U. Mortensen

Diskriminanzanalyse

Kompaktkurs Multivariate Methoden
Institut für Psychologie der Universität Mainz
WS 2010/2011

Version 18. 01. 2011

Inhaltsverzeichnis

1 Einführung

Das Ziel vieler Messungen (im allgemeinsten Sinn des Wortes: Messungen also auch als Regristrierung von Merkmalen oder Symptomen) ist es, eine Person oder ein Objekt einer bestimmten Klasse zuzuordnen. So soll etwa ein Altphilologe die Frage beantworten, ob ein Text von Platon oder von einem seiner Schüler stammt, ein Kunsthistoriker wird gefragt, ob ein Gemälde von Rembrandt persönlich oder von einem Epigonen gemalt wurde, ein Arzt soll anhand von Symptomen entscheiden, ob ein Patient an einem grippalen Infekt, an einer Hirnhautentzündung oder an sonst einer Erkrankung leidet. Eine Psychologin muß entscheiden, ob ein Patient, der Symptome der Depression zeigt, auf Grund situativer Umstände nur depressiv verstimmt ist oder ob hirnphysiologische Prozesse die Ursache für den Zustand des Patienten sind, oder ob ein Bewerber für eine Managementposition für die zu bewältigenden Aufgaben geeignet ist oder nicht. Die Entscheidungen sind nicht immer fehlerfrei, es kommt darauf an, die "Evidenz" optimal zu nutzen, so dass die Wahrscheinlichkeit eines Fehlers möglichst gering wird.

In all diesen Situationen werden mehrere Merkmale – "features", Symptome – benutzt, um zu einer Klassifikation zu kommen. Experten benutzen dabei eine Gewichtung dieser Merkmale, oft implizit und ohne sich bewußt zu sein, welcher Art diese Gewichtung ist. So kann es vorkommen, dass die Gewichtung suboptimal ist (man denke an Fehlbeurteilungen in Prüfungssituationen oder in Einstellungsgesprächen, wie etwa im Beispiel der von Experten geführten Aufnahmegespräche der Texas Medical School, oder in therapeutischen Situationen, die zu Fehleinschätzungen der psychischen Problematik von Patienten führen können). Die Frage ist dann, wie man zu einer möglichst guten Gewichtung gelangen kann.

Liegen Messungen von Merkmalen in numerischer Form vor, kann man unter Umständen von bestimmten Wahrscheinlichkeitsverteilungen für die Messungen ausgehen, um die Entscheidungsfindung zu optimieren. Allerdings können Annahmen über Wahrscheinlichkeisverteilungen falsch sein, oder die Daten weisen nicht hinreichend eindeutig auf eine bestimmte Wahrscheinlichkeitsverteilung hin. Dann liegt es nahe, zu versuchen, ohne eine solche Annahme auszukommen. Einen solchen Ansatz hat Fisher (1936)[1] vorgelegt, und dieser Ansatz soll hier vorgestellt werden.

Es sei ω eine Person oder ein Objekt, für das die Messungen $x_1, x_2, \ldots, x_p$ von p Merkmalen vorliegen. Die Aufgabe ist nun, anhand dieser Werte ω einer Klasse $\mathcal{K}_k$ zuzuordnen, woebei es K Klassen geben möge, also $k = 1, \ldots, K$. Von einem geometrischen Standpunkt aus gesehen wird ω durch einen Punkt mit den Koordinaten $x_1, x_2, \ldots, x_p$ in einem p-dimensionalen Raum mit den Koordinatenachsen $X_1, X_2, \ldots, X_p$ repräsentiert. Im allgemeinen hat man viele Personen oder Objekte $\omega_1, \omega_2, \ldots, \omega_m$, so dass die Werte der Koordinaten, durch die ω_i durch einen Punkt repräsentiert repräsentiert wird, entsprechend doppelt indiziert werden müssen:

$$\omega_i \rightarrow (x_{i1}, x_{i2}, \ldots, x_{ip}), \quad i = 1, 2, \ldots, m \tag{1}$$

Die ω_i bilden dann "Punktwolken". Sind die zu verschiedenen Klassen korrespondierenden Teilpunktwolken wohlsepariert, so sind fehlerfreie Zuordnungen der ω_i zur jeweiligen Klasse möglich. Man muß dann nur die Teilräume bestimmen, die zu den einzelnen Klassen

[1] Fisher, R.A. (1936) The use of multiple measurements in Taxonomic Problems. Annals of Eugenics, 7, 179-188

korrespondieren. Es zeigt sich, dass in den meisten Fällen eine solche Wohlseparierheit in Bezug auf die Koordinatenachsen $X_1, X_2, \ldots, X_p$ nicht gegeben ist, so dass die Messungen $(x_{i1}, x_{i2}, \ldots, x_{ip})$ für ω_i so, wie sie gegeben sind, nicht immer zu einer korrekten Zuordnung führen. Fisher (1936) hat nun gezeigt, dass bestimmte Transformationen der Messungen zu einer Minimierung der Fehlerwahrscheinlichkeiten führen können. Diese Transformationen werden im folgenden Abschnitt genauer vorgestellt.

2 Darstellung des Verfahrens

2.1 Diskriminanzfunktionen und Diskriminanzkriterium

Fishers Ansatz besteht in der Annahme, dass die Kategorien auf zumindest einer, eventuell neu zu definierender Skala angeordnet werden können, wobei diese Skala so geartet ist, dass sie für gegebene Variablen (Symptome etc) eine optimale, d.h. minimal fehlerhafte Zuordnung von Personen oder Objekten zu den Klassen gestattet.

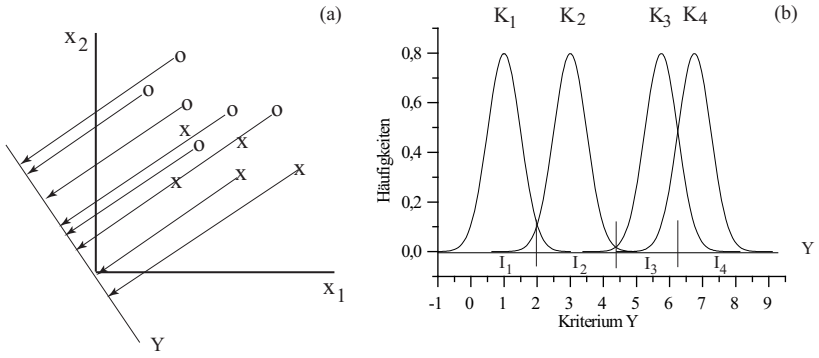
Dazu definiert Fisher die Funktion

$$Y = u_1 X_1 + \cdots + u_p X_p, \quad p \geq 2. \tag{2}$$

Warum die Forderung $p \geq 2$ aufgestellt wird, wird später diskutiert. Y ist als als Linearkombination der X_j, $j = 1, \cdots, p$ definiert. Der Variationsbereich der Y-Werte wird in K Intervalle aufgeteilt, die die einzelnen Kategorien repräsentieren, vergl. Abbildung 1, (b). $\mathcal{K}_1$ bis $\mathcal{K}_4$ sind dort die vorgegebenen Kategorien, denen Intervalle I_1 bis I_4 auf der Y-Skala entsprechen. Die Häufigkeitsverteilungen der Y-Werte sind typischerweise nicht auf die Intervalle I_k beschränkt, es gibt Überlappungen, die zu Fehlurteilen führen. Da die Y-Werte aus den zur Verfügung stehenden X_j-Werten berechnet werden und die $u_1, \ldots, u_p$ in noch zu zeigender Weise optimal bestimmt werden, muß man mit diesen möglichen Fehlern leben. Eine Reduktion der Anzahl der Fehler auf Null wird dann nur möglich sein, wenn man zu anderen Symptomen, d.h. anderen Variablen übergehen kann.

Die Frage ist nun, wie die Gewichte $u_1, \ldots, u_p$ bestimmt werden können. Das Ziel ist sicherlich, Y so zu bestimmen, dass sich die Verteilungen (verg. Abb. 1 (b)) möglichst wenig überschneiden. Dies bedeutet, dass sich die Mittelwerte für die Klassen, im Beispiel $\bar{y}_1, \ldots, \bar{y}_4$ so ausgeprägt wie möglich unterscheiden sollen, – relativ zu den Varianzen der Verteilungen innerhalb der einzelnen Klassen. Nimmt man für den Moment an, dass die Varianzen der Häufigkeitsverteilungen im Prinzip gleich groß sind, sich also nur wegen der Stichprobenfehler unterscheiden, so kann man die Varianz *zwischen* den Kategorien mit der Varianz der Mittelwerte in Verbindung bringen, und für die Varianz *innerhalb* der Kategorien die gemittelte Varianz innerhalb der Kategorien (*pooled variance* wie in der Varianzanalyse). In der Tat liegt hier eine Beziehung zur Varianzanalyse (ANOVA) vor: während bei der ANOVA die gemessenen Mittelwertsunterschiede auf Signifikanz getestet werden, sollen sie hier auf der Basis geeignet gewählter Gewichte $u_1, \ldots, u_p$ so groß wie möglich gemacht werden. Natürlich kann man jetzt eine ANOVA in Bezug auf die $\bar{y}_j$-Werte rechnen, um zu prüfen, ob sich die $\bar{y}_j$-Werte signifikant unterscheiden, – man würde damit prüfen, ob die Messwerte überhaupt Information über die Unterschiedlichkeit der Klassen enthalten.

Abbildung 1: Diskriminanzanalyse: Fishers Ansatz.In Bezug auf die ursprünglichen Koordinaten werden die Objekte o und x kaum getrennt, in Bezug auf die neue Variable Y ist die Trennung deutlich.



Statt die Varianz der Mittelwerte und dann die gemittelte Varianz innerhalb der Gruppen zu berechnen kann man gleich, analog zum Vorgehen bei der ANOVA, die Quadratsumme der Abweichungen in eine Teilsumme QS_{inn} zu Lasten der Variation der Y-Werte innerhalb und in eine Teilsumme QS_{zw} zu Lasten der Variation der Mittelwerte zerlegen; in Abschnitt 2.2 wird diese Zerlegung geliefert. Da die Y-Werte von den mit u_j gewichteten X_j-Werten abhängen, $j = 1, \cdots, p$, werden auch diese Quadratsummen von den u_j abhängen. Diese Gewichte sollen so bestimmt werden, daß QS_{zw} groß im Vergleich zu QS_{inn} ist. Die Aufgabe, die Werte der u_j in dieser Weise zu bestimmen, ist dann gleichbedeutend mit der Aufgabe, den Wert des Quotienten

$$\lambda = \lambda(u_1, \cdots, u_p) = \frac{QS_{zw}(u_1, \cdots, u_p)}{QS_{inn}(u_1, \cdots, u_p)} \tag{3}$$

zu maximieren, denn je größer $QS_{zw}(u_1, \cdots, u_p)$ im Vergleich zu $QS_{inn}(u_1, \cdots, u_p)$, desto größer wird der Wert von λ sein.

Definition 2.1 *Die Funktion (2) heißt* lineare Diskriminanzfunktion *oder* kanonische Variable*. Die in (3) definierte Größe λ heißt* Diskriminanzkriterium.

Die tatsächliche Bestimmung der $u_1, \cdots, u_p$ ist dann die *Diskriminanzanalyse*.

Die Diskriminanzfunktion wurde zuerst von Fisher (1936) eingeführt. Die alternative Bezeichnung *kanonische Variable* ergibt sich aus einem Zusammenhang mit der Kanonischen Korrelation, auf die in einem gesonderten Skript eingegangen wird.

Natürlich kann es sein, daß 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.

Nach (2) wird anhand der Y-Werte entschieden. Die die Personen oder Objekte repräsentierenden Punkte im p-dimensionalen Raum werden auf einen Raum mit deutlich

geringerer Dimension projiziert; hat $W^{-1}B$ nur einen Eigenvektor, so werden sie auf einen 1-dimensionalen Raum, eben die Achse Y projiziert. Es werde nun angenommen, man habe nur einen Prädiktor oder nur ein Symptom, so dass $p = 1$. Dann können Unterschiede zwischen den Objekten nur hinsichtlich dieses einen Merkmals regristriert werden. Es kann dann auch nur ein Gewicht $u_1 = u$ geben, so dass $Y = uX$. Y repräsentiert dann nur eine Streckung oder Stauchung der X-Skala, womit sich zwar die Mittelwerte, aber auch die Varianzen ändern, und in Bezug auf die Optimierung der Kategorisierung ist nichts gewonnen. Deshalb wird $p \geq 2$ gefordert.

Es sei $p = 2$. Man kann nun fragen, wie die Werte von X_1 und X_2 variieren dürfen, ohne dass es zu einer anderen Kategorisierung kommt. Dazu sei $Y = y \in I_k$, also ein fester Wert im Intervall I_k, und es gilt

$$y = u_1 X_1 + u_2 X_2.$$

Daraus folgt

$$X_2 = \frac{y}{u_2} - \frac{u_1}{u_2} X_1.$$

Dies ist die Gleichung einer Geraden: Alle Punkte (X_1, X_2), für die sich der Y-Wert y ergibt, liegen auf dieser Geraden, weil sie dieser Beziehung genügen müssen. Diese Beziehung gilt für jeden Punkt $y \in I_k$; die Geradengleichungen unterscheiden sich nur durch den von y abhängenden Wert der additiven Konstanten. Die Menge der Punkte (X_1, X_2), die auf die Kategorie $\mathcal{K}_k$ führen, liegen also alle in einem Rechteck der Breite I_k, das senkrecht über I_k auf der Y-Achse steht. Für $p = 3$ liegen die Punkte (X_1, X_2, X_3), die zu einem bestimmten y-Wert führen, auf einer Ebene, die Punkte, die zu einem Intervall I_k führen, liegen in einem Kubus. Für $p > 3$ ergeben sich Hyperebenen und Hyperräume.

2.2 Die Varianzzerlegung und die Maximierung des Diskriminanzkriteriums

Es sollen zunächst explizite Ausdrücke für die Quadratsummen QS_{inn} und QS_{zw} hergeleitet werden, um die Maximierung des Kriteriums λ durchführen zu können. Dazu ist es nützlich, die Gleichung (2) etwas ausführlicher anzuschreiben.

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:

$$Y = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_1 1} \\ \hline y_{12} \\ y_{22} \\ \vdots \\ y_{n_2 2} \\ \hline \vdots \\ \hline y_{1K} \\ y_{2K} \\ \vdots \\ y_{n_K K} \end{pmatrix}, \quad X = \left(\begin{array}{cccc} X_{111} & X_{112} & \cdots & X_{11p} \\ X_{211} & X_{212} & \cdots & X_{21p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_1 11} & X_{n_1 12} & \cdots & X_{n_1 1p} \\ \hline X_{121} & X_{122} & \cdots & X_{12p} \\ X_{221} & X_{222} & \cdots & X_{22p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_2 21} & X_{n_2 22} & \cdots & X_{n_2 2p} \\ \hline \vdots & \vdots & \cdots & \vdots \\ \hline X_{1K1} & X_{1K2} & \cdots & X_{1Kp} \\ X_{2K1} & X_{2K2} & \cdots & X_{2Kp} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_K K1} & X_{n_K K2} & \cdots & X_{n_K Kp} \end{array}\right), \tag{4}$$

Für die Y-Werte gelte

$$y_{ik} = u_1 x_{ik1} + u_2 x_{ik2} + \cdots + u_p x_{ikp}, \ \ i = 1, \ldots, n_k \tag{5}$$

$$\bar{y}_k = u_1 \bar{x}_{k1} + u_2 \bar{x}_{k2} + \cdots + u_p \bar{x}_{kp} \tag{6}$$

$$\bar{y} = u_1 \bar{x}_1 + u_2 \bar{x}_2 + \cdots + u_p \bar{x}_p \tag{7}$$

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

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

$$QS_{ges} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ik} - \bar{y})^2, \quad QS_{inn} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (y_{ik} - \bar{y}_k)^2, \quad QS_{zw} = \sum_{k=1}^{K} n_k (\bar{y}_k - \bar{y})^2. \tag{8}$$

Dann gilt

$$QS_{ges} = QS_{zw} + QS_{inn} \tag{9}$$

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

□

Um die Abhängigkeit von den u_j, $1 \le j \le p$ explizit zu machen, muß (3) umgeschrieben werden. Durch Einsetzen ergibt sich

$$QS_{inn} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (u_1 (X_{k1i} - \bar{x}_{k1}) + \cdots + u_p (X_{kpi} - \bar{x}_{kp}))^2 \tag{10}$$

$$QS_{zw} = \sum_{k=1}^{K} n_k(u_1(\bar{x}_{k1} - \bar{x}_1) + \cdots + u_p(\bar{x}_{kp} - \bar{x}_p))^2 \quad (11)$$

Man kann nun die rechten Seiten von (10) und (11) in den Ausdruck (3) 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 (3) und damit die Ausdrücke für QS_{zw} und QS_{inn} in Matrixform angeschrieben werden. Betrachtet man den Ausdruck für QS_{inn} in (10), so sieht man, dass QS_{inn} sich durch Summation der Ausdrücke

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

ergibt. Offenbar gilt

$$=$$

$$\begin{aligned}(u_1(X_{k1i} - \bar{x}_{k1}) + \cdots + u_p(X_{kpi} - \bar{x}_{kp}))^2 &= \sum_{j=1}^{p} u_j^2((X_{kji} - \bar{x}_{kj}))^2 \\ &= +\sum_{j \neq j'} (u_j(X_{kji} - \bar{x}_{kj})(u_{j'}(X_{kj'i} - \bar{x}_{kj'}).\end{aligned}$$

Um hieraus QS_{inn} zu gewinnen, muß noch über die Indices i und k summiert werden. Da es auf die Reihenfolge der Summation nicht ankommt, erhält man

$$QS_{inn} = \sum_{j=1}^{p} u_j^2 \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_{kji} - \bar{x}_{kj}))^2 + \sum_{j \neq j'} u_j u_{j'} \sum_{k=1}^{K} \sum_{i=1}^{n_k} (X_{kji} - \bar{x}_{kj})((X_{kj'i} - \bar{x}_{kj'}).$$

Die erste Summe auf der rechten Seite enthält die Quadratsumme für die zusammengefasste Varianzen für die p Prädiktoren, und die zweite Summe enthält die zusammengefassten Kovarianzen zwischen den p Prädiktorwerten. Es kann nun tatsächlich gezeigt werden, dass QS_{inn} sich als Quadratische Form

$$QS_{inn} = \vec{u}\,' W \vec{u} \quad (12)$$

darstellen lät, wobei W die Matrix der Varianzen (in den Diagonalzellen) und Kovarianzen "innerhalb" ist, und $\vec{u} = (u_1, u_2, \ldots, u_p)'$. Für QS_{zw} ergibt sich der analoge Ausdruck

$$QS_{zw} = \vec{u}\,' B \vec{u}, \quad (13)$$

wobei B die Matrix der Varianzen und Kovarianzen "zwischen" den Kategorien ist. Eine explizite Herleitung der Gleichungen (12) und (13) wird im Anhang geliefert. Für das Diskriminanzkriterium ergibt sich jedenfalls der Ausdruck

$$\lambda = \frac{\vec{u}\,' B \vec{u}}{\vec{u}\,' W \vec{u}}. \quad (14)$$

Man maximiert λ als Funktion der u_j, indem man diese Gleichung nach den u_j ableitet und die Ableitungen gleich Null setzt. Das Ergebnis ist

$$W^{-1} B \vec{u}_{\max} = \lambda_{\max} \vec{u}_{\max}, \quad (15)$$

wobei $\vec{u}_{\max}$ der Vektor mit den gesuchten optimalen Gewichten ist und $\lambda_{\max}$ ist der maximale Wert des Diskriminanzkriteriums. $\vec{u}_{\max}$ ist offenbar ein Eigenvektor von $W^{-1}B$ und $\lambda_{\max}$ ist der zugehörige Eigenwert.

2.3 Die Zuordnung von Personen oder Objekten zu Klassen

Die Messung des Objekts oder der Person ω habe den Vektor $\vec{x}$ ergeben. Dann sind die ω bezüglich der Diskriminanzfunktionen $\vec{Y}_1, \ldots, \vec{Y}_s$ maximal diskriminierbar, wobei $\vec{Y}_r = X\vec{u}_r$, $r = 1, \ldots, s$.

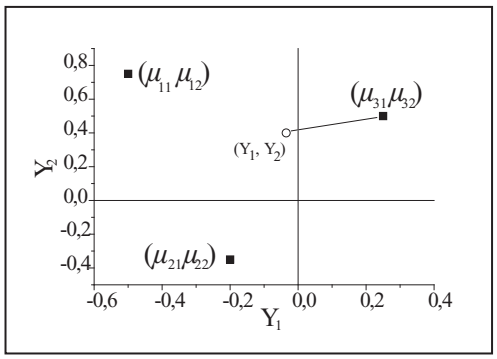
ω soll nun einer Klasse Ω_k zugeordnet werden. Es wird die folgende Entscheidungsregel eingeführt:

> **Regel:** Es sei $\vec{y} = (y_1, \ldots, y_s)'$, und $E(\vec{y}|\Omega_j) = \vec{\mu}_{y_j} = (\mu_{y_{j1}}, \ldots, \mu_{y_{js}})'$ sei der Vektor der Erwartungswerte der y_i, $1 \leq i \leq s$ wenn $\omega \in \Omega_j$, wenn also das zu klassifizierende Objekt oder die Person tatsächlich zur Klasse Ω_j gehört. Weiter sei
>
> $$(\vec{y} - \vec{\mu}_{yj})'(\vec{y} - \vec{\mu}_{yj}) = \sum_{r=1}^{s}(y_r - \mu_{Yjr})^2 = \|\vec{y} - \vec{\mu}_{yj}\|^2; \tag{16}$$
>
> $\mu_{y_{jr}}$ ist die r-te Komponente des Vektors μ_{y_j}; $\|Y - \mu_{Yj}\|^2$ ist das Quadrat der Euklidischen Distanz zwischen den Punkten $\vec{y}$ und μ_{Y_j}. Die Zuordnung von ω zu einer Klasse Ω_k erfolgt gemäß der Bedingung
>
> $$\omega \to \Omega_k \text{ genau dann, wenn } \|\vec{y} - \vec{\mu}_{y_k}\| = \min_j \|\vec{y} - \vec{\mu}_{y_j}\|. \tag{17}$$

Nach Definition des Vektors $\vec{y}$ gelten für die Komponenten y_r die Beziehungen $y_r = \vec{u}_r'\vec{x}$, $\vec{x}$ der Vektor der Beobachtungen, d.h. der p Messungen, aufgrund derer ω zu klassifizieren ist. Es sei $\vec{\mu}_j = \vec{E}(X_j)$, $1 \leq j \leq p$, der Vektor der Erwartungswerte der Variablen X_j. Sicher gilt dann für die r-te Komponente μ_{rj} die Beziehung $\mu_{rj} = \vec{u}_r'\vec{\mu}_j$, so daß

$$\sum_{r=1}^{s}(Y_r - \mu_{yjr})^2 = \sum_{r=1}^{s}(\vec{u}_r'(\vec{x} - \vec{\mu}_j))^2. \tag{18}$$

Abbildung 2: Klassifikation nach dem Fisher-Ansatz; der Punkt (y_1, y_2) wird der Klasse Ω_3 zugeordnet, da die (euklidische) Distanz von (y_1, y_2) zu (μ_{31}, μ_{32}) die kleinste ist.



Das Entscheidungskriterium ist in Bezug auf die Vektoren *vecy* definiert worden. Man kann nun fragen, in welcher Weise die unmittelbar beobachteten x-Werte in die Entscheidung eingehen. Dazu wird der folgende Satz bewiesen:

Satz 2.2 *Es sei* $y_j = u_j'\vec{x}$, $\vec{y} = (y_1, \ldots, y_s)'$, *wobei* $\vec{x} = (x_1, x_2, \ldots, x_p)'$ *der Vektor der Messwerte ist und* $u_j = (u_{j1}, \ldots, u_{jp})'$ *der* j*-te Eigenvektor von* $W^{-1}B$, $j = 1, \ldots, s$, s *die Anzahl der Eigenwerte von* $W^{-1}B$ *ungleich Null, und es sei. Weiter sei* $\mu_{kj}(y) = E(y|\Omega_k)$ *der Erwartungswert von* y_j, *wenn das zu klassifizierende Objekt zur Klasse* Ω_k *gehört;* $\vec{\mu}_k(\vec{y}) = (\mu_{k1}(y), \ldots, \mu_{ks}(Y))'$. *Dann gilt*

$$\|\vec{y} - \vec{\mu}_k(y)\|^2 = (\vec{x} - \vec{\mu}_k)'\Sigma^{-1}(\vec{x} - \vec{\mu}_k), \tag{19}$$

Beweis: Siehe Anhang, Abschnitt 2.2. □

Der Ausdruck auf der rechten Seite von (19) definiert eine spezielle Art von Distanz:

Definition 2.2 *Die Größe*

$$\delta = \sqrt{(\vec{x} - \vec{\mu}_k)'\Sigma^{-1}(\vec{x} - \vec{\mu}_k)} \tag{20}$$

heißt Mahalanobis-Distanz[2] *zwischen dem Endpunkt des Vektors* $\vec{x}$ *und dem Endpunkt des Vektors* $\vec{\mu}_k$.

Der Euklidische Abstand $\|\vec{y} - \vec{\mu}_k(y)\|$ zwischen dem Endpunkt des Vektors $\vec{y}$) und dem von $\vec{\mu}_k(Y)$ ist also nach (19) gleich der Mahalanobis-Distanz des Vektors der Messwerte vom Punkt $\vec{\mu}_k$. Die Mahalanobis-Distanz ist durch die *quadratische Form*

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

erklärt, die bekanntlich ein Ellipsoid definiert, da ja Σ^{-1} eine symmetrische Matrix ist. Der Abstand zwischen den Endpunkten von $\vec{x}$ und $\vec{\mu}_k$ hängt demnach nicht nur, wie bei der Euklidischen Distanz, von der Länge der Vektoren $\vec{x}$ und $\vec{\mu}_k$ ab, sondern auch von deren Orientierung. Hat der Vektor $(\vec{x} - \vec{\mu}_k)$ eine Orientierung, die der der ersten Hauptachse des Ellipsoids entspricht, so ist die Distanz der Endpunkte "größer" (im Sinne der Mahalanobis-Distanz), als wenn dieser Vektor in Richtung der zweiten Hauptachse weist. Eine Spezialfall liegt vor, wenn Σ und damit auch $Sigma^{-1}$ eine Diagonalmatrix ist, d.h. wenn alle Prädiktoren unkorreliert sind; dann wird die Mahalanobis-Distanz zu einer Euklidischen Distanz.

Die Mahalanobis-Distanz tritt bei der Definition der multivariaten Normalverteilung auf. In Satz 2.2 wird aber die Normalverteilung *nicht* vorausgesetzt. Die Beziehung (19) gilt also auch, wenn die Daten nicht normalverteilt sind. Sie zeigt, dass bei optimalen Klassifikationen (im Sinne der Fisherschen Diskrtiminanzanalyse) die Differenzen $\vec{x} - \vec{\mu}_k$ auf eine komplexe Weise mit den Varianzen und Kovarianzen von $\vec{x}$ gewichtet werden müssen, was dazu beitragen dürfte, dass "intuitive", auf "Erfahrung" beruhende Klassifikationen häufig so suboptimal ausfallen.

Zur Anzahl der Diskriminanten: Es gibt so viele 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

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

$\lambda_k > 0$. Für jede Klasse Ω_j, $j = 1, 2, \ldots, g$ existiert ein Vektor μ_j, dessen Komponenten die Erwartungswerte über die Variablen für die j-te Klasse sind. Weiter sei

$$\vec{\bar{\mu}} = \frac{1}{g} \sum_{j=1}^{g} \vec{\mu}_j. \tag{21}$$

Es werden nun die g Vektoren

$$\vec{\mu}_1 - \vec{\bar{\mu}}, \vec{\mu}_2 - \vec{\bar{\mu}}, \ldots, \vec{\mu}_g - \vec{\bar{\mu}} \tag{22}$$

betrachtet. Dann folgt

$$(\vec{\mu}_1 - \vec{\bar{\mu}}) + (\vec{\mu}_2 - \vec{\bar{\mu}}) + \cdots + (\vec{\mu}_g - \vec{\bar{\mu}}) = g\vec{\bar{\mu}} - g\vec{\bar{\mu}} = \vec{0}.$$

Also gilt z.B.

$$(\vec{\mu}_1 - \vec{\bar{\mu}}) = -(\vec{\mu}_2 - \vec{\bar{\mu}}) - \cdots - (\vec{\mu}_g - \vec{\bar{\mu}}),$$

d.h. $(\vec{\mu}_1 - \vec{\bar{\mu}})$ kann als Linearkombination der restlichen Differenzen dargestellt werden. Linearkombinationen der Vektoren (22) definieren Hyperebenen der Dimension $q \leq g-1$. Es sei $\vec{v}$ ein Vektor, der senkrecht auf jedem Vektor $\vec{\mu}_j - \vec{\bar{\mu}}$ und damit auf den Hyperebenen steht. Dann hat man

$$B\vec{v} = \sum_{j=1}^{g} (\vec{\mu}_j - \vec{\bar{\mu}})(\vec{\mu}_j - \vec{\bar{\mu}})'\vec{v} = \sum_{j=1}^{g} (\vec{\mu}_j - \vec{\bar{\mu}})\vec{0} = \vec{0}, \tag{23}$$

denn $(\vec{\mu}_j - \vec{\bar{\mu}})'\vec{v} = \vec{0}$ nach Voraussetzung. Daraus folgt

$$W^{-1}B\vec{v} = \vec{v}, \tag{24}$$

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

Tabelle 1: Maximalzahl zu betrachtender Diskriminanten

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

2.4 Klassifikation nach Fisher versus Klassifikation nach Gauss

Auch wenn die multivariate Normalverteilung zunächst nicht vorausgesetzt wird, so ist es doch nützlich, wenn man sie voraussetzen kann, da dann bestimmte Tests möglich werden. Es wird zunächst die multivariate Normalverteilung eingeführt.

Die p-dimensionale Normalverteilung ist durch

$$f(\vec{x}|\Omega_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu}_k)'\Sigma_k^{-1}(\vec{x}-\vec{\mu}_k)\right) \tag{25}$$

definiert. $\vec{\mu}_k$ ist der Vektor der Erwartungs- (Mittel-)werte der Komponenten von $\vec{x}$ (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[3].

Die Einführung einer Wahrscheinlichkeitsverteilung läßt eine allgemeinere Definition von Diskriminanzfunktionen zu. Zunächst sei an den Satz von Bayes (bedingte Wahrscheinlichkeiten) erinnert: sine A udn B irgendzwei zufällige Ereignisse, so gilt bekanntlich

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}. \tag{26}$$

Setzt man nun für A das Ereignis, dass die Person oder das Objekt zu Klasse Ω_k gehört, und für B, dass der Vektor $\vec{x}$ beobachtet wird, so erhält man aus dieser Beziehung

$$f(\Omega_k|\vec{x}) = f(\vec{x}|\Omega_k)\frac{P(\Omega_k)}{f(\vec{x})}. \tag{27}$$

Hierin ist $P(\Omega_k)$ die *a priori*-Wahrscheinlichkeit dafür, dass eine Person oder ein Objekt aus der Kategorie oder Klasse Ω_k stammt. Nur um die Notation zu vereinfachen werde angenommen, dass nur zwischen zwei Klassen entschieden werden muß (die Betrachtung läßt sich sofort auf den Fall von K Klassen verallgemeinern). Dann kann man den Quotienten

$$\frac{f(\Omega_2|\vec{x})}{f(\Omega_1|\vec{x})} = \frac{f(\vec{x}|\Omega_2)}{f(\vec{x}|\Omega_1}\frac{P(\Omega_2)}{P(\Omega_1)} \tag{28}$$

betrachten; die unbedingte Dichte für $\vec{x}$ hat sich hier herausgekürzt. 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 $p(\Omega_2|x) > p(\Omega_1|\vec{x})$, andernfalls für Ω_1.

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

   $$\text{Entscheide für } \Omega_k, \text{ wenn } p(\Omega_k|\vec{x}) = \max_{1\leq j\leq g} p(\Omega_j|\vec{x}). \tag{29}$$

   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 (**??**) heraus und man erhält die Maximum-Likelihood (ML)-Regel

   $$\text{Entscheide für } \Omega_k, \text{ wenn } f(\vec{x}|\Omega_k) = \max_{1\leq j\leq g} f(\vec{x}|\Omega_j). \tag{30}$$

[3]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, d.h. wenn alle Eigenwerte von Σ und damit Σ^{-1} ungleich Null sind. Determinanten werden im Folgenden nicht weiter benötigt, so dass keine weitere Definition dieser Größe gegeben wird.

Diskriminanzfunktionen: Nach (29) entscheidet man sich für Ω_2, wenn der Quotient $f(\vec{x}|\Omega_2)p(\Omega_2)/(f(\vec{x}|\Omega_1)p(\Omega_1) > 1$ ist, andernfalls entscheidet man für Ω_1, d.h. man entscheidet sich für Ω_2, wenn

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

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

$$\log f(x|\Omega_2) + \log p(\Omega_2) > \log f(x|\Omega_1) + \log p(\Omega_1) \tag{31}$$

geschrieben werden. Es wird die folgende Funktion eingeführt:

Definition 2.3 *Es sei*

$$d_k(\vec{x}) = \log f(\vec{x}|\Omega_k) + \log p(\Omega_k), \quad 1 \le k \le g. \tag{32}$$

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

$$\text{Entscheide für } \Omega_k \text{ (d.h. } \vec{x} \in R_k\text{), wenn } d_k(\vec{x}) = \max_{1 \le j \le g} d_j(x). \tag{33}$$

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

In (19) wird die Fishersche Kriteriumsgröße zur Mahalanobis-Distanz in Beziehung gesetzt. Klassifiziert man gemäß der Gaussverteilung, so ist ergibt sich nach Logarithmierung[4]

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

die Diskriminanzfunktion, wobei $\vec{x}$ der Vektor $\vec{x} = (\vec{x}_1, \ldots, \vec{x}_n)'$ der Beobachtungen ist. Für f wird die multivariate Gauss-Verteilung eingesetzt. Man erhält dann

$$d_k(\vec{x}) = \frac{1}{2}\vec{x}'\Sigma^{-1}\vec{x} - \vec{\mu}_k'\Sigma^{-1}\vec{x} + \frac{1}{2}\vec{\mu}_k'\Sigma^{-1}\vec{\mu}_k - \frac{1}{2}(\log|\Sigma^{-1}| - \log p(\Omega_k)), \tag{34}$$

bzw.

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

[4]Der Logarithmus $\log x$ ist eine monotone Funktion von x,d.h. wird x größer, so auch der Logarithmus von x. Gelegentlich erleichtert bzw. vereinfacht die Logaritmierung die Betrachtungen, wie bei der Normalverteilung: Man muß nicht mehr mit $e^{f(x)}$ rechnen, sondern kann wegen $\log e^{f(x)} = f(x)$ gleich mit $f(x)$ rechnen, wen es nur auf die Größenordnungen ankommt.

da der Term $\vec{x}'\Sigma^{-1}\vec{x}$ für alle d_k identisch ist und daher für die Diskrimination keine Information liefert. Hier ist es sinnvoll, doch noch einmal den ursprünglichen Ausdruck (34) für d_k zu btrachten: subtrahiert man $\vec{x}'\Sigma^{-1}\vec{x}/2$ und addiert man $\log p(\Omega_k)$ auf beiden Seiten, so erhält man

$$d_k(\vec{x}) - \frac{1}{2}\vec{x}'\Sigma^{-1}\vec{x} + \log p(\Omega_k) = -\frac{1}{2}(\vec{x} - \vec{\mu}_k)'\Sigma^{-1}(\vec{x} - \vec{\mu}_k). \tag{35}$$

Nach (19) erhält man dann die Beziehung

$$-d_k(\vec{x}) + \frac{1}{2}\vec{x}'\Sigma^{-1}\vec{x} + \log p(\Omega_k) = \sum_{j=1}^{s}(y_j - \vec{\mu}_j(Y))^2 = \|\vec{y} - \vec{\mu}_k(y)\|^2, \tag{36}$$

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

2.5 Stichprobenumfang und korrelierte Prädiktoren

In den üblichen Lehrbüchern zur Regressions- und Diskriminanzanalyse wird bemerkenswert wenig auf die Problematik (i) eines zu geringen Stichprobenumfangs, und (ii) korrelierender Prädiktoren hingewiesen, obwohl die Problematik seit langem bekannt ist. Das mag daran liegen, dass in vielen Anwendungen, etwa in der Psychologie, die Anzahl der Prädiktoren oft ziemlich klein ist im Vergleich zur Anzahl der "Fälle". Bei der multiplen Regression läßt sich zeigen, dass die Varianzen der Schätzungen für die Regressionsgewichte und die Kovarianzen zwischen ihnen durch $\sigma^2(X'X)^{-1}$ gegeben sind[5], wobei X die Matrix der Prädiktoren ist und σ^2 die Fehlervarianz ist. Sind die Prädiktoren unkorreliert, so ist $X'X$ und damit auch $(X'X)^{-1}$ eine Diagonalmatrix, d.h. die Schätzungen $\hat{b}_j$ für die Regressionsgewichte sind ebenfalls unkorreliert. Für die Interpretation der Regressionsgewichte ist diese Eigenschaft von Bedeutung, da man nun ein Regressionsgewicht unabhängig von den anderen interpretieren kann. Sind sie dagegen korreliert, so sind die Abhängigkeiten zwischen den Schätzungen nur schwer zu durchschauen, was die Interpretation erschwert. Darüber hinaus wird die Varianz der Schätzungen erhöht, was ebenfalls die Interpretation sehr erschwert: bekanntlich ist $X'X = P\Lambda P'$, wobei P die Matrix der Eigenvektoren von $X'X$ ist und Λ die Diagonalmatrix der zugehörigen Eigenwerte. Es läßt sich zeigen, dass korrelierende Prädiktoren implizieren, dass einige Eigenwerte *klein* werden. Nun ist

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

wobei nun in Λ^{-1} die Reziprokwerte der Eigenwerte stehen. Werden einge Eigenwerte klein, so werden die Reziprokwerte *groß*, und damit werden die Abweichungen $\hat{b}_j - b_j$ groß ($\hat{b}_j$ die Schätzung für den j-ten Regressionskoeffizienten b_j). Ein häufig gewählter Ausweg aus dieser Krise ist die Ridge-Regression[6]. In Seber (1977) findet man eine kurze Diskussion.

[5] vergl. etwa Seber, G. A. F.: Linear Regression Analysis, New York 1977, p. 48

[6] Hoerl, A. E., Kennard, R. W. (1970) Ridge Regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1), 55 – 67

Diese Situation überträgt sich auf die Diskriminanzanalyse. Die Gewichte $\vec{u}$ ergeben sich als Eigenvektoren von $W^{-1}B$ (vergl. (15)). W ist die Matrix der Vareianzen und Kovarianzen *innerhalb*, und B die der Varianzen und Kovarianzen *zwischen* den Klassen. Werden die Stichprobenumfänge n_k klein in Relation zu p, so werden die Eigenvektoren Weise beliebig in dem Sinne, dass sie eine Klassifikationsmöglichkeit suggerieren, obwohl diese gar nicht existieren muß ('beliebig' heißt also, dass sie mit den wahren Werten, die die tatsächliche Klassifikationsmöglichkeit anzeigen, kaum noch etwas zu tun haben). Die Varianzen der Gewichte werden viel zu groß, so dass eine Interpretation der Gewichte erschwert, wenn nicht unmöglich wird. Dies ist ein sehr unangenehmes Resultat, denn man möchte ja wissen, welche Prädiktoren tatsächlich zur Klassifikation beitragen und welche nicht.

Eine Faustregel besagt, dass man mindest drei- bis viermal so viele "Fälle" wie Prädiktoren haben sollte. Sind die Prädiktoren korreliert – dies ist ein in Anwendungen in der Medizin und Psychologie häufig vorkommender Fall – so sollte man versuchen, *regularisierte* bzw. *penalisierte* Diskriminanzanalysen zu rechnen[7]. Das Programm R hält ein entsprechendes Modul bereit.

2.6 Statistische Tests

Sind das Kriterium λ und die Gewichte $\vec{u}$ 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

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g, \tag{37}$$

   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 $\vec{y}$ betrachten, für die $\|\vec{y} - \vec{\mu}_k\|^2 = \sum_{j=1}^{s}(y_j - \vec{\mu}_{yjk})^2$ =konstant gilt. Offenbar liegen die Endpunkte all dieser Vektoren auf einer Hyperkugel. Betrachtet man die zu den Y korrespondierende Menge der $\vec{x}$, für die die Mahalanobis-Distanzen $(\vec{x} - \vec{\mu}_j)'\Sigma^{-1}(\vec{x} - \vec{\mu}_j)$ konstant sind, so liegen die Endpunkte der $\vec{x}$ 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

[7] Friedman,J.H. (1989) Regularized Discriminant Analysis. *Journal of the American Statistical Association*, 84, 165 – 175.

Hastie, T., Buja, A., Tibshirani, R. (1995) Penalized disciminant analysis. *The Annals of Statistics*, 23 (1), 73 – 102

Dimensionen interpretiert werden können, unterschiedlich große Varianzanteile haben; sind diese Anteile gleich, so definiert die Mahalanobis-Distanz eine Menge von Hyperkugeln.

3. Die Beziehung (19) gilt andererseits unabhängig von der Annahme der Normalverteilung, denn sie besagt ja nur, daß $\|\vec{y}_j - \vec{\mu}_{y_j}\|^2$ gleich der Mahalanobis-Distanz des durch $\vec{x}$ definierten Punktes von $\vec{\mu}_j$ 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 $\vec{x}$ 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 $\vec{x}$ 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 tatsächllihc 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

$$H_0: \qquad \mu_1 = \mu_2 = \cdots = \mu_g, \tag{38}$$

$$H_1: \qquad \mu_i \neq \mu_j, \text{ für mindestens ein Paar } (i,j) \text{ mit } i \neq j \tag{39}$$

In der einfachen Varianzanalyse hat man den bekannten Test

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

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

$$\Lambda = \frac{|W|}{|B+W|} = |I + W^{-1}B|^{-1}, \tag{40}$$

Wilk's Λ; unter H_0 gilt

$$\Lambda \sim \Lambda(q, N-g, g-1) \tag{41}$$

(Λ-Verteilung von Wilks).

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

$$\epsilon = p(\Omega_1)\epsilon_{12} + p(\Omega_2)\epsilon_{21} \tag{42}$$

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:

$$\epsilon_{12} = \Phi\left(\frac{\log(\pi_1/\pi_2) - \delta^2/2}{\delta}\right) \tag{43}$$

$$\epsilon_{21} = \Phi\left(-\frac{\log(\pi_2/\pi_1) + \delta^2/2}{\delta}\right), \tag{44}$$

wobei

$$\delta = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \tag{45}$$

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

Für die ML-Regel ergeben sich die Fehlerraten gemäß

$$\epsilon_{12} = \epsilon_{21} = \Phi\left(-\frac{\delta}{2}\right). \tag{46}$$

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

$$\hat{d}(x) = (x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2))'S^{-1}((\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2) \tag{47}$$

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:

$$\hat{d}(\bar{x}_1) = (\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2) \tag{48}$$

$$\hat{d}(\bar{x}_2) = -(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2). \tag{49}$$

Dann ist für die Bayes-Regel

$$\hat{\epsilon} = \pi_1\hat{\epsilon}_{12} + \pi_2\epsilon_{21} \tag{50}$$

mit

$$\hat{\epsilon}_{12} = \Phi\left(\frac{\log(\pi_2/\pi_1) - D^2/2}{D}\right), \quad \hat{\epsilon}_{21} = \Phi\left(\frac{-\log(\pi_2/\pi_1) - D^2/2}{D}\right) \tag{51}$$

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

$$\hat{\epsilon}_{12} = \hat{\epsilon}_{21} = \Phi(-D/2). \tag{52}$$

2.7 Diskriminanzanalyse bei kategorialen Daten

2.7.1 Volles multinomiales Modell

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

$$m = \prod_{i=1}^{p} m_i \tag{53}$$

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

$$d_k(x) = p(x|k)p(k), \tag{54}$$

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

$$\hat{\pi}(x, k) = \frac{n(x, k)}{N}, \tag{55}$$

und diese Schätzungen liefern die Diskriminanzfunktionen

$$\hat{d_k}(x) = \hat{\pi}(x, k), \tag{56}$$

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!

2.7.2 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

$$p(x_1, \ldots, x_p|k) = \prod_{i=1}^{p} \pi_{ki}^{x_i}(1 - \pi_{ik}^{1-x_i}) \tag{57}$$

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

$$\begin{aligned} d_k(x) &= \log p(x|k) + \log p(k) \\ &= \sum_{i^1}^{p} x_i \log \pi_{ik} + \sum_{i=1}^{p} (1-x_i)\log(1-\pi_{ik}) + \log p(k) \\ &= \sum_{i=1}^{p} \nu_i x_i + \nu_0, \end{aligned} \tag{58}$$

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.

2.7.3 Parametrisierung in Modellfamilien I: 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

$$\begin{aligned} \log n^{(k)}_{i_1 i_2 i_3} &= \mu^{(k)} + \mu^{(k)}_{1(i_1)} + \mu^{(k)}_{2(i_2)} + \mu^{(k)}_{3(i_3)} \\ &+ \mu^{(k)}_{12(i_1 i_2)} + \mu^{(k)}_{13(i_1 i_3)} + \mu^{(k)}_{23(i_2 i_3)} + \mu^{(k)}_{123(i_1 i_2 i_3)} \end{aligned} \tag{59}$$

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

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

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

$$\begin{aligned} \log n(x|k) &= \nu^{(k)} + \nu^{(k)}_1 x_1 + \nu^{(k)}_2 x_2 + \nu^{(k)}_3 x_3 \\ &+ \nu^{(k)}_{12} x_1 x_2 + \nu^{(k)}_{13} x_3 + \nu^{(k)}_{23} x_2 x_3 + \nu^{(k)}_{123} x_1 x_2 x_3, \end{aligned} \tag{60}$$

wobei $x_i = 0$ oder $x_i = 1$; alternativ kann auch $x_i = 1$ oder $x_i = -1$ gesetzt werden (Effektskalierung). Der Vergleich mit (59) 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

$$d_k(x) = \log p(k) - \log n_k + \log n(x|k), \tag{61}$$

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

(60) 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.

2.7.4 Parametrisierung in Modellfamilien I: Logit-Modelle

Es soll entschieden werden, ob ein Objekt der k-ten Klasse zugeordnet werden soll oder nicht. Dazu kann man die a-posteriori-Wahrscheinlichkeit $p(k|x)$ betrachten. Das entsprechende Logit ist $\log p(k|x)/(1-p(k|x))$, und man erhält

$$\begin{aligned} \log \frac{p(k|x)}{1-p(k|x)} &= \lambda + \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_p x_p \\ &\quad + \lambda_{12} x_1 x_2 + \cdots + \lambda_{12\ldots p} x_1 x_2 \cdots x_p. \end{aligned} \tag{62}$$

Dies ist wieder das saturierte Modell. Über die logistische Regression bestimmt man nun das bestpassende und sparsamste Modell und testet es gegen das vollständige Unabhängigkeitsmodell

$$\log \frac{p(k|x)}{1-p(k|x)} = \lambda + \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_p x_p.$$

Man entscheidet sich für die k-te Klasse, wenn $\log p(k|x)/(1-p(k|x)) > 0$ ist, sonst für die Komplementärklasse.

3 Beispiele

Beispiel 3.1 Bei 832 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 2: 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

$$S = \begin{pmatrix} 214.26 & 72.37 & 195.61 \\ 72.73 & 212.44 & 175.53 \\ 195.61 & 175.53 & 1820.61 \end{pmatrix}. \tag{63}$$

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

$$S^{-1} = \begin{pmatrix} 214.26 & .72.37 & 195.61 \\ 72.73 & 212.44 & 175.53 \\ 195.61 & 175.53 & 1820.61 \end{pmatrix}. \tag{64}$$

Die Gewichte der Variablen sind Es ergibt sich ein Gesamt-F-Wert von $F = 17.605$;

Tabelle 3: 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		

bei $df = n_1 + n_2 - 3 - 1 = 828$ 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%.

Tabelle 4: Klassifikationen

Klassifikation nach ML			
	Klassifizierung		Σ
Lernstichprobe	krank	gesund	
Kranke	51	20	71
Gesunde	272	489	767
Klassifikation nach Bayes			
	Klassifizierung		Σ
Lernstichprobe	krank	gesund	
Kranke	0	71	71
Gesunde	0	761	761

Die Mahalanobis-Distanz ist gemäß Tabelle 3 $\delta^2 = .815$. 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 3.2 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 5. Ein hoher Wert auf einer Skala reflektiert eine hohe Präferenz für die entsprechende Aktivität.Die Tabelle 6 gibt die Mittelwerte der drei Variablen für die einzelnen Gruppen.

Die zusammengefasste ("pooled") Varianz-Kovarianz-Matrix wird in der Tabelle 7 angegeben. Die Matrix B der Varianzen-Kovarianzen zwischen den Mittelwerten findet man in der Tabelle 8. Die Tabelle 9 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 Bemerkungen über die Signifikanz der einzelnen Eigenwerte machen! Der Abbildung 3 entsprechend kann man folgern, dass in erster Linie die Gruppe

Abbildung 3: Projektion auf die maximal diskriminierende Achse u_1

o, also die Operational Conrol Agents, von den übrigen beiden Gruppen unterscheidet, während sich die Gruppen p (Passenger Agents) und m (Mechanics) kaum voneinander unterscheiden. Nach Tabelle 9 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". □

Tabelle 5: Die Freizeitinteressen von Angestellten einer Fluglinie, p: Passenger Agents, m Mechanics, o Operations control (Beispiel 3.2)

Person	X_1	X_2	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 6: Mittelwerte der drei Variablen für die drei Gruppen

	X_1	X_2	X_3
p	13.200	23.500	9.00
m	16.461	21.000	13.231
o	13.214	14.714	13.857

Tabelle 7: Varianz-Kovarianz-Matrix W

	X_1	X_2	X_3
X_1	609.188	-10.143	13.044
X_2	-10.143	343.357	148.429
X_3	13.044	-217.996	154.086

Tabelle 8: 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

Tabelle 9: Eigenvektoren $\vec{u}$ und Eigenwerte λ

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

4 Anhang: Ungleichungen, Maxima und Beweise

4.1 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 man 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 Eigenwerte. Dann gilt

$$\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}' = \sum_{k=1}^{n} \lambda_k \mathbf{p}_k \mathbf{p}_k'. \tag{65}$$

Es ist

$$\mathbf{A}^{-1} = (\mathbf{P}\Lambda\mathbf{P}')^{-1} = (\mathbf{P}')^{-1}\Lambda^{-1}\mathbf{P}^{-1}, \tag{66}$$

und da $\mathbf{P}^{-1} = \mathbf{P}'$ und $(\mathbf{P}')^{-1} = \mathbf{P}$, hat man

$$\mathbf{A}^{-1} = \mathbf{P}\Lambda^{-1}\mathbf{P}'. \tag{67}$$

Es sei $\Lambda^{-1/2} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$. Man definiert nun

$$\mathbf{A}^{1/2} = \sum_{k=1}^{n} \sqrt{\lambda_k}\mathbf{p}_k\mathbf{p}' = \mathbf{P}\Lambda^{-1/2}\mathbf{P}'. \tag{68}$$

Die Matrix $A^{1/2}$ hat die zu $\sqrt{a}$, $a \in \mathbb{R}$, analogen Eigenschaften:

$$(\mathbf{A}^{1/2})' = \mathbf{A}^{1/2}, \quad (\mathbf{A}^{1/2} \text{ ist symmetrisch}) \tag{69}$$

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A} \tag{70}$$

$$(\mathbf{A}^{1/2})^{-1} = \sum_{k=1}^{n} \frac{1}{\sqrt{\lambda_k}}\mathbf{p}_k\mathbf{p}_k' = \mathbf{P}\Lambda^{-1/2}\mathbf{P}' \tag{71}$$

$$\mathbf{A}^{1/2}\mathbf{A}^{-1/2} = \mathbf{I}, \quad \mathbf{A}^{-1/2}\mathbf{A}^{-1/2} = \mathbf{A}^{-1}. \tag{72}$$

4.2 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

$$(\mathbf{a}'\mathbf{b}) \leq (\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b}), \tag{73}$$

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

$$\begin{aligned} 0 < |\mathbf{a} - x\mathbf{b}|^2 = (\mathbf{a} - x\mathbf{b})'(\mathbf{a} - x\mathbf{b}) &= \mathbf{a}'\mathbf{a} - x\mathbf{b}'\mathbf{a} + \mathbf{a}'x\mathbf{b} + x\mathbf{b}'x\mathbf{b} \\ &= \mathbf{a}'\mathbf{a} - 2x(x\mathbf{b}'\mathbf{a}) + x^2(\mathbf{b}'x\mathbf{b}). \end{aligned} \tag{74}$$

$|\mathbf{a} - x\mathbf{b}|^2$ 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 $\mathbf{b} \neq x\mathbf{a}$. □

4.3 Verallgemeinerte Cauchy-Schwarzsche Ungleichung

Es seien $\mathbf{a}$ und $\mathbf{b}$ irgend zwei n-dimensionale Vektoren, und $\mathbf{A}$ sei eine positiv-definite $(n \times n)$-Matrix, d.h. alle Eigenwerte λ_k, $1 \leq k \leq n$ sind größer als Null. Dann gilt

$$(\mathbf{a}'\mathbf{x})' \leq (\mathbf{a}'\mathbf{B}\mathbf{a})(\mathbf{b}'\mathbf{B}^{-1}\mathbf{b}), \tag{75}$$

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

$$\mathbf{a}'bvec = \mathbf{a}'Ibvec = \mathbf{a}'B^{1/2}B^{-1/2}bvec = (\mathbf{B}^{1/2}\mathbf{a})'(\mathbf{B}^{-1/2}bvec).$$

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

$$\max_{\mathbf{x}\neq 0} \frac{\mathbf{x}'\mathbf{a}}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}, \tag{76}$$

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

$$(\mathbf{x}\mathbf{a})^2 \leq (\mathbf{x}\mathbf{B}\mathbf{x})(\mathbf{a}'\mathbf{B}^{-1}\mathbf{a}).$$

Es ist $\mathbf{x}'\mathbf{B}\mathbf{x} > 0$, denn $\mathbf{B}$ ist positiv-definit und $\mathbf{x} \neq 0$. Nimmt man den Vektor $\mathbf{x}$, für den das Maximum erreicht wird, erhält man die obere Grenze

$$\frac{(\mathbf{x}'\mathbf{a})^2}{\mathbf{x}'\mathbf{B}\mathbf{x}} \leq \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}.$$

Für $\mathbf{x} = c\mathbf{B}^{-1}\mathbf{a}$ folgt (76). □

4.4 Die Maximierung quadratischer Formen

Es sei $\mathbf{B}$ eine postiv-definite $(n \times n)$-Matrix mit den Eigenwerten $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ und den dazu koauf die Länge 1 korrespondierenden Eigenvektoren $\mathbf{p}_1, \ldots, \mathbf{p}_n$. Dann gilt

$$\max_{\mathbf{x}\neq 0} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_1 \quad \text{für } \mathbf{x} = \mathbf{p}_1 \tag{77}$$

$$\min_{\mathbf{x}\neq 0} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_n, \quad \text{für } \mathbf{x} = \mathbf{p}_n \tag{78}$$

$$\max_{\mathbf{x}\perp \mathbf{p}_1,\ldots,\mathbf{p}_k} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_{k+1}, \quad \text{für } \mathbf{x} = \mathbf{p}_{k+1}, \quad k = 1, \ldots n-1, \tag{79}$$

wobei "$\perp$" für "ist orthogonal" steht.

Beweis: Es sei $\mathbf{P}$ die Matrix der Eigenvektoren von $\mathbf{B}$ und Λ die Diagonalmatrix der Eigenwerte von $\mathbf{B}$. Dann gilt $\mathbf{B} = \mathbf{P}\Lambda\mathbf{P}'$ und $\mathbf{B}^{1/2} = \mathbf{P}\Lambda^{1/2}\mathbf{P}'$. Es sei $\mathbf{y} = \mathbf{P}'\mathbf{x}$, $xvwc \neq 0$. Dann ist

$$\begin{aligned}
\frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} &= \frac{\mathbf{x}'\mathbf{B}^{1/2}\mathbf{B}^{1/2}\mathbf{x}}{\mathbf{x}'\mathbf{P}'\mathbf{P}\mathbf{x}} \\
&= \frac{\mathbf{x}'\mathbf{P}\Lambda^{1/2}\mathbf{P}'\mathbf{x}}{\mathbf{y}'\mathbf{y}} \\
&= \frac{\mathbf{y}'\Lambda\mathbf{y}}{\mathbf{y}'\mathbf{y}} \\
&= \frac{\sum_{k=1}^n \lambda_k y_k^2}{\sum_{k=1}^n y_k^2} \leq \lambda_1 \frac{\sum_{k=1}^n y_k^2}{\sum_{k=1}^n y_k^2} = \lambda_1.
\end{aligned} \tag{80}$$

Für $\mathbf{x} = \mathbf{p}_1$ erhält man

$$\mathbf{y} = \mathbf{P}'\mathbf{p}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

denn

$$\mathbf{p}_k'\mathbf{p}_1 = \begin{cases} 1, & k = 1 \\ 0, & k \neq 1 \end{cases}$$

Für diese Wahl von $\mathbf{x}$ hat man $\mathbf{y}'\Lambda\mathbf{y}'\mathbf{y} = \lambda_1/1 = \lambda_1$, d.h.

$$\frac{\mathbf{p}_1'\mathbf{B}\mathbf{p}_1}{\mathbf{p}_1'\mathbf{p}_1} = \mathbf{p}_1'\mathbf{B}\mathbf{p}_1 = \lambda_1. \tag{81}$$

Die Gleichung $\mathbf{x} = \mathbf{P}\mathbf{y}$ kann in der Form

$$\mathbf{x} = \mathbf{P}\mathbf{y} = y_1\mathbf{p}_1 + y_2\mathbf{p}_2 + \cdots + y_n\mathbf{p}_n$$

geschrieben werden, so dass $\mathbf{x} \perp \mathbf{p}_1, \ldots, \mathbf{p}_k$ die Gleichung

$$0 = \mathbf{p}_j'\mathbf{x} = y_1\mathbf{p}_j'\mathbf{p}_1 + y_2\mathbf{p}_j'\mathbf{p}_2 + \cdots + y_n\mathbf{p}_j'\mathbf{p}_n = y_j, \quad j \leq k$$

impliziert. Ist also $\mathbf{x}$ orthogonal zu den ersten k Eigenvektoren $\mathbf{p}_j$, so nimmt die linke Seite von (81) die Form

$$\frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \frac{\sum_{j=k+1}^{n} \lambda_j y_j^2}{\sum_{j=k+1}^{n} y_j^2}.$$

Für $y_{k+1} = 1$, $y_{k+2} = \cdots = y_n = 0$ folgt die Behauptung. □

4.5 Beweis von Satz 2.2

Beweis: W ist eine Schätzung von Σ. Es wird zuerst gezeigt, daß $W^{-1}B = \Sigma^{-1}B$ und $\Sigma^{-1/2}B\Sigma^{-1/2}$ die gleichen Eigenwerte haben, und die Eigenvektoren $\vec{e}_i$ von $\Sigma^{-1/2}B\Sigma^{-1/2}$ in der Beziehung $u_i = \Sigma^{-1/2}\vec{e}_i$ zu den Eigenvektoren u_i von $\Sigma^{-1}B$ stehen. Es sei also $(\lambda, \vec{e})$ ein Eigenwert und der zugehörige Eigenvektor von $\Sigma^{-1/2}B\Sigma^{-1/2}$, d.h. es gelte

$$\Sigma^{-1/2}B\Sigma^{-1/2}\vec{e} = \lambda\vec{e}.$$

Multiplikation von links mit $\Sigma^{-1/2}$ liefert dann

$$\Sigma^{-1/2}\Sigma^{-1/2}B\Sigma^{-1/2} = \Sigma^{-1}B\Sigma^{-1/2}\vec{e} = \lambda\Sigma^{-1/2}\vec{e}.$$

Aber $\Sigma^{-1/2}\vec{e}$ ist ein Vektor, und die letzte Gleichung bedeutet dann, dass es sich um einen Eigenvektor von $\Sigma^{-1}B$ handelt, d.h. um einen Eigenvektor $\vec{u}$ von $\Sigma^{-1}B$; ist $E = [\vec{e}_1, \ldots, \vec{e}_s]$ die Matrix der normierten Eigenvektoren von $\Sigma^{-1/2}B\Sigma^{-1/2}$, so ist $\Sigma^{-1/2}E$ die Matrix der Eigenvektoren von $\Sigma^{-1}B$, und λ ist offenbar ein Eigenwert sowohl von $\Sigma^{-1}B$ als auch von $\Sigma^{-1/2}B\Sigma^{-1/2}$. Dann ist

$$\begin{aligned}(\vec{x} - \vec{\mu}_j)'\Sigma^{-1}(\vec{x} - \vec{\mu}_j) &= (\vec{x} - \vec{\mu}_j)'\Sigma^{-1/2}\Sigma^{-1/2}(\vec{x} - \vec{\mu}_j) \\ &= (\vec{x} - \vec{\mu}_j)'\Sigma^{-1/2}EE'\Sigma^{-1/2}(\vec{x} - \vec{\mu}_j), \quad EE' = I\end{aligned}$$

Nach dem vorher gezeigten ist aber $\Sigma^{-1/2}E = U = [u_1, \ldots, u_p]$ die Matrix der Eigenvektoren von $\Sigma^{-1}B$, so daß man

$$E'\Sigma^{-1/2}(\vec{x} - \vec{\mu}_j) = \begin{pmatrix} \vec{u}_1\,'(\vec{x} - \vec{\mu}_j) \\ \vec{u}_2\,'(\vec{x} - \vec{\mu}_j) \\ \vdots \\ \vec{u}_p\,'(\vec{x} - \vec{\mu}_j) \end{pmatrix}$$

schreiben kann. Dies heißt aber

$$\begin{aligned}(\vec{x} - \vec{\mu}_j)'\Sigma^{-1/2}EE'\Sigma^{-1/2}(\vec{x} - \vec{\mu}_j) &= (\vec{x} - \vec{\mu}_j)'\Sigma^{-1/2}\Sigma^{-1/2}(\vec{x} - \vec{\mu}_j) \\ &= (\vec{x} - \vec{\mu}_j)'\Sigma^{-1}(\vec{x} - \vec{\mu}_j) \\ &= \sum_{j=1}^{p}(u_j'(\vec{x} - \vec{\mu}_j))^2\end{aligned}$$

□