 = \cdots = \mu_K, \tag{1.39}$$

$$H_1: \quad \mu_i \neq \mu_j, \text{ für mindestens ein Paar } (i,j) \text{ mit } i \neq j \tag{1.40}$$

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

$$\Lambda = \frac{|W|}{|B+W|} = |I + W^{-1}B|^{-1}, \tag{1.41}$$

Wilk's Λ; unter H_0 gilt

$$\Lambda \sim \Lambda(q, N-K, K-1) \tag{1.42}$$

(Λ-Verteilung von Wilks).

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

$$\epsilon = p(\Omega_1)\epsilon_{12} + p(\Omega_2)\epsilon_{21} \tag{1.43}$$

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:

$$\epsilon_{12} = \Phi\left(\frac{\log(\pi_1/\pi_2) - \delta^2/2}{\delta}\right) \tag{1.44}$$

$$\epsilon_{21} = \Phi\left(-\frac{\log(\pi_2/\pi_1) + \delta^2/2}{\delta}\right), \tag{1.45}$$

wobei

$$\delta = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \tag{1.46}$$

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

Für die ML-Regel ergeben sich die Fehlerraten gemäß

$$\epsilon_{12} = \epsilon_{21} = \Phi\left(-\frac{\delta}{2}\right). \tag{1.47}$$

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

$$\hat{d}(x) = (x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2))' S^{-1}((\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2) \tag{1.48}$$

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:

$$\begin{aligned} \hat{d}(\bar{x}_1) &= (\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2) && (1.49)\\ \hat{d}(\bar{x}_2) &= -(\bar{x}_1 - \bar{x}_2)' S^{-1}(\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2). && (1.50) \end{aligned}$$

Dann ist für die Bayes-Regel

$$\hat{\epsilon} = \pi_1 \hat{\epsilon}_{12} + \pi_2 \epsilon_{21} \tag{1.51}$$

mit

$$\hat{\epsilon}_{12} = \Phi\left(\frac{\log(\pi_2/\pi_1) - D^2/2}{D}\right), \quad \hat{\epsilon}_{21} = \Phi\left(\frac{-\log(\pi_2/\pi_1) - D^2/2}{D}\right) \tag{1.52}$$

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

$$\hat{\epsilon}_{12} = \hat{\epsilon}_{21} = \Phi(-D/2). \tag{1.53}$$

2 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

$$\begin{aligned} \mathbf{u} &= a_1\mathbf{x}_1 + \cdots + a_p\mathbf{x}_p && (2.1)\\ \mathbf{v} &= b_1\mathbf{y}_1 + \cdots + b_q\mathbf{y}_q, && (2.2) \end{aligned}$$

und $\mathbf{u}'\mathbf{v} = \max$ gilt; bei geeigneter Normnalisierung 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

$$\begin{aligned} \mathbf{u} &= X\mathbf{a} && (2.3) \\ \mathbf{v} &= Y\mathbf{b} && (2.4) \end{aligned}$$

schreiben. Dann ist

$$R_{uv} = \mathbf{u}'\mathbf{v} = \mathbf{a}'X'Y\mathbf{b} = \mathbf{a}'R_{xy}\mathbf{b}. \tag{2.5}$$

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

$$Q(\mathbf{a}, \mathbf{b}) = \mathbf{a}'X'Y\mathbf{b} - \lambda(\mathbf{a}'X'X\mathbf{a} - 1) - \mu(\mathbf{b}'Y'Y\mathbf{b} - 1), \tag{2.6}$$

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

$$\begin{aligned} \frac{\partial Q}{\partial \mathbf{a}} &= R_{xy}\mathbf{b} - \lambda R_{xx}\mathbf{a} = 0 && (2.7) \\ \frac{\partial Q}{\partial \mathbf{b}} &= R_{yx}\mathbf{b} - \mu R_{yy}\mathbf{b} = 0 && (2.8) \end{aligned}$$

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

$$\begin{aligned} \mathbf{a}'R_{xy}\mathbf{b} - \lambda\mathbf{a}'R_{xx}\mathbf{a} &= 0 && (2.9) \\ \mathbf{b}'R_{yx}\mathbf{b} - \mu\mathbf{b}'R_{yy}\mathbf{b} &= 0 && (2.10) \end{aligned}$$

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

$$R_{uv} = \lambda = \mu, \tag{2.11}$$

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

$$\begin{aligned} \lambda^2\mathbf{a} &= R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx}\mathbf{a}. && (2.12) \\ \lambda^2\mathbf{b} &= R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}\mathbf{b}. && (2.13) \end{aligned}$$

$\mathbf{a}$ 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 kann man im Anhang **??** nachlesen. 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.

+

Index