Klassifikations- und Diskriminanzanalyse[1]

Skript zur Veranstaltung *Evaluation und Forschungsmethoden*

U. Mortensen

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

[1] Version 1. September 2014

Inhaltsverzeichnis

1 Klassifizieren und diskrimininieren

1.1 Einführung

Aus praktischen oder theoretischen Gründen kann es von Interesse sein, Personen oder Objekte einer von mehreren möglichen Klassen $\Omega_1, \ldots, \Omega_g$ zuzuordnen. So muß ein Arzt entscheiden, ob ein Patient "krank" (Klasse Ω_1) oder "gesund" (Klasse Ω_2) ist, oder ein Therapeut muß entscheiden, ob die Depression einer Patientin "endogen" (Klasse Ω_1) oder situationsbedingt (Klasse Ω_2) ist. Die jeweilige Zuordnung wird von der Beobachtung bestimmter Merkmale abhängen. Es seien insbesonderer p solche Merkmale in Form von Messungen $x_1, \ldots, x_p$ gegeben. Die Sicherheit der Zuordnung wiederum hängt davon ab, wie gut sich die Objekte oder die Personen anhand der beobachteten Merkmale in bezug auf die Klassen trennen lassen, d.h. wie gut sich die Objekte oder Personen aufgrund der Messungen jeweils einer der Klassen zuordnen lassen. Es gibt zwei mögliche Ansätze, diese Frage zu diskutieren:

1. Man nimmt eine bestimmte bedingte Wahrscheinlichkeitsdichte

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

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

2. Man kann eine verteilungsfreie Methode zur Entscheidung über die Zugehörigkeit eines Objekts oder einer Person wählen. Diese Methode wurde von Fisher (1936) eingeführt. Diese Methode besteht im Wesentlichen darin, eine neue Skala Y zu bestimmen, in bezug auf die sich die Klassen maximal unterscheiden, wobei $Y = u_1x_1 + u_2x_2 + \cdots + u_px_p$ gelten soll. Die "Gewichte" $u_1, u_2, \ldots, u_p$ sind zunächst unbekannt und werden so bestimmt, dass die maximal mögliche Separation der Klassen in bezug auf die Y-Werte erreicht wird. Dabei kann es sein, dass nicht nur eine solche Variable Y bestimmt werden muß, sondern mehrere $Y_1, \ldots, Y_r$.

Es soll zunächst der erste Ansatz beschrieben werden, weil mit diesem Ansatz die allgemeinen entscheidungstheoretischen Grundlagen verbunden sind, in bezug auf die auch die Entscheidungsfindung nach Maßgabe des zweiten Ansatzes diskutiert werden kann.

1.2 Entscheidungsregeln

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

folgt angeschrieben werden:

$$D(\vec{x}) = \begin{cases} D_1, & \vec{x} \in R_1 \\ & \\ D_2, & \vec{x} \in R_2 \end{cases}, \quad R_1 \cup R_2 = R, \quad R_1 \cap R_2 = \emptyset, \tag{1}$$

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

$$\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_g, \quad \Omega_i \cap \Omega_j = \emptyset, \quad i \neq j. \tag{2}$$

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

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

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

$$\begin{aligned} E(K) &= K_{11}P(D_1|\Omega_1)p(\Omega_1) + K_{12}p(D_2|\Omega_1)p(\Omega_1) \\ &\quad + K_{21}P(D_1|\Omega_2)p(\Omega_2) + K_{22}P(D_2|\Omega_2)p(\Omega_2). \end{aligned} \tag{3}$$

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

$$P(D_1|\Omega_1) = \int_{R_1} f_1(\vec{x})\, dX, \quad P(D_1|\Omega_2) = \int_{R_1} f_2(\vec{x})\, dx. \tag{4}$$

$$P(D_2|\Omega_1) = 1 - P(D_1|\Omega_2), \quad P(D_2|\Omega_2) = 1 - P(D_1|\Omega_2). \tag{5}$$

Diese Ausdrücke können in (3) eingesetzt werden; fasst man die korrespondierenden Ausdrücke zusammen, so ergibt sich

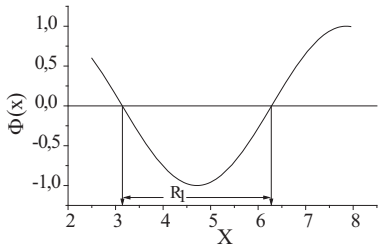
$$\begin{aligned} E(K) &= P(\Omega_1)K_{21} + (1 - P(\Omega_1))K_{22} \\ &+ \int_{R_1} ((1 - P(\Omega_1))(K_{12} - K_{22})f_2(x) - p(\Omega_1)(K_{21} - K_{11})f_1(x)\, dX. \end{aligned} \tag{6}$$

Minimierung der erwarteten Kosten: Da die Kosten und die a priori Wahrscheinlichkeiten festliegen, sind die ersten beiden Terme auf der rechten Seite fest. Um $E(K)$ zu minimieren, muß der Bereich R_1 geeignet gewählt werden. Die Funktion, für die das Integral über R_1 in (6) berechnet werden soll, ist

$$\phi(\vec{x}) = ((1 - P(\Omega_1))(K_{12} - K_{22})f_2(\vec{x}) - p(\Omega_1)(K_{21} - K_{11})f_1(\vec{x}). \tag{7}$$

Ein möglicher Verlauf von $\phi(\vec{x})$ wird in Abb. 1 gezeigt. Das Integral der Funktion ϕ ist die Differenz der Teilintegrale über (i) den Bereich, in dem $\phi > 0$ ist, und (ii) den Bereich,

Abbildung 1: Zur Bestimmung des Integrationsbereichs



in dem $\phi < 0$ ist. Das Integral wird dann minimal, wenn man nur über den Bereich integriert, in dem $\phi < 0$ ist; dies ist der Bereich R_1. Ist ϕ wie in (7) definiert, so ist $\phi < 0$ wenn die Ungleichung

$$(1 - P(\Omega_1))(K_{12} - K_{22})f_2(\vec{x}) < p(\Omega_1)(K_{21} - K_{11})f_1(\vec{x}) \tag{8}$$

gilt. Für jeden x-Wert aus dem so definierten Bereich R_1 gilt demnach (8). Die Ungleichung läßt sich wie folgt umformen:

$$\frac{f_2(\vec{x})}{f_1(\vec{x})} < \frac{p(\Omega_1)}{1 - p(\Omega_1)} \frac{(K_{21} - K_{11})}{(K_{12} - K_{22})} \tag{9}$$

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

Definition 1.1 *Es sei $f(x|\Omega_i)$ die Dichte von x unter der Bedingung Ω_i, d.h. die* Likelihood *von x unter der Bedingung Ω_i, $i = 1, 2$. Dann heißt*

$$\lambda(\vec{x}) = \frac{f(\vec{x}|\Omega_2)}{f(\vec{x}|\Omega_1)} \tag{10}$$

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

Die Entscheidung nach Maßgabe von (9) ist dann eine Entscheidung anhand des Likelihood-Quotienten: setzt man

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

so entscheidet man nach der Regel

$$\lambda(\vec{x}) < \lambda_0 \Rightarrow D(\vec{x}) = D_1, \quad \lambda(\vec{x}) \geq \lambda_0 \Rightarrow D(\vec{x}) = D_2, \tag{12}$$

wobei nach (1) D_1 die Entscheidung für Ω_1 und D_2 die Entscheidung für Ω_2 bdeutet, so wird man *332im Durchschnitt* die Kosten minimieren.

Entscheidungsregeln: Die Beziehung (10) definiert bereits die allgemeine Entscheidungsregel: entscheide auf der Basis der Daten x für Ω_2, wenn der Likelihood-Quotient $\lambda(\vec{x})$ größer als λ_0 ist, und für Ω_1, wenn er kleiner als λ_0 ist. Der "kritische" Wert λ_0 wird (i) durch die a priori-Wahrscheinlichkeiten $p_1 = p(\Omega_1)$ und $p_2 = p(\Omega_2)$, und (ii) durch die Kosten K_{ij} bestimmt. Allerdings hat man das Problem, die Kosten explizit angeben zu müssen, damit die Regel (10) angewendet werden kann. Dies ist in vielen Fällen nicht möglich. Ein Ausweg aus dieser Lage ergibt sich, wenn man die Annahme

$$\left(\frac{K_{21} - K_{11}}{K_{12} - K_{22}}\right) = 1. \tag{13}$$

macht. Diesen Fall hat man insbesondere dann, wenn man $K_{ii} = 0$ und $K_{ij} = K_{ji}$ annehmen kann, wenn also korrekte Entscheidungen keine Kosten und Fehlentscheidungen gleiche Kosten verursachen. Diese Annahmen sind nicht in allen Fällen plausibel, aber sie sind gleichbedeutend mit dem Ansatz, die Kosten nicht explizit in Rechnung zu stellen. Jedenfalls ist, wenn (13) gilt, $\lambda_0 = p(\Omega_1)/p(\Omega_2)$. Gilt $\lambda(\vec{x}) > \lambda_0$, so folgt aus (10)

$$\lambda(\vec{x})\frac{p(\Omega_2)}{p(\Omega_1)} = \frac{f(\vec{x}|\Omega_2)}{f(\vec{x}|\Omega_1)}\frac{p(\Omega_2)}{p(\Omega_1)} > 1; \tag{14}$$

eine analoge Aussage gilt für $\lambda(\vec{x}) \leq \lambda_0$. Aber $f(\vec{x}|\Omega_1)p(\Omega_1)$ und $f(\vec{x}|\Omega_2)p(\Omega_2)$ entsprechen nach dem Satz von Bayes den a posteriori-Wahrscheinlichkeiten $f(\Omega_1|\vec{x})$ und $f(\Omega_2|\vec{x})$, so daß (14) dem Quotienten

$$\frac{f(\Omega_2|\vec{x})}{f(\Omega_1|\vec{x})} = \frac{f(\vec{x}|\Omega_2)}{f(\vec{x}|\Omega_1)}\frac{p(\Omega_2)}{p(\Omega_1)} \tag{15}$$

entspricht. Dieser Quotient führt zu den beiden folgenden Entscheidungsregeln:

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

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

   $$\text{Entscheide für } \Omega_k, \text{ wenn } p(\Omega_k|\vec{x}) = \max_{1\leq j\leq g} p(\Omega_j|\vec{x}). \tag{16}$$

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

2. **Maximum-Likelihood (ML)-Regel:** Gelegentlich sind die a priori-Wahrscheinlichkeiten nicht bekannt; man kann dann den Fall gleicher a priori-Wahrscheinlichkeiten annehmen. Die a priori-Wahrscheinlichkeiten kürzen sich dann in (15) heraus und man erhält die Maximum-Likelihood (ML)-Regel

   $$\text{Entscheide für } \Omega_k, \text{ wenn } f(\vec{x}|\Omega_k) = \max_{1\leq j\leq g} f(\vec{x}|\Omega_j). \tag{17}$$

Diskriminanzfunktionen: Nach (14) und (16) entscheidet man h 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

$$\log f(x|\Omega_2) + \log p(\Omega_2) > \log f(x|\Omega_1) + \log p(\Omega_1) \tag{18}$$

geschrieben werden. Es wird die folgende Funktion eingeführt:

Definition 1.2 *Es sei*

$$d_k(\vec{x}) = \log f(\vec{x}|\Omega_k) + \log p(\Omega_k), \quad 1 \leq k \leq g. \tag{19}$$

d_k *heißt dann* Diskriminanzfunktion.

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

$$\text{Entscheide für } \Omega_k \text{ (d.h. } \vec{x} \in R_k), \text{ wenn } d_k(\vec{x}) = \max_{1 \leq j \leq g} d_j(x). \tag{20}$$

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

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

$$d_j(\vec{x}_0) = d_k(\vec{x}_0), \quad j \neq k. \tag{21}$$

$\vec{x}_0$ trennt dann die Bereiche von Datenvektoren x, für die man sich für Ω_j oder für Ω_k entscheidet. Die Menge der Vektoren x, die der Gleichung (21) genügt, bildet im allgemeinen Fall eine (Hyper-)Fläche im p-dimensionalen Raum, wenn p die Anzahl der Komponenten des Vektors $\vec{x}_0$ ist. Diese Flächen werden durch die Art der Dichten $f(x|\Omega)$ definiert, wobei man sich i.a. auf die multivariate Normalverteilung konzentriert, die in Abschnitt 1.3 eingeführt wird.

Im allgemeinen Fall ist $g > 2$; um sich für eine Klasse Ω_k zu entscheiden, muß man im Prinzip $\binom{g}{2} = g(g-1)/2$ Vergleiche durchführen. Andererseits wird durch die Bedingungen (21) der Raum in Teilräume R_k, $k = 1, 2, \ldots, g$ aufgeteilt; findet man $\vec{x} \in R_k$, so wird man sich für Ω_k entscheiden. Ist z.B. $g = 3$, so gibt es die Teilräume R_1, R_2 und R_3. Die R_k sind durch die Bedingungen

$$d_1(\vec{x}) = d_2(\vec{x}) \text{ und } d_1(\vec{x}) = d_3(\vec{x}) \tag{22}$$

$$d_2(\vec{x}) = d_1(\vec{x}) \text{ und } d_2(\vec{x}) = d_3(\vec{x}) \tag{23}$$

$$d_3(\vec{x}) = d_1(\vec{x}) \text{ und } d_3(\vec{x}) = d_2(\vec{x}) \tag{24}$$

definiert.

Fehlerraten: Alle hier betrachteten Entscheidungen sind probabilistisch und damit kann die Möglichkeit einer Fehlentscheidung nicht ausgeschlossen werden. Dementsprechend

kann man die Fehlerrate bestimmen. Dazu sei T die Menge der Werte, die x überhaupt annehmen kann. Jede Entscheidungsregel definiert implizit einen Teilbereich T_k derart, dass man für Ω_k entscheidet, wenn $\vec{x} \in T_k$. Für den Fall, dass man nur zwischen zwei Möglichkeiten entscheiden muß, macht man also einen Fehler, wenn man für Ω_1 entscheidet, obwohl Ω_2 zutrifft, und umgekehrt. Die Wahrscheinlichkeit eines Fehlers ist dann durch

$$\epsilon = \int_{T_2} f(\vec{x}|\Omega_1)p(\Omega_1)d\vec{x} + \int_{T_1} f(\vec{x}|\Omega_2)p(\Omega_2)d\vec{x} \tag{25}$$

gegeben.

1.3 Klassifikation und die multivariate Normalverteilung

1.3.1 Multivariate Normalverteilung und Mahalanobis-Distanz

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

$$f(\vec{x}|\Omega_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu}_k)'\Sigma_k^{-1}(\vec{x} - \vec{\mu}_k)\right) \tag{26}$$

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[2].

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

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

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

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

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

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

Definition 1.3 *Die Größe*

$$\delta(\vec{x}, \vec{\mu}_k) = \sqrt{(\vec{x} - \vec{\mu}_k)' \Sigma_k^{-1} (\vec{x} - \vec{\mu}_k)} \tag{27}$$

heißt Mahalanobis-Distanz[3] *zwischen den durch die Endpunkte der Vektoren $\vec{x}$ und $\vec{\mu}_k$ definierten Punkten.*

Anmerkungen zum Begriff der Mahalanobis-Distanz:

1. Die Menge der $\vec{x}$, für die $\delta(\vec{x}, \vec{\mu}_k) =$ konstant ist, hat nach (26) die gleiche Dichte, d.h. $\delta(\vec{x}, \vec{\mu}_k) =$ konstant definiert einen geometrischen Ort gleicher Wahrscheinlichkeit[4].

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

   $$\delta^2 = (\vec{x} - \vec{\mu}_k)' \Sigma_k^{-1} (\vec{x} - \vec{\mu}_k) = \vec{\xi}_k{}' \Sigma_k^{-1} \vec{\xi}_k \tag{28}$$

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

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

   $$\Sigma = (Y\Lambda Y')^{-1} = (Y')^{-1} \Lambda^{-1} Y^{-1} = Y\Lambda^{-1} Y', \tag{29}$$

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

[3] Mahalanobis, P.C. (1936) On the generalized distance in statistics. Proc. Nat. Inst. Sci. Calcutta, 12, 49-55

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

Σ^{-1}. Insbesondere gilt dann

$$\Sigma^{-1}\vec{y}_1 = (1/\lambda_1)\vec{y}_1, \quad \Sigma^{-1}\vec{y}_2 = (1/\lambda_2)\vec{y}_2. \tag{30}$$

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

$$\delta_1^2 = \vec{y}_1'\Sigma^{-1}\vec{y}_1 = 1/\lambda_1, \quad \delta_2^2 = \vec{y}_2'\Sigma^{-1}\vec{y}_2 = 1/\lambda_2, \tag{31}$$

d.h. die Quadrate der Mahalanobis-Distanzen für die Eigenwerte $\vec{y}_1$ und $\vec{y}_2$ sind gerade durch die Reziprokwerte der Eigenwerte von Σ gegeben, d.h. aber $\delta_1 = 1/\sqrt{\lambda_1}$ und $\delta_2 = 1/\sqrt{\lambda_2}$. Da λ_1 und λ_2 Eigenwerte von Σ sind, gilt $\sqrt{\lambda_1} \geq \sqrt{\lambda_2}$, und mithin $\delta_1 < \delta_2$. Nun sind die Längen der Hauptachsen der durch die Varianz-Kovarianz-Matrix Σ definierten Ellipsen stets proportional zu $1/\lambda_k$, $k = 1, 2$, vergl. das Skriptum Faktorenanalyse, Seite 28, d.h. die Länge der ersten Hauptachse ist gleich $a_1 = \sqrt{k_0/\lambda_1}$, die der zweiten ist gleich $a_2 = \sqrt{k_0/\lambda_2}$. Die Mahalanobis-Distanzen für die Eigenvektoren sind gerade proportional zu den Längen der Hauptachsen. Die Konstante $k_0 = 1$ korrespondiert dann zur Länge der Eigenvektoren.

Obwohl also $\vec{y}_1$ und $\vec{y}_2$ die gleiche Länge haben, sind die zugehörigen Mahalanobis-Distanzen verschieden. Die Mahalanobis-Distanz zum Endpunkt von $\vec{y}_1$ ist kleiner als die zum Endpunkt von $\vec{y}_2$; es läßt sich zeigen, dass die Mahalanobis-Distanz für einen Punkt, der auf der durch $\vec{y}_1$ liegenden Geraden liegt und einen euklidischen Abstand von d_e vom Zentrum der Ellipse hat, minimal ist relativ zu den Mahalanobis-Distanzen für Punkte mit gleichem euklidischen Abstand d_e vom Zentrum dr Ellipse, aber mit einer von $\vec{y}_1$ abweichenden Orientierung. Die Mahalanobis-Distanz wird maximal, wenn der Punkt mit Abstand d_e auf der Geraden liegt, deren Richtung mit der von $\vec{y}_2$ zusammen fällt. Die Lage der Ellipse, die durch Σ beschrieben wird, wird ja durch die korrelative Beziehung zweier Variablen, etwa X_1 und X_2, bestimmt. Für einen gegebenen Wert x_1 von X_1 ist dann derjenige X_2-Wert am wahrscheinlichsten, für den $x_2 = b_1x_1 + b_0$ gilt, wobei $b_1 = r(s_1/s_2)$ und b_0 die Regressionskoeffizienten sind. Der Punkt hat die euklidische Distanz $d_e = [(x_1 - \mu_1)^2 + (x_2 - \mu_2)^2]^{1/2} = (\xi_1^2 + \xi_2^2)^{1/2}$. Punkte (ξ_1', ξ_2'), die auf einer Geraden liegen, deren Richtung sich von der der Regressionsgeraden unterscheidet, haben eine geringere Wahrscheinlichkeit, wenn sie die gleiche euklidische Distanz d_e von (μ_1, μ_2) haben. Die Wahrschlichkeit wird minimal (relativ zu d_e), wenn sie auf einer Geraden liegen, deren Orientierung orthogonal zu der der Regressionsgeraden ist.

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

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}, \quad \sigma_{12} = \sigma_{21}. \tag{32}$$

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

$$\Sigma^{-1} = \begin{pmatrix} \frac{\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, & -\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \\ -\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, & \frac{\sigma_1^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma_1^2(1-r^2)} & -\frac{\sigma_2/\sigma_1}{1-r^2} \\ -\frac{\sigma_2/\sigma_1}{1-r^2} & \frac{1}{\sigma_2^2(1-r^2)} \end{pmatrix} \tag{33}$$

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

dann

$$\delta^2 = \vec{\xi}'\Sigma^{-1}\vec{\xi} = \frac{\xi_1^2\sigma_2^2 + \xi_2^2\sigma_1^2 - 2\xi_1\xi_2\sigma_{12}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, \tag{34}$$

wobei wieder $\vec{\xi} = (\xi_1, \xi_2)' = ((x_1 - \mu_1), (x_2 - \mu_2))'$ gesetzt wurde; (34) definiert, wie oben schon angedeutet, für $\delta =$ konstant eine Ellipse, die für $\sigma_{12} = 0$ achsenparallel wird. Für gegebenen Wert von σ_{12} liegt damit jeder Punkt auf einer vom Ellipsen haben die gleiche, durch σ_{12} defnierte Orientierung.

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

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

$$\delta^2 = \frac{\xi_1^2\sigma_2^2 + \xi_2^2\sigma_1^2}{\sigma_1^2\sigma_2^2} = \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} = z_1^2 + z_2^2. \tag{35}$$

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

(b) $\sigma_{12} \neq 0$. In diesem Fall hängt der Wert von δ einerseits von $\xi_1\sigma_2$ und $\xi_2\sigma_1$ ab, andererseits vom Produkt $\xi_1\xi_2\sigma_{12}$ ab. Wichtig dabei ist die Lage der Punkte (x_1, x_2) relativ zum Punkt (μ_1, μ_2). Um zu einer Veranschaulichung der Verhältnisse zu gelangen, faßt man $\xi_1 = x_1 - \mu_1$ und $\xi_2 = x_2 - \mu_2$ als Konstante auf, hält ebenfalls σ_1 und σ_2 konstant und variiert σ_{12} in (34), etwa im Intervall[5]

$$-\sigma_1\sigma_2 < \sigma_{12} < \sigma_1\sigma_2. \tag{36}$$

Für $\sigma_{12} \to \sigma_1\sigma_2$ folgt $\delta \to \infty$, da dann der Nenner in (34) gegen Null strebt. Um das Verhalten von δ in Abhängigkeit von σ_{12} zu verdeutlichen, ist es illustrativer, ein kleineres Intervall zu betrachten, in dem $|\sigma_{12}| < \sigma_1\sigma_2$ gilt, etwa

$$-\min(\sigma_1, \sigma_2) \leq \sigma_{12} \leq \min(\sigma_1, \sigma_2). \tag{37}$$

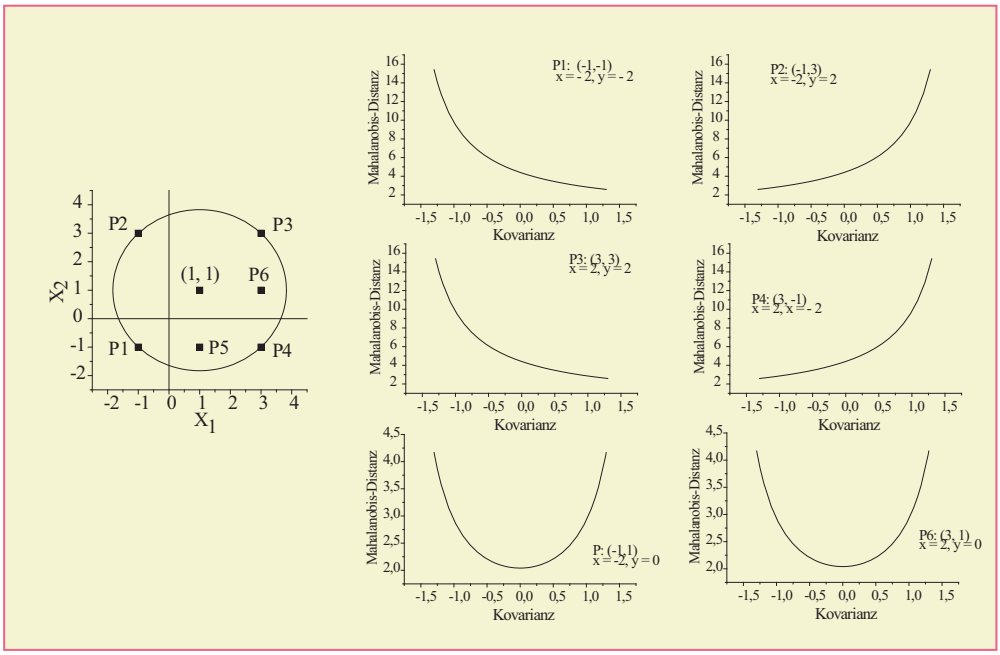
Die Inspektion von (34) zeigt, daß der Wert von δ als Funktion von σ_{12} vom Vorzeichen der Differenzen ξ_1 und ξ_2 abhängt; sind die Vorzeichen gleich, wird sich ein anderer Verlauf ergeben als wenn sie ungleich sind. In Abbildung 2 sind verschiedene Verläufe der Mahalanobis-Distanz in Abhängigkeit von der Kovarianz σ_{12} dargestellt worden. Die Form des Verlaufs und der Wertebereich von δ hängen von den Positionen des jeweiligen Punktepaares ab. Hier war einer der beiden Punkte, der Punkt mit den Koordinaten (1, 1), stets der Mittelpunkt eines Kreises und die Punkte P_1, P_2, P_3 und P_4 liegen auf dem Umfang des Kreises; sie haben deshalb alle den gleichen (euklidischen) Abstand von (1, 1). Kovariieren die x-Komponenten von x positiv, so müßten alle Punkte in der Nachbarschaft der Geraden durch P_1 und P_3 liegen,

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

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

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

Abbildung 2: Mahalanobis-Distanzen zwischen $(1, 1)$ und verschiedenen Punkten für verschiedene Kovarianzen



kovariieren sie negativ, so liegen sie in der Nachbarschaft der Geraden durch P_2 und P_4 liegen. Für die Punkte P_1 und P_3 ergibt sich demnach der gleiche Zusammenhang zwischen δ und σ_{12}: Für negative Werte von σ_{12} ist δ groß, d.h. $\sigma_{12} \to -\sigma_1\sigma_2$ folgt $\delta \to \infty$, und für $\sigma_{12} \to \sigma_1\sigma_2$ folgt $\delta \to 0$, d.h. je größer die Kovarianz, desto kleiner wird δ. Für die Punkte P_2 und P_4 ergibt sich der umgekehrte Zusammenhang: je größer der Wert der Kovarianz, desto größer wird auch die Mahalanobis-Distanz δ; für $\sigma_{12} \to \sigma_1\sigma_2$ folgt $\delta \to \infty$.

Die Punkte P_5 und P_6 liegen nicht auf den Geraden, die einer perfekten Kovarianz entsprechen, es ergeben sich die in Abb. 2 gezeigten U-förmigen Zusammenhänge. Die Mahalanobis-Distanzen sind hier minimal, wenn die Kovarianz gleich Null ist.

1.3.2 Ungleiche Varianz-Kovarianzmatrizen

Für die multivariate Normalverteilung sind die Diskriminanzfunktionen für die Maximum-a posteriori-Regel gemäß (19) durch

$$d_k(\vec{x}) = \frac{1}{2}(\vec{x} - \vec{\mu}_k)'\Sigma_k^{-1}(\vec{x} - \vec{\mu}_k) - \log|\Sigma_k^{-1}| - \log p(\Omega_k) \tag{38}$$

gegeben. Man sieht, dass d_k u.a. durch die in (27) eingeführte Mahalanobis-Distanz

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

definiert ist, denn offenbar gilt

$$d_k(\vec{x}) = \frac{1}{2}\delta(\vec{x}, \vec{\mu}_k) - \log|\Sigma_k^{-1}| - \log p(\Omega_k). \tag{39}$$

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

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

so dass

$$d_k(\vec{x}) = \frac{1}{2}\vec{x}\,'\Sigma_k^{-1}\vec{x} - \vec{\mu}_k'\Sigma_k^{-1}\vec{x} + \frac{1}{2}\vec{\mu}_k\,'\Sigma_k^{-1}\vec{\mu}_k - \frac{1}{2}(\log|\Sigma_k^{-1}| - \log p(\Omega_k)), \tag{40}$$

oder

$$d_k(\vec{x}) = \vec{x}\,'A_k\vec{x} - B_k\vec{x} + C_{k0}, \tag{41}$$

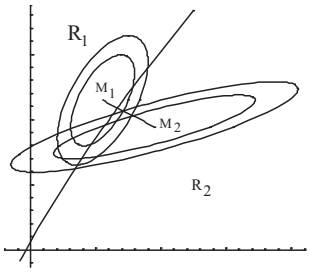
mit $A_k = (1/2)\Sigma_k^{-1}$, $B_k = \vec{\mu}_k\,'\Sigma_k^{-1}$ und $C_{k0} = -\frac{1}{2}(\log|\Sigma_k^{-1}| - \log p(\Omega_k))$. A_k ist eine symmetrische Matrix, weshalb $\vec{x}\,'A_k\vec{x}$ eine *quadratische Form*[6] ist. Die Diskriminanzfunktion hängt also von den Quadraten der Komponenten von $\vec{x}$ ab.

Die Trennflächen sind Lösungen der Gleichungen $d_j(\vec{x}) - d_k(\vec{x}) = 0$. (41) liefert

$$\begin{aligned} 0 = d_j(\vec{x}) - d_k(\vec{x}) &= \vec{x}\,'A_j\vec{x} - B_j\vec{x} + C_{j0} - \vec{x}\,'A_k\vec{x} + B_k\vec{x} - C_{k0} \\ &= \vec{x}\,'(A_j - A_k)\vec{x} - (B_j - B_k)\vec{x} + C_{j0} - C_{k0} \end{aligned} \tag{42}$$

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

Abbildung 3: Gaussverteilungen mit ungleichen Varianz-Kovarianz-Matrizen und nichtlinearer Trennung der Bereiche; M_1 und M_2 Mittelpunkte der Ellipsen.



1.3.3 Gleiche Varianz-Kovarianzmatrizen

Es gelte nun $\Sigma_k = \Sigma$ für alle k. Der Spezialfall gleicher Kovarianzmatrizen ist für die Praxis von großer Bedeutung, da einerseits nicht für jedes Ω_k eine besondere Varianz-Kovarianzmatrix geschätzt werden muß, und andererseits sich einfachere Diskriminanzfunktionen ergeben. Das Quadrat der Mahlanobis-Distanz ist nun

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

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

d.h. es ist $\Sigma_k = \Sigma$ für alle k, so dass nach (39)

$$d_k(\vec{x}) = \frac{1}{2}\delta(\vec{x}, \vec{\mu}_k) - \log|\Sigma^{-1}| - \log p(\Omega_k). \tag{43}$$

Aus (40) erhält man für $\Sigma_k = \Sigma$ sofort

$$d_k(\vec{x}) = \frac{1}{2}x'\Sigma^{-1}\vec{x} - \mu_k'\Sigma^{-1}\vec{x} + \frac{1}{2}\vec{\mu}_k\,'\Sigma^{-1}\vec{\mu}_k - \frac{1}{2}(\log|\Sigma^{-1}| - \log p(\Omega_k)). \tag{44}$$

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

$$d_k(\vec{x}) = -\vec{\mu}_k\,'\Sigma^{-1}\vec{x} + \frac{1}{2}\vec{\mu}_k\,'\Sigma^{-1}\vec{\mu}_k - \log p(\Omega_k)). \tag{45}$$

Flächen gleicher Distanz: Diese Flächen sind nach (21) durch die Gleichungen $d_j(\vec{x}) = d_k(\vec{x})$ definiert, d.h. es soll $d_j(\vec{x}) - d_k(\vec{x}) = 0$ gelten. Man findet

$$d_j(\vec{x}) - d_k(\vec{x}) = \vec{\mu}_k\,'\Sigma^{-1}\vec{x} - \vec{\mu}_j\,'\Sigma^{-1}\vec{x} + \frac{1}{2}\vec{\mu}_j\,'\Sigma^{-1}\vec{\mu}_j - \frac{1}{2}\vec{\mu}_k\,'\Sigma^{-1}\vec{\mu}_k - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right). \tag{46}$$

Diese Gleichung läßt sich zu

$$d_j(\vec{x}) - d_k(\vec{x}) = (\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}\vec{x} - \frac{1}{2}(\vec{\mu}_k - \vec{\mu}_j)\,'\Sigma^{-1}(\vec{\mu}_k + \vec{\mu}_j) - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right) \tag{47}$$

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

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

d.h.

$$(\vec{\mu}_k - \vec{\mu}_j)\,'\Sigma^{-1}\vec{x} = \frac{1}{2}(\vec{\mu}_j - \vec{\mu}_k)'\Sigma^{-1}(\vec{\mu}_j + \vec{\mu}_k) - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right) \tag{48}$$

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

$$(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1} = \vec{b}_{kj}' = (b_{kj}^{(1)}, \ldots, b_{kj}^{(p)}). \tag{49}$$

Dann ist $(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}\vec{x}$ ein Skalarprodukt, und (48) kann in der Form

$$(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}\vec{x} = b_{kj}^{(1)}x_1 + b_2x_2 + \cdots + b_{kj}^{(p)}x_p = K_{jk} = \text{konstant} \tag{50}$$

geschrieben werden, wobei die Konstante K_{jk} durch die rechte Seite von (48), also

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

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

$$b_{kj}^{(1)}x_1 + b_{kj}^{(2)}x_2 + b_{kj}^{(3)}x_3 = K_{jk}, \tag{51}$$

d.h. etwa

$$x_3 = K_{jk}/b_{kj}^{(3)} - (b_{kj}^{(1)}/b_{kj}^{(3)})x_1 - (b_{kj}^{(2)}/b_{kj}^{(3)})x_2, \tag{52}$$

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

$$b_{kj}^{(1)} x_1 + b_{kj}^{(2)} x_2 = K_{jk}. \tag{53}$$

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

$$x_2 = K_{jk}/b_{kj}^{(2)} - (b_{kj}^{(2)}/b_{kj}^{(1)})x_1, \tag{54}$$

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

Position und Orientierung der Hyperebenen: Erweitert man in (48) den Term $\log P(\Omega_k)/P(\Omega_j)$ in (47) mit der Mahalanobis-Distanz $\delta(\vec{\mu}_j, \vec{\mu}_k) = (\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}(\vec{\mu}_k - \vec{\mu}_j)$ zwischen $\vec{\mu}_j$ und $\vec{\mu}_k$,

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

und setzt man diesen Ausdruck in (46) für $\log P(\Omega_k)/P(\Omega_j)$ ein und zieht dann den Faktor $(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}$ heraus, erhält man

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

Zur Vereinfachung kann man

$$\vec{x}_0 = \frac{1}{2}(\vec{\mu}_j + \vec{\mu}_k) - \frac{\log(p(\Omega_k)/p(\Omega_j))}{\delta(\vec{\mu}_j, \vec{\mu}_k)}(\vec{\mu}_j - \vec{\mu}_k)) \tag{55}$$

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

$$(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1}(\vec{x} - \vec{x}_0) = 0 \tag{56}$$

schreiben kann.

Der Vektor $(\vec{\mu}_k - \vec{\mu}_j)$ entspricht der Geraden, die die Punkte $\vec{\mu}_j$ und $\vec{\mu}_k$ verbindet. (56) bedeutet dann, dass die Vektoren $\vec{x}$, die in der trennenden Hyperebenen liegen, nicht orthogonal zu $(\vec{\mu}_k - \vec{\mu}_j)$ sind. Orthogonal sind nur die Vektoren $(\vec{\mu}_k - \vec{\mu}_j)'\Sigma^{-1} = \vec{b}$ und $\vec{x} - \vec{x}_0$. Die Hyperebene $b_1 x_1 + \cdots + b_p x_p$ schneidet die Gerade $(\vec{\mu}_k - \vec{\mu}_j)$ auch nicht notwendig in deren Mittelpunkt.

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

$$\hat{\vec{\mu}_k} = \vec{\bar{x}_k}, \quad \hat{\Sigma}_k = S_k, \tag{57}$$

wobei S_k die empirische Matrix der Varianzen und Kovarianzen ist.

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

$$S = \frac{1}{N-g} \sum_{s=1}^{g} \sum_{n=1}^{n_k} (x_{kn} - \bar{x}_k)(x_{kn} - \bar{x}_k)', \tag{58}$$

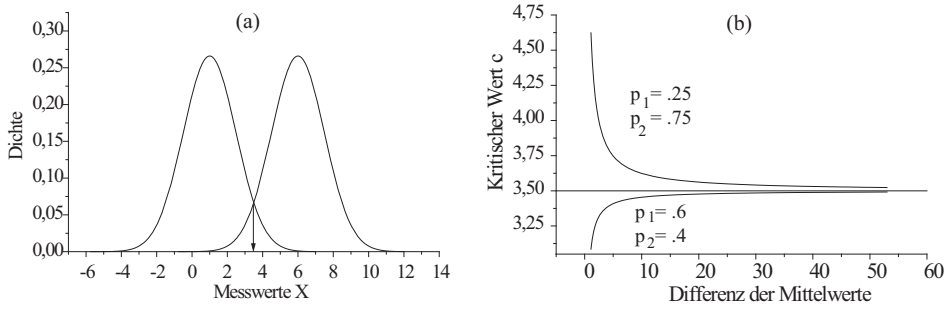
S_k bzw. S und $\bar{x}_k$ werden dann für die Größen Σ_k, Σ und μ_k eingesetzt ("plug-in"-Schätzungen).

1.3.4 Klassifikationen und Fehlklassifikationen

1.3.5 Beispiele

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

Abbildung 4: (a) Gaussverteilungen mit gleicher Varianz; der Pfeil zeigt auf $c_0 = (\mu_1 + \mu_2)/2$. (b) Kritische Werte für verschiedene Werte von p_1/p_2 in Abhängigkeit von der Differenz $\mu_1 - \mu_2$ bei konstantem $c_0 =$.



Der Variablenraum ist hier 1-dimensional, d.h. ist eine Gerade. Die Hyperebene, die die Bereiche R_1 und R_2 trennt, ist jetzt ein Punkt auf dieser Gerade, d.h. eine reelle Zahl c. Dieser Punkt entspricht der Lösung $x = c$ der Gleichung (56), in der x_0 ebenfalls ein Skalar und kein Vektor ist. Dann ist aber (56) genau dann erfüllt, wenn $c = x_0$ ist, und x_0 ist durch (55) gegeben. Man findet $\delta = (\bar{x}_1 - \bar{x}_2)/\sigma^2$ und damit

$$c = \frac{1}{2}(\bar{x}_1 + \bar{x}_2) - \frac{\sigma^2}{\bar{x}_1 - \bar{x}_2} \log\left(\frac{p(\Omega_2)}{p(\Omega_1)}\right), \tag{59}$$

und c trennt die Bereiche R_1 und R_2; beobachtet man einen $x(\omega)$-Wert größer als c so wird man ω der Klasse Ω_2 zuordnen, andernfalls Ω_2. Für $p(\Omega_1) = p(\Omega_2)$ wird $\log(p(\Omega_k)/p(\Omega_j)) = 0$ und c halbiert gerade das Intervall zwischen $\bar{x}_1$ und $\bar{x}_2$. Der Effekt unterschiedlicher a priori-Wahrscheinlichkeiten hängt vom Wert der Differenz $\mu_1 - \mu_2$ ab; für größer werdende Differenz strebt c gegen $c_0 = (\mu_1 + \mu_2)/2$ (vergl. Abb. 4). □

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

$$S = \begin{pmatrix} 32.948, & 10.743 \\ 10.743, & 10.24 \end{pmatrix}, \quad S^{-1} = 365 \begin{pmatrix} 0.046, & -.048 \\ -.048, & .148 \end{pmatrix}. \tag{60}$$

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

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

	Brahmanen	Handwerker	Varianz
Größe	164.51	160.53	32.948
Sitzhöhe	86.43	81.47	10.240

für die Handwerkers, so hat man

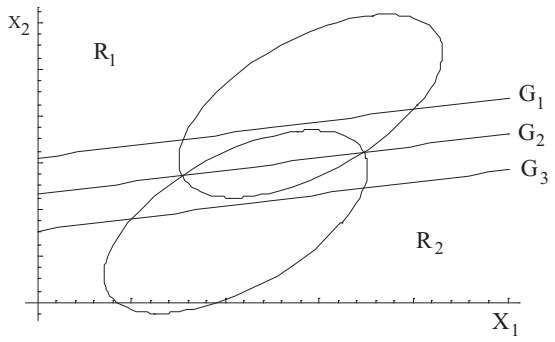
$$\bar{x}_B = \begin{pmatrix} 164.51 \\ 86.43 \end{pmatrix}, \quad \bar{x}_A = \begin{pmatrix} 160.53 \\ 81.47 \end{pmatrix}. \tag{61}$$

Nach (48) ist die Hyperebene, die die Bereiche trennt, durch

$$(\bar{x}_B - \bar{x}_A)' S^{-1} x = \frac{1}{2}(\bar{x}_B - \bar{x}_A)' S^{-1}(\bar{x}_B + \bar{X}_A) - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right). \tag{62}$$

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

Abbildung 5: Brahmanen und Handwerker (Beispiel 1.2): Trenn"flächen" für verschiedene a priori-Wahrscheinlichkeiten; G_1: $p_1/p_2 = .3/.7$; G_2: $p_1/p_2 = 1$, G_3: $p_1/p_2 = .7/.3$.



$$.056x_1 - .544x_2 = -36.460 - \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right) \tag{63}$$

gegeben, bzw. durch

$$x_2 = \left(36.460 + \log\left(\frac{p(\Omega_k)}{p(\Omega_j)}\right)\right)/.544 - (.056/.544)x_1. \tag{64}$$

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

Die ursprünglichen Messwerte sind nicht gegeben, aber man kann die Ellipsen, die den jeweiligen 2-dimensionalen Normalverteilungen der beiden Gruppen entsprechen, bestimmen, denn sie sind durch die Mahalanobis-Distanz $\delta(x, \bar{x}_i)$, $i = 1, 2$ gegeben.

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

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

Tabelle 2: Scores für verschiedene Berufe

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

die durchschnittliche Varianz-Kovarianz-Matrix S und deren Inverse S^{-1} hat man

$$S = \begin{pmatrix} 100 & 30 & 32 \\ 30 & 100 & 44 \\ 32 & 44 & 100 \end{pmatrix}, \quad S^{-1} = \begin{pmatrix} .0115 & \text{-}.0023 & \text{-}.0027 \\ \text{-}.0023 & .0129 & \text{-}.0049 \\ \text{-}.0027 & \text{-}.0049 & .0130 \end{pmatrix} \tag{65}$$

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 (45) (Seite 15) 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

$$d_1 = (108 - 114, 112 - 111, 101 - 110)S^{-1} \begin{pmatrix} 108 \text{ - } 114 \\ 112 \text{ - } 111 \\ 101 \text{ - } 110 \end{pmatrix} \tag{66}$$

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. □

1.4 Der Ansatz von Fisher (1936)

1.4.1 Diskriminanzfunktionen

Bei diesem Ansatz[7] wird die Annahme der p-dimensionalen Normalverteilung (zunächst) nicht gemacht.

Es werden wieder p Merkmale beobachtet, und die 5 Beobachtungen mögen als Messungen $X_1, \cdots, X_p$ vorliegen. Der allgemeine Ansatz ist nun, Gewichte $u_1, \cdots, u_p$ zu finden derart, daß die Klassifikation anhand der linearen Funktion

$$Y = u_1 X_1 + \cdots + u_p X_p \tag{67}$$

durchgeführt werden kann. Y ist ein Wert auf einer Skala, die offenbar als Linearkombination der X_j, $j = 1, \cdots, p$ definiert ist und die die Eigenschaft haben soll, daß die Y-Werte von Mitgliedern verschiedener Gruppen maximal verschieden sind. Anders ausgedrückt soll die Variation der Y-Werte *innerhalb* der Gruppen so klein wie möglich im Vergleich zur Variation der Y *zwischen* den Gruppen sein, oder umgekehrt: die Variation zwischen den Gruppen soll so groß wie möglich im Vergleich zur Variation innerhalb der Gruppen sein. Hat man also für eine Person einen Y-Wert bestimmt, so soll man einigermaßen sicher sein 541 können, daß dieser Wert in einem Bereich liegt, der für die Gruppe oder Klasse, zu der die Person gehört, charakteristisch ist.

Die Variation der Y-Werte kann durch eine Varianz spezifiziert werden. Die Variation innerhalb der Gruppen ist dementsprechend durch die Varianz innerhalb der Gruppen definierbar, d.h. durch die Varianzen der Werte innerhalb der Gruppen in bezug auf den jeweiligen Gruppen- oder Klassenmittelwert $\bar{y}_k$, $k = 1, \cdots, K$, K die Anzahl der Gruppen. Die Varianz zwischen den Gruppen ist dann durch die Varianz der Mittelwerte repräsentierbar. Aber Varianzen sind durch entsprechende Quadratsummen definiert, etwa durch die Quadratsumme "zwischen", QS_{zw}, und durch die Quadratsumme "innerhalb", QS_{inn}. 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

$$\lambda = \lambda(u_1, \cdots, u_p) = \frac{QS_{zw}(u_1, \cdots, u_p)}{QS_{inn}(u_1, \cdots, u_p)} \tag{68}$$

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 1.4 *Die Funktion (67) heißt* lineare Diskriminanzfunktion *oder* kanonische Variable. *Die in (68) 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 den später (vergl. Abschnitt 2) noch eingegangen wird.

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

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.

1.4.2 Die Varianzzerlegung

Es sollen zunächst explizite Ausdrücke für die Quadratsummen QS_{inn} und QS_{zw} hergeleitet werden, um die Maximierung des 612 teriums λ durchführen zu können. Dazu ist es nützlich, die Gleichung (67) 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} 622 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:

$$Y = \begin{pmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{n_1 1} \\ \hline y_{12} \\ y_{22} \\ \vdots \\ y_{n_2 2} \\ \hline \vdots \\ \hline y_{1K} \\ y_{2K} \\ \vdots \\ y_{n_K K} \end{pmatrix}, \quad X = \begin{pmatrix} X_{111} & X_{112} & \cdots & X_{11p} \\ X_{211} & X_{212} & \cdots & X_{21p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_1 11} & X_{n_1 12} & \cdots & X_{n_1 1p} \\ \hline X_{121} & X_{122} & \cdots & X_{12p} \\ X_{221} & X_{222} & \cdots & X_{22p} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_2 21} & X_{n_2 22} & \cdots & X_{n_2 2p} \\ \hline \vdots & \vdots & \cdots & \vdots \\ \hline X_{1K1} & X_{1K2} & \cdots & X_{1Kp} \\ X_{2K1} & X_{2K2} & \cdots & X_{2Kp} \\ \vdots & \vdots & \cdots & \vdots \\ X_{n_K K1} & X_{n_K K2} & \cdots & X_{n_K Kp} \end{pmatrix}, \tag{69}$$

Für die Y-Werte gelte

$$\begin{aligned} y_{ik} &= u_1 x_{ik1} + u_2 x_{ik2} + \cdots + u_p x_{ikp}, \;\; i = 1, \ldots, n_k & (70) \\ \bar{y}_k &= u_1 \bar{x}_{k1} + u_2 \bar{x}_{k2} + \cdots + u_p \bar{x}_{kp} & (71) \\ \bar{y} &= u_1 \bar{x}_1 + u_2 \bar{x}_2 + \cdots + u_p \bar{x}_p & (72) \end{aligned}$$

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*

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

$$QS_{ges} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik}-\bar{y})^2, \quad QS_{inn} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_{ik}-\bar{y}_k)^2, \quad QS_{zw} = \sum_{k=1}^{K} n_k(\bar{y}_k-\bar{y})^2. \tag{73}$$

Dann gilt

$$QS_{ges} = QS_{zw} + QS_{inn} \tag{74}$$

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

□

Um die Abhängigkeit von den u_j, $1 \leq j \leq p$ explizit zu machen, muß (68) umgeschrieben werden. Durch Einsetzen ergibt sich

$$QS_{inn} = \sum_{k=1}^{K}\sum_{i=1}^{n_k}(u_1(X_{k1i} - \bar{x}_{k1}) + \cdots + u_p(X_{kpi} - \bar{x}_{kp}))^2 \tag{75}$$

$$QS_{zw} = \sum_{k=1}^{K} n_k(u_1(\bar{x}_{k1} - \bar{x}_1) + \cdots + u_p(\bar{x}_{kp} - \bar{x}_p))^2 \tag{76}$$

Man kann nun die rechten Seiten von (75) und (76) in den Ausdruck (68) 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 (68) und damit die Ausdrücke für QS_{zw} und QS_{inn} in Matrixform angeschrieben werden.

1.4.3 Die Matrixschreibweise für die Quadratsummen

Es sind Meßwerte auf p Prädiktorvariablen $X_1, \cdots, X_p$ gegeben. Gemäß (75) und (76) müssen die Mittelwerte $\bar{x}_{kj}$ und $\bar{x}_j$ gebildet werden, $k = 1, \cdots, K$ und $j = 1, \cdots, p$. Die $\bar{x}_{kj}$ sind die Mittelwerte der $j = 1, \cdots, p$ Prädiktorvariablen in der k-ten Gruppe; sie werden in einer $(K \times p)$-Matrix M zusammengefaßt:

$$M = \begin{pmatrix} \bar{x}_{11} & \bar{x}_{12} & \cdots & \bar{x}_{1p} \\ \bar{x}_{21} & \bar{x}_{22} & \cdots & \bar{x}_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \bar{x}_{K1} & \bar{x}_{K2} & \cdots & \bar{x}_{Kp} \end{pmatrix} \tag{77}$$

Es sei

$$\bar{x}_j = \frac{1}{K}\sum_{k=1}^{K}\bar{x}_{kj}; \tag{78}$$

$\bar{x}_j$ ist also der Mittelwert für die j-te Variable über alle Gruppen. Dann sei

$$\vec{x}_{ki} = (X_{k1i}, X_{k2i}, \ldots, X_{kpi})' \tag{79}$$

ein aus der i-ten Zeile in der k-ten Gruppe gebildete Vektor, und

$$\vec{\bar{x}} = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_p)' \tag{80}$$

sei der Vektor dieser Mittelwerte. Schließlich sei

$$\vec{\bar{x}}_{k\cdot} = (\bar{x}_{k1}, \bar{x}_{k2}, \cdots, \bar{x}_{kp})' \tag{81}$$

der Vektor der Gruppenmittelwerte für die j-te Prädiktorvariable; die Komponenten dieses Vektors sind die Elemente der k-ten Zeile der Matrix M.

Darstellung von Quadratsummen: Die Quadratsummen können leicht vektoriell dargestellt werden. Dabei wird von der folgenden Schreibweise Gebrauch gemacht: es seien $\vec{a} = (a_1, \ldots, a_m)'$ und $\vec{b} = (b_1, \ldots, b_n)'$ irgend zwei Vektoren. Dann kann das Produkt $\vec{a}\vec{b}'$ gebildet werden, das *kein* Skalarprodukt ist:

$$\vec{a}\vec{b}' = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} (b_1, b_2, \ldots, b_n) = \begin{pmatrix} a_1b_1 & a_1b_2 & \cdots & 793a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ & & \vdots & \\ a_mb_1 & a_mb_2 & \cdots & a_mb_n \end{pmatrix} \tag{82}$$

Gilt speziell $\vec{a} = \vec{b}$, so erhält man hieraus

$$\vec{a}\vec{a}' = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix} (a_1, a_2, \ldots, a_m) = \begin{pmatrix} a_1^2 & a_1a_2 & \cdots & a_1a_m \\ a_2a_1 & a_2^2 & \cdots & a_2a_m \\ & & \vdots & \\ a_ma_1 & a_ma_2 & \cdots & a_m^2 \end{pmatrix}. \tag{83}$$

Es sei nun $\vec{u} = (u_1, \ldots, u_p)'$ und $\vec{z} = (z_1, \ldots, z_p)'$ ein beliebiger Vektor. Dann gilt

$$(u_1z_1 + u_2z_2 + \cdots + u_pz_p)^2 = \vec{u}'\vec{z}\vec{z}'\vec{u}. \tag{84}$$

Für die linke Seite gilt

$$(u_1z_1 + u_2z_2 + \cdots + u_pz_p)^2 = u_1^2z_1^2 + \cdots + u_p^2z_p^2 + \sum_{i \neq j} u_iu_jz_iz_j. \tag{85}$$

Andererseits ist

$$\vec{z}\vec{z}' = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{pmatrix} (z_1, z_2, \ldots, z_p) = \begin{pmatrix} z_1^2 & z_1z_2 & \cdots & z_1z_p \\ z_2z_1 & z_2^2 & \cdots & z_2z_p \\ \vdots & \vdots & \ddots & \vdots \\ z_pz_1 & z_pz_2 & \cdots & z_p^2 \end{pmatrix},$$

und man rechnet leicht nach, dass dann

$$\vec{u}' \begin{pmatrix} z_1^2 & z_1z_2 & \cdots & z_1z_p \\ z_2z_1 & z_2^2 & \cdots & z_2z_p \\ \vdots & \vdots & \ddots & \vdots \\ z_pz_1 & z_pz_2 & \cdots & z_p^2 \end{pmatrix} \vec{u} = u_1^2z_1^2 + \cdots + u_p^2z_p^2 + \sum_{i \neq j} u_iu_jz_iz_j$$

ist, dass also die Gleichung (84) gilt.

Es sei nun $x_{kji} = X_{kji} - \bar{x}_{kj}$, und $\vec{x}_{ki} = (x_{k1i}, \ldots, x_{kpi})'$. Dann ist (vergl. (75))

$$(u_1x_{k1i} + u_2x_{k2i} + \cdots + x_{kpi})^2 = \vec{u}'\vec{x}_{ki}\vec{x}_{ki}{}'\vec{u}. \tag{86}$$

und

$$\begin{aligned} QS_{inn} &= \sum_{k=1}^{K}\sum_{i=1}^{n_k} \vec{u}\,'\vec{x}_{ki}\vec{x}_{ki}\,'\vec{u} \\ &= \vec{u}\,'\left(\sum_{k=1}^{K}\sum_{i=1}^{n_k} \vec{x}_{ki}\vec{x}_{ki}\,'\right)\vec{u} \end{aligned} \tag{87}$$

Es werde

$$W = \sum_{k=1}^{K}\sum_{i=1}^{n_k} \vec{x}_{ki}\vec{x}_{ki}\,' \tag{88}$$

gesetzt; W ist die Matrix der zusammengefassten ("pooled") Varianz-Kovarianz-Quadratsummen, und statt (87) kann man

$$QS_{inn} = \vec{u}\,'W\vec{u} \tag{89}$$

schreiben. Für QS_{zw} erhält man analog

$$QS_{zw} = \sum_{k=1}^{K} n_k\vec{u}\,'(\bar{x}_{k\cdot} - \bar{x})(\vec{\bar{x}}_{k\cdot} - \vec{\bar{x}})'\vec{u} = \vec{u}\,'\left(\sum_{k=1}^{K} n_k(\vec{\bar{x}}_{k\cdot} - \vec{\bar{x}})(\vec{\bar{x}}_{k\cdot} - \vec{\bar{x}})'\right)\vec{u}. \tag{90}$$

Setzt man

$$B = \sum_{k=1}^{K} n_k(\vec{\bar{x}}_{k\cdot} - \vec{\bar{x}})(\vec{\bar{x}}_{k\cdot} - \vec{\bar{x}})' \tag{91}$$

so kann man

$$QS_{zw} = \vec{u}\,'B\vec{u} \tag{92}$$

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

Die Matrix M läßt sich in Matrixschreibweise definieren. Dazu werde die Matrix G eingeführt:

$$G = \left(\begin{array}{ccccc} 1 & 0 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 \\ \hline 0 & 1 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 1 & 0 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & \cdots & \vdots \\ \hline 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{array}\right) \tag{93}$$

G ist eine *Indikatormatrix*: die Spalten repräsentieren die verschiedenen Gruppen, und eine 1 zeigt an, in welche Gruppe eine Person gehört. Personen, die zu einer Gruppe

gehören, sind zusammengefasst worden: die horizontalen Geraden trennen die Meßwerte der verschiedenen Gruppen. Es ist

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

$n_1, \ldots, n_K$ die Anzahl der Messwerte in den verschiedenen Gruppen. Die in (77) eingeführte Matrix M der Mittelwerte läßt sich dann wie folgt darstellen: die Matrix $G'X$ ist offenbar eine $(K \times p)$-Matrix, deren k-te Zeile und j-te Spalte gerade die Summe $\sum_{i=1}^{n_k} x_{ijk}$ enthält. Also ist

$$G'X = (G'G)M, \quad X'G = M'(G'G),$$

und damit

$$M = (G'G)^{-1}G'X. \tag{94}$$

denn als Diagonalmatrix ist $G'G$ natürlich symmetrisch, d.h. $(G'G)' = G'G$.

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

$$W = (X - GM)'(X - GM), \quad B = (GM - 1\bar{x}')'(GM - 1\bar{x}') \tag{95}$$

entsprechen. Mit

$$T = (X - \vec{1}\vec{\bar{x}}')'(X - \vec{1}\vec{\bar{x}}') \tag{96}$$

gilt[8]

$$T = W + B, \quad QS_{ges} = \vec{u}\,'T\vec{u} = \vec{u}\,'(W + B)\vec{u} \tag{97}$$

Es sei $x = X - \vec{1}\vec{\bar{x}}\,'$; die Matrix x heißt *zentriert*. Für den Fall zentrierter Daten Daten vereinfachen sich diese Ausdrücke etwas. Da $\vec{\bar{x}} = \vec{0}$ ergeben sich die Ausdrücke für den zentrierten Fall, indem man $\vec{1}\vec{\bar{x}}\,'$ gleich Null setzt bzw. einfach fortläßt. Dementsprechend erhält man aus (95)

$$W = (x - GM)'(x - GM), \quad B = (GM)'(GM) = M'G'GM \tag{98}$$

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

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

Ausmultipliziert ergibt sich

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

Es ist aber

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

Aus $x'G = M'G'G$ (vergl. (94)) folgt durch Multiplikation mit G^{-1} 961von rechts $x' = M'G'$, so daß $x'GM - M'G'GM = M'G'GM - M'G'GM = 0$, und ebenso $x'1\bar{x}' - M'G'1\bar{x}' = M'G'1\bar{x}' - M'G'1\bar{x}' = 0$. Der letzte Term in der Zerlegung für T ist dann ebenfalls gleich Null, da $(x - 1\bar{9}68x')'(x - GM) = ((x - GM)'(x - 1\bar{x}))'$; so daß tatsächlich $T = W + B$.

wobei M natürlich anhand von x, nicht von X berechnet wird. (96) liefert

$$T = x'x \tag{99}$$

und (97) gilt natürlich nach wie vor. Zusammenfassend ergibt sich

$$\lambda = \frac{QS_{zw}}{QS_{inn}} = \frac{\vec{u}\,'B\vec{u}}{\vec{u}\,'W\vec{u}} \tag{100}$$

1.4.4 Bestimmung der Lösung für $\vec{u}$ und λ

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

$$\lambda = \frac{\mathbf{x}'A\mathbf{x}}{\mathbf{x}'\mathbf{x}} \tag{101}$$

ein *Raleigh-Quotient*. Es gilt der

Satz 1.2 *(Satz von Courant-Fischer) Es sei A eine symmetrische Matric mit Eigenwerten $\lambda_1 \geq \cdots \geq \lambda_n$ und Eigenvektoren $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n$. Der Raleigh-Quotient λ nimmt den maximalen Wert*

$$\max_{\boldsymbol{x}\neq\vec{0}} \frac{\boldsymbol{x}'A\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \lambda_1 = \frac{\boldsymbol{p}_1'A\boldsymbol{p}_1}{\boldsymbol{p}_1'\boldsymbol{p}_1} \tag{102}$$

und den minimalen Wert

$$\min_{\boldsymbol{x}\neq\vec{0}} \frac{\boldsymbol{x}'A\boldsymbol{x}}{\boldsymbol{x}'\boldsymbol{x}} = \lambda_n = \frac{\boldsymbol{p}_n'A\boldsymbol{p}_n}{\boldsymbol{p}_n'\boldsymbol{p}_n} \tag{103}$$

an.

Beweis: Da A symmetrisch existiert die Darstellung $A = P\Lambda P'$, P die orthonormalen Eigenvektoren $\mathbf{p}_1, \ldots, \mathbf{p}_n$ und Λ die Diagonalmatrix der Eigenwerte $\lambda_1, \ldots, \lambda_n$ von A. Dann gilt

$$\lambda = \frac{\mathbf{x}'P\Lambda P\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \frac{\mathbf{x}'P\Lambda P\mathbf{x}}{\mathbf{x}'PP'\mathbf{x}},$$

denn $PP' = I$ die Einheitsmatrix. Es werde $P'\mathbf{x}$ ist ein Vektor, und es werde $\mathbf{v} = P'\mathbf{x}$ gesetzt, so dass

$$\lambda = \frac{\mathbf{v}'\Lambda\mathbf{v}}{\mathbf{v}'\mathbf{v}} = \frac{\sum_j \lambda_j \mathbf{v}_j^2}{v_j^2}. \tag{104}$$

Hierin kann man die λ_j durch $\lambda_{\mathbf{X}} = \lambda_1$ ersetzen; man hat dann

$$\lambda = \frac{\sum_j \lambda_j v_j^2}{v_j^2} \leq \frac{\sum_j \lambda_1 v_j^2}{v_j^2} = \frac{\lambda_1 \sum_j v_j^2}{v_j^2} = \lambda_1.$$

Der Wert von λ ist also höchstens gleich dem größten Eigenwert λ_1 von A, – gleich, welchen Vektor $\mathbf{x}$ man wählt. Inbesondere kann man also den zu λ_1 korrespondierenden Eigenvektor $\mathbf{p}_1$ wählen. Man erhält dann

$$\lambda = \frac{\mathbf{p}_1'P\Lambda P'\mathbf{p}_1}{\mathbf{p}_1'\mathbf{p}_1} = \lambda_1,$$

denn die Eigenvektoren in P sind orthonormal, so dass $\mathbf{p}_1' P = (1, 0, \ldots, 0)'$, denn $\mathbf{p}_1'\mathbf{p}_j = 0$ für $j \neq 1$ und $\mathbf{p}_1'\mathbf{p}_j = 1$ für $j = 1$. λ nimmt also den maximal möglichen Wert λ_1 an genau dann, wenn $\mathbf{x} = \mathbf{p}_1$.

Ersetzt man in (104) die λ_j durch den kleinsten Eigenwert λ_n, so wird man auf analoge Weise auf (103) geführt. □

Es gilt der

Satz 1.3 *Der gesuchte Vektor $\vec{u}$ von Gewichten ist ein Eigenvektor der Matrix $W^{-1}B$, und das Diskriminanzkriterium λ ist der zugehörige Eigenwert dieser Matrix, d.h. es gilt*

$$W^{-1}B\boldsymbol{u} = \lambda\boldsymbol{u} \tag{105}$$

Beweis: Der Satz kan bewiesen werden, indem man die partiellen Ableitungen $\partial\lambda/\partial u_j$ bildet und gleich Null setzt; die Lösungen des entstehenden Gleichungssystems liefern die $\hat{u}_j$, für die λ maximal wird. Dieser Beweis findet sich im Anhang, Abschnitt 4.5. Hier wird ein Beweis vorgeführt, der auf dem Satz von Courant-Fischer beruht. Er hat den Vorteil, auf eine Aussage über die Beziehung zwischen Diskriminanzfunktionen zu führen.

Gesucht ist

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

Sowohl B wie auch W sind symmetrische Matrizen. Also existiert für W die Spektralzerlegung $W = P\Lambda P'$. Mit $\Lambda^{1/2}$ werde die Diagonalmatrix bezeichnet, in deren Diagonalen die Wurzeln $\sqrt{\lambda_j}$ der Eigenwerte λ_j von W stehen; offenbar gilt $\Lambda 11/2\Lambda^{1/2} = \Lambda$, so dass man $W = P\Lambda^{1/2}\Lambda^{1/2}P'$ schreiben kann. Man definiert $W^{1/2} = P\Lambda^{1/2}$ als die *Wurzel* der Matrix W; in der Tat findet man

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

Weiter sei

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

Dann ist

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

denn $\mathbf{u} = W^{-1/2}\mathbf{v}$. Damit ist λ ein Raleigh-Quotient in Bezug auf die symmetrische Matrix $A = W^{-1/2}BW^{-1}$, und λ nimmt den maximalen Wert λ_1 an, der sich als größter Eigenwert von A ergibt, wenn $\mathbf{v} = \mathbf{v}_1$ der zugehörige Eigenvektor von A ist. Es muß demnach

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

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

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

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

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

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

Fasst man alle Eigenvektoren $\mathbf{u}$ zu einer Matrix U und alle Eigenwerte zu einer Diagonalmatrix Λ zusammen, so kann (105) in der Form

$$W^{-1}BU = U\Lambda \tag{106}$$

schreiben.

Anmerkung: Satz 1.3 bezieht sich zunächst nur auf einen Gewichtsvektor $\vec{u}$. Andererseits ist es möglich, dass die Matrix $W^{-1}B$ mehr als einen von Null verschiedenen Eigenwert λ hat, so dass mehr als ein Vektor $\mathbf{u}$ existiert. Es ist also möglich, dass $r > 1$ Eigenwerte $\lambda_r \neq 0$ und damit r Eigenvektoren $\mathbf{u}_r$ existieren. Da die Matrix $W^{-1}B$ nicht symmetrisch ist, sind die $\mathbf{u}_r$ zwar linear unabhängig, aber nicht orthogonal.

Satz 1.4 *$W^{-1}B$ 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 $\boldsymbol{y}_j = u_{1j}\boldsymbol{x}_1 + \cdots + u_{pj}\boldsymbol{x}_p$, $j = 1, \ldots, r$ und es gilt*

$$\boldsymbol{y}_j'\boldsymbol{y}_k = 0, \quad j \neq k, \tag{107}$$

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

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

$$\mathbf{y}_j = X\mathbf{u}_j. \tag{108}$$

Es werde angenommen, dass die Spalten von X durch $\mathbf{x}_j = \mathbf{X}_j - \bar{\mathbf{x}}_j$ definiert sind, 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 $\Sigma = X'X$ die Matrix der Kovarianzen zwischen den $\mathbf{x}_j$ ist. Nun wird Σ aber durch die Matrix W geschätzt, $\mathbf{u}_j$ 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. □

1.4.5 Klassifikation von Beobachtungen

Für ein Object wird der Vektor $\mathbf{x} = (x_1, \ldots, x_p)'$ von Messungen bestimmt. Es stellt sich nun die Frage, welcher Klasse das Objekt zugeordnet werden soll.

Dazu wird zunächst der zugehörige Vektor $\mathbf{y}$ bestimmt:

$$\mathbf{y} = U'\mathbf{x} = \begin{pmatrix} u_{11}, & u_{21}, & \cdots & u_{ps} \\ u_{12}, & u_{22}, & \cdots & u_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1s}, & u_{2s}, & \cdots & u_{ps} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_s \end{pmatrix}, \tag{109}$$

y_j ist die Komponente von $\mathbf{y}$ auf der j-ten kanonischen Variablen, $j = 1, \ldots, s$. Weiter sei

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

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

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

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

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

gegeben. Dann ist

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

d.h.

$$\bar{y}_{jk} = \bar{x}_{1k}u_{1j} + \bar{x}_{2k}u_{2j} + \cdots + \bar{x}_{pj}u_{pj}, \quad j = 1, \ldots, s \tag{110}$$

Dann ist

$$\mathbf{y} - \bar{\mathbf{y}}_k = U'\mathbf{x} - U'\bar{\mathbf{x}}_k = U'(\mathbf{x} - \bar{\mathbf{x}}_k). \tag{111}$$

Man hat dann den

Satz 1.5 *Es gilt*

$$\|\boldsymbol{y} - \bar{\boldsymbol{y}}_k\|^2 = (\boldsymbol{x} - \bar{\boldsymbol{x}}_k)'W^{-1}(\boldsymbol{x} - \bar{\boldsymbol{x}}_k), \tag{112}$$

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

Beweis: Es sei $W^{-1/2}$ die Wurzel der Matrix W^{-1}, so dass $W^{-1/2}W^{-1/2} = W^{-1}$ und $W^{-1/2}W^{1/2} = I$ (s.Anhang, Abschnitt 4.1). Aus der Gleichung (106) folgt dann

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

P ist die Matrix der orthonormalen Eigenvektoren der symmetrischen Matrix $W^{-1/2}BW^{-1/2}$, so dass $P'P = PP' = I$ und $U = W^{-1/2}P$, so dass $UU' = W^{-1/2}PP'W^{-1/2} = W^{-1}$, und wegen (111) folgt

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

und dies war zu zeigen. □

Anmerkung: Der Beziehung (112) liegt *nicht* die Annahme der multivariaten Normalverteilung zugrunde, die ja ebenfalls durch die Mahalanobis-Distanz $(\mathbf{x}-\bar{\mathbf{x}}_k)'W^{-1}(\mathbf{x}-\bar{\mathbf{x}}_k)$ definiert ist; die Beziehung (112) folgt rein algebraisch. □

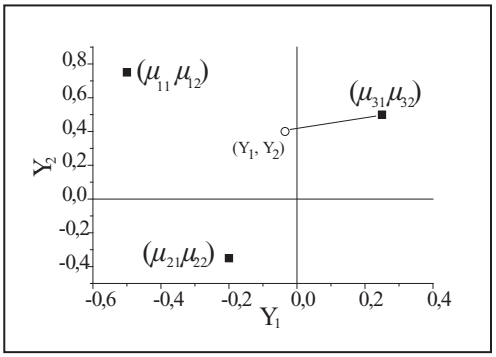
Das Objekt ω mit dem Vektor $\mathbf{x} = \mathbf{x}(\omega)$ und dem zugehörigen Vektor $\mathbf{y} = U'\mathbf{x}$ soll nun einer Klasse Ω_k zugeordnet werden. Es wird die folgende Entscheidungsregel eingeführt:

Regel:

$$\omega \to \Omega_k \text{ genau dann, wenn } \|\mathbf{y}-\bar{\mathbf{y}}_k\|^2 = \min_j \|\mathbf{y}-\bar{\mathbf{y}}_j\|^2 = \min_j (\mathbf{x}-\bar{\mathbf{x}}_j)'W^{-1}(\mathbf{x}-\bar{\mathbf{x}}_j), \quad (113)$$

d.h. man ordnet ω derjenigen Klasse G_k zu, für die $\|\mathbf{y} - \bar{\mathbf{y}}_k\|$ minimal ist, vergl. Abb. 6. Das ist gleichzeitig diejenige Klasse, für die die Mahalanobis-Distanz zwichen $\mathbf{x}$ und $\bar{\mathbf{x}}_k$ minimal ist. Man bemerke, dass die Beziehung (112) *nicht* durch Annahme der Normalverteilung hergeleitet wurde, sondern als rein algebraische Beziehung.

Abbildung 6: 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.



1.4.6 Zur Anzahl der kanonischen Variablen

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

$$\bar{\mathbf{y}} = \frac{1}{g}\sum_{k=1}^{g} \bar{\mathbf{y}}_k. \quad (114)$$

Es werden nun die g Vektoren

$$\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}, \bar{\mathbf{y}}_2 - \bar{\mathbf{y}}, \ldots, \bar{\mathbf{y}}_g - \bar{\mathbf{y}} \quad (115)$$

betrachtet. Dann folgt

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

Also gilt z.B.

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

d.h. $(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}})$ kann als Linearkombination der restlichen Differenzen dargestellt werden. Linearkombinationen der Vektoren (115) definieren Hyperebenen der Dimension $q \leq g-1$. Es sei $\mathbf{v}$ ein Vektor, der senkrecht auf jedem Vektor $\bar{\mathbf{y}}_k - \bar{\mathbf{y}}$ und damit auf den Hyperebenen steht. Dann hat man

$$B\mathbf{v} = \sum_{k=1}^{g} (\bar{\mathbf{y}}_k - \bar{\mathbf{y}})(\bar{\mathbf{y}}_k - \bar{\mathbf{y}})'\mathbf{v} = \sum_{k=1}^{g} (\bar{\mathbf{y}}_k - \bar{\mathbf{y}})\vec{0} = \vec{0}, \tag{116}$$

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

$$W^{-1}B\mathbf{v} = \mathbf{v}, \tag{117}$$

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 3: 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

1.5 Klassifikation nach Fisher versus Klassifikation nach Gauss

Der Fishersche Ansatz setzt nicht die Annahme der multivariaten Normalverteilung voraus. Gleichwohl liefert Satz 1.5 eine Beziehung zwischen der Klassifikation nach Fisher einerseits und Gauss andererseits.

In (112) wird die Fishersche Kriteriumsgröße zur Mahalanobis-Distanz in Beziehung gesetzt. Klassifiziert man gemäß der Gaussverteilung, so ist nach (19)

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

die Diskriminanzfunktion, woei $\vec{x}$ der Vektor $\vec{x} = (\vec{x}_1, \ldots, \vec{x}_n)'$ der Beobachtungen ist. Für f wird die multivariate Gauss-Verteilung eingesetzt. Nach (44) erhält man ann

$$d_k(\vec{x}) = \frac{1}{2}\vec{x}'\Sigma^{-1}\vec{x} - \vec{\mu}_k'\Sigma^{-1}\vec{x} + \frac{1}{2}\vec{\mu}_k'\Sigma^{-1}\vec{\mu}_k - \frac{1}{2}(\log|\Sigma^{-1}| - \log p(\Omega_k)), \tag{118}$$

bzw. nach (45)

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

da der Term $\vec{x}'W^{-1}\vec{x}$ für alle d_k identisch ist und daher für die Diskrimination keine Information liefert. Hier ist es sinnvoll, doch noch einmal den ursprünglichen Ausdruck

(118) für d_k zu btrachten: subtrahiert man $\vec{x}'W^{-1}\vec{x}/2$ und addiert man $\log p(\Omega_k)$ auf beiden Seiten, so erhält man

$$d_k(\vec{x}) - \frac{1}{2}\vec{x}'W^{-1}\vec{x} + \log p(\Omega_k) = -\frac{1}{2}(\vec{x} - \vec{\mu}_k)'\Sigma^{-1}(\vec{x} - \vec{\mu}_k). \tag{119}$$

Man erhält man dann die Beziehung

$$-d_k(\vec{x}) + \frac{1}{2}\vec{x}'W^{-1}\vec{x} + \log p(\Omega_k) = \sum_{j=1}^{s}(y_j - \vec{\mu}_j(Y))^2 = \|\vec{y} - \vec{\mu}_k(y)\|^2, \tag{120}$$

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

1.6 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

   $$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g, \tag{121}$$

   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

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

   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 $(\mathbf{x} - \bar{\mathbf{y}}_k)'W^{-1}(\mathbf{x} - \bar{\mathbf{y}}_k)$ 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 (112) 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 $\vec{x}$ 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

$$H_0: \qquad \mu_1 = \mu_2 = \cdots = \mu_K, \tag{122}$$

$$H_1: \qquad \mu_i \neq \mu_j, \text{ für mindestens ein Paar } (i,j) \text{ mit } i \neq j \tag{123}$$

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

$$\Lambda = \frac{|W|}{|B+W|} = |I + W^{-1}B|^{-1}, \tag{124}$$

Wilk's Λ; unter H_0 gilt

$$\Lambda \sim \Lambda(q, N-K, K-1) \tag{125}$$

(Λ-Verteilung von Wilks).

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

$$\epsilon = p(\Omega_1)\epsilon_{12} + p(\Omega_2)\epsilon_{21} \tag{126}$$

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:

$$\epsilon_{12} = \Phi\left(\frac{\log(\pi_1/\pi_2) - \delta^2/2}{\delta}\right) \tag{127}$$

$$\epsilon_{21} = \Phi\left(-\frac{\log(\pi_2/\pi_1) + \delta^2/2}{\delta}\right), \tag{128}$$

wobei

$$\delta = (\mu_1 - \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) \tag{129}$$

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

Für die ML-Regel ergeben sich die Fehlerraten gemäß

$$\epsilon_{12} = \epsilon_{21} = \Phi\left(-\frac{\delta}{2}\right). \tag{130}$$

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

$$\hat{d}(x) = (x - \frac{1}{2}(\bar{x}_1 + \bar{x}_2))'S^{-1}((\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2) \tag{131}$$

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:

$$\hat{d}(\bar{x}_1) = (\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2) \tag{132}$$

$$\hat{d}(\bar{x}_2) = -(\bar{x}_1 - \bar{x}_2)'S^{-1}(\bar{x}_1 - \bar{x}_2) - \log(\pi_1/\pi_2). \tag{133}$$

Dann ist für die Bayes-Regel

$$\hat{\epsilon} = \pi_1\hat{\epsilon}_{12} + \pi_2\epsilon_{21} \tag{134}$$

mit

$$\hat{\epsilon}_{12} = \Phi\left(\frac{\log(\pi_2/\pi_1) - D^2/2}{D}\right), \quad \hat{\epsilon}_{21} = \Phi\left(\frac{-\log(\pi_2/\pi_1) - D^2/2}{D}\right) \tag{135}$$

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

$$\hat{\epsilon}_{12} = \hat{\epsilon}_{21} = \Phi(-D/2). \tag{136}$$

1.7 Diskriminanzanalyse bei kategorialen Daten

1.7.1 Volles multinomiales Modell

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

$$m = \prod_{i=1}^{p} m_i \tag{137}$$

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

$$d_k(x) = p(x|k)p(k), \tag{138}$$

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

$$\hat{\pi}(x, k) = \frac{n(x, k)}{N}, \tag{139}$$

und diese Schätzungen liefern die Diskriminanzfunktionen

$$\hat{d}_k(x) = \hat{\pi}(x, k), \tag{140}$$

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!

1.7.2 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

$$p(x_1, \ldots, x_p|k) = \prod_{i=1}^{p} \pi_{ki}^{x_i}(1 - \pi_{ik}^{1-x_i}) \tag{141}$$

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

$$\begin{aligned} d_k(x) &= \log p(x|k) + \log p(k) \\ &= \sum_{i^1}^{p} x_i \log \pi_{ik} + \sum_{i=1}^{p} (1 - x_i) \log(1 - \pi_{ik}) + \log p(k) \\ &= \sum_{i=1}^{p} \nu_i x_i + \nu_0, \end{aligned} \tag{142}$$

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.

1.7.3 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

$$\begin{aligned} \log n^{(k)}_{i_1 i_2 i_3} &= \mu^{(k)} + \mu^{(k)}_{1(i_1)} + \mu^{(k)}_{2(i_2)} + \mu^{(k)}_{3(i_3)} \\ &+ \mu^{(k)}_{12(i_1 i_2)} + \mu^{(k)}_{13(i_1 i_3)} + \mu^{(k)}_{23(i_2 i_3)} + \mu^{(k)}_{123(i_1 i_2 i_3)} \end{aligned} \tag{143}$$

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

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

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

$$\begin{aligned} \log n(x|k) &= \nu^{(k)} + \nu^{(k)}_1 x_1 + \nu^{(k)}_2 x_2 + \nu^{(k)}_3 x_3 \\ &\quad + \nu^{(k)}_{12} x_1 x_2 + \nu^{(k)}_{13} x_3 + \nu^{(k)}_{23} x_2 x_3 + \nu^{(k)}_{123} x_1 x_2 x_3, \end{aligned} \tag{144}$$

wobei $x_i = 0$ oder $x_i = 1$; alternativ kann auch $x_i = 1$ oder $x_i = -1$ gesetzt werden (Effektskalierung). Der Vergleich mit (143) 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

$$d_k(x) = \log p(k) - \log n_k + \log n(x|k), \tag{145}$$

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

(144) 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, 704die die Zuordnung von Objekten bzw. Personen zu Klassen ermöglicht.

1.7.4 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 ent-

sprechende Logit ist $\log p(k|x)/(1-p(k|x))$, und man erhält

$$\begin{aligned}\log \frac{p(k|x)}{1-p(k|x)} &= \lambda + \lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_p x_p \\ &\quad + \lambda_{12} x_1 x_2 + \cdots + \lambda_{12\ldots p} x_1 x_2 \cdots x_p. \end{aligned} \tag{146}$$

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.

2 Die Beziehung zwischen Diskriminanzanalyse und Kanonischer Korrelation

Diese Beziehung wurde um ersten Mal von Bartlett (1938) hergestellt.

Bei der Kanonischen Korrelation 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.

Man betrachte die Matrix G in (93), Seite 24. Korrespondierend zu den ersten n_1 Zeilen der ersten Gruppe in der Matrix X enthalten die ersten n_1 Zeilen von G den p-dimensionalen Einheitsvektor $(1, 0, \cdots, 0)'$; die folgenden n_2, zur zweiten Gruppe korrespondierenden Zeilen enthalten die 1 an der zweiten Stelle (d.h. in der zweiten Spalte), etc. Allgemein zeigt für eine gegebene Zeile von G die 1 in der k-ten Spalte an, daß das Objekt oder die Person in der entsprechenden Zeile von X zur k-ten Gruppe gehört. Die Elemente der Matrix G sind "Meßwerte", die einfach nur die Zugehörigkeit zu einer der K Gruppen anzeigen, (d.h. sie sind "Dummy"-Variablen).

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

Satz 2.1 *Zwischen den Kanonischen Korrelation R und dem Diskriminanzkriterium λ besteht die Beziehung*

$$\frac{\vec{u}\,' B\,\vec{u}}{\vec{u}\,' W\,\vec{u}} = \lambda = \frac{R^2}{1-R^2}, \quad \textit{d.h. } R^2 = \frac{\lambda}{1+\lambda} \tag{147}$$

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

$$(x'x)^{-1}(x'G)(G'G)^{-1}(G'x)\vec{u} = \lambda\vec{u} \tag{148}$$

wobei G für y substituiert wurde.

Beweis: Die Kanonische Korrelation ist durch

$$R_{uv} = \frac{Kov(\vec{u}, \vec{v})}{s_u s_v} = \frac{\vec{a}\,' x' y \vec{b}}{\sqrt{(\vec{a}\,' x' x \vec{a})(\vec{b}\,' y' y \vec{b})}}$$

gegeben, wobei $\vec{u} = x\vec{a}$, $\vec{v} = y\vec{b}$ die latenten Variablen sind, die so bestimmt werden, daß sie maximal miteinander korrelieren[9].

Es gilt

$$(x'x)^{-1}(x'y)(y'y)^{-1}(y'x)\vec{a} = R^2\vec{a} \tag{149}$$

wobei R^2 statt λ^2 geschrieben wurde, um eine Konfusion mit dem hier behandelten Diskriminanzkriterium λ zu vermeiden. Setzt man $y = G$, so kann man diesen Ausdruck in Termen der Matrizen T und B schreiben: nach (98) und (99), Seite 25, gilt ja

$$B = M'G'GM, \quad T = x'x$$

Nach (94) ist aber $x'y = x'G = M'(G'G)$ und $y'x = G'x = (G'G)M$ und natürlich $(y'y)^{-1} = (G'G)^{-1}$. In (149) eingesetzt ergibt sich

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

d.h. aber

$$T^{-1}B\,\vec{a} = R^2\vec{a} \tag{150}$$

Wegen $T = W + B$ folgt hieraus

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

so daß

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

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

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

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

$$\frac{\vec{a}\,' B\,\vec{a}}{\vec{a}\,' W\,\vec{a}} = \lambda = \frac{R^2}{1 - R^2}, \quad \text{d.h. } R^2 = \frac{\lambda}{1 + \lambda} \tag{151}$$

wobei λ natürlich das Diskriminanzkriterium ist, vergl. (100), p. 26; dies ist Gleichung (147). Der Vergleich mit (100) zeigt weiter, daß sich der Gewichtevektor $a = u$ als Eigenvektor der Matrix $(x'x)^{-1}(x'G)(G'G)^{-1}(G'x)$ ergibt. □

Anmerkungen:

[9] Im Skript zur Kanonischen Korrelation wurden diese Gleichungen (die Gleichungen (14) und (15)) in der Form $R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx}\vec{a}_k = \lambda_k^2\vec{a}_k$ und $R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}\vec{b}_k = \lambda_k^2\vec{b}_k$ angegeben. λ_k^2 ist der k-te Eigenwert von $R_{xx}^{-1}R_{xy}R_{yy}^{-1}R_{yx}$ bzw. $R_{yy}^{-1}R_{yx}R_{xx}^{-1}R_{xy}$. Es wurde λ_k^2 geschrieben, weil λ_k ein Hilfsterm bei der Ableitungen der Gleichungen ist und sich dessen Quadrat als Eigenwert der genannten Matrizen erweist.

1. Die Beziehung $R^2 = \lambda/(1+\lambda)$, d.h. (147), ist auch als *Roy'sches Kriterium* bekannt.

2. Aus der Beziehung (150), d.h. $T^{-1}B\,a = R^2 a$, folgt durch Multiplikation von links mit T

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

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

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

so daß wegen der Definition von B und T

$$R^2 = \frac{\vec{a}\,'B\,\vec{a}}{\vec{a}\,'T\,\vec{a}} = \frac{QS_{zw}}{QS_{ges}}. \tag{152}$$

Diese Beziehung entspricht der für Korrelationen bereits bekannten Beziehung $r^2 = QS_{zw}/QS_{ges}$. (Zur Erinnerung: $\lambda = QS_{zw}/QS_{inn}$.) R^2 entspricht also einem Determinationskoeffizienten.

3. Die Beziehung (147) ist für die Interpretation des Diskriminanzkoeffizienten λ von Interesse. Offenbar ist ja $\lambda \geq 0$, da QS_{zw} und QS_{inn} als Quadratsummen nicht kleiner als 0 werden können. Für $QS_{inn} > 0$, aber $QS_{zw} \to 0$ (d.h. die Unterschiede zwischen den Gruppen verschwinden) folgt $\lambda \to 0$. Umgekehrt strebt λ für $QS_{zw} > 0$ und $QS_{inn} \to 0$, (d.h. die "Fehler"varianz geht gegen Null) gegen unendlich, $\lambda \to \infty$, so daß $0 \leq \lambda < \infty$. Die Beziehung $R^2 = \lambda/(1+\lambda)$ erlaubt nun, die Unterschiede zwischen den Gruppen auf das Intervall $(0, 1)$ abzubilden, d.h. wie einen Determinationskoeffizienten zu interpretieren, denn für $\lambda \to \infty$ folgt $R^2 \to 1$, und $\lambda \to 0$ impliziert $R^2 \to 0$.

4. In Definition 1.4 wurde angemerkt, daß statt des Ausdrucks *Diskriminanzfunktion* auch der Ausdruck *kanonische Variable* üblich ist. Der Hintergrund dieses Sprachgebrauchs wird in Satz 2.1 deutlich; der Vektor $\vec{u}$ ergibt sich als der Eigenvektor $\vec{a}$.

3 Beispiele

Beispiel 3.1 Bei gesunden Männern Wurden die Variablen x_1: Alter, x_2: Blutdruck, und x_3 Cholesterinspiegel gemessen. Die Frage ist, ob sich aus diesen Werten das Risiko für eine spätere Herzgefäßkranzerkrankung feststellen läßt. Am Ende einer Beobachtungsperiode waren 71 der Männer erkrankt. Es ergaben sich die folgenden Messungen:

Tabelle 4: Mittelwerte und Varianzen für Gesunde und Erkrankte

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

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

$$S = \begin{pmatrix} 214.26 & 72.37 & .61 \\ 72.73\ 985 & 212.44 & 175.53 \\ 195.61 & 175.53 & 1820.61 \end{pmatrix}. \tag{153}$$

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

$$S^{-1} = \begin{pmatrix} 214.26 & .72.37 & 195.61 \\ 72.73 & 212.44 & 175.53 \\ 195.61 & 175.53 & 1820.61 \end{pmatrix}. \tag{154}$$

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

Tabelle 5: 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 $df = n_1033 + n_2 - 3 - 1 = 828$ ist er hochsignifikant. Dies bedeutet, daß sich die beiden Gruppen (Erkrankte und Gesund) anhand der Risikofaktoren x_1 , x_2 und x_3 gut trennen lassen. Da der partielle F-Wert für x_3 relativ klein ist, ist es möglich, daß der Cholesteringehalt kaum zur Trennung der Gruppen beiträgt.

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

Die Mahalanobis-Distanz ist gemäß Tabelle 5 $\delta^2 = .815$. 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

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

Tabelle 6: Klassifikation nach der ML-Regel

	Klassifizierung		Σ
Lernstichprobe	krank	gesund	
Kranke	51	20	71
Gesunde	272	489	767

Tabelle 7: Klassifikation nach der Bayes-Regel

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

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 082 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) 155erhoben: X_1 = "Outdoor", X_2 = "Convivial", und $X_3$156 = "Conservative". Die Angestellten Wurden in drei Klassen eingeteilt: p "Passenger Agents", m "Mechanics". und o "Operations Control Agents", vergl. Tabelle 8. Ein hoher Wert auf einer Skala reflektiert eine hohe Präferenz für die entsprechende Aktivität.Die Tabelle 9 gibt die Mittelwerte der drei Variablen für die einzelnen Gruppen.

Die zusammengefasste ("pooled") Varianz-Kovarianz-Matrix wird in der Tabelle 10 angegeben. 181Die Matrix B der Varianzen-Kovarianzen zwischen den Mittelwerten findet man in der Tabelle 11. Die Tabelle 12 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 7 entsprechend kann man folgern, dass in erster Linie die Gruppe 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 12 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". □

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 245man die Wurzel $\mathbf{A}^{1/2}$ einer positiv definiten, quadratischen

Tabelle 8: 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

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

Tabelle 9: 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 10: 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 11: Varianz-Kovarianz-Matrix B

	$\bar{x}_1$	$\bar{1}90x_2$	$\bar{x}_3$
$\bar{x}_1$	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 12: Eigenvektoren $\vec{u}$ und Eigenwerte λ

Variable	u_1 207	u_2
X_1	.091	-.974
X_2	.988	.0386
X_3	-.124	-.221
λ	$\lambda_1 = 1.675$	$\lambda_2 = .155$

$\mathbf{P}'\mathbf{P} = \mathbf{I}$ die Einheitsmatrix. Λ ist die Diagonalmatrix der igenwerte. Dann gilt

$$\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}' = \sum_{k=1}^{n} \lambda_k \mathbf{p}_k \mathbf{p}_k'. \tag{155}$$

Es ist

$$\mathbf{A}^{-1} = (\mathbf{P}\Lambda\mathbf{P}')^{-1} = (\mathbf{P}')^{-1}\Lambda^{-1}\mathbf{P}^{-1}, \tag{156}$$

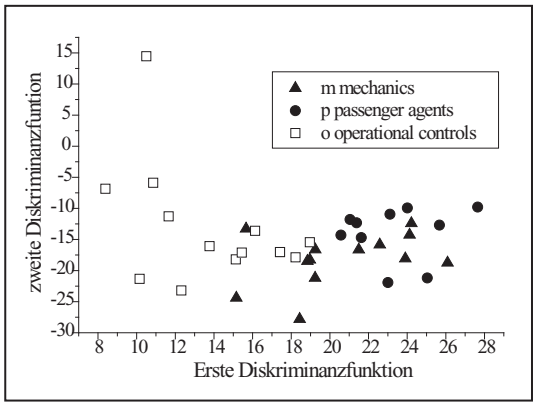
und da $\mathbf{P}^{-1} = \mathbf{P}'$ und $(\mathbf{P}')^{-1} = \mathbf{P}$, hat man

$$\mathbf{A}^{-1} = \mathbf{P}\Lambda^{-1}\mathbf{P}'. \tag{157}$$

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

$$\mathbf{A}^{1/2} = \sum_{k=1}^{n} \sqrt{\lambda_k}\mathbf{p}_k\mathbf{p}' = \mathbf{P}\Lambda^{-1/2}\mathbf{P}'. \tag{158}$$

Abbildung 7: Projektion auf die maximal diskriminierende Achse u_1



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

$$(\mathbf{A}^{1/2})' = \mathbf{A}^{1/2}, \quad (\mathbf{A}^{1/2} \text{ ist symmetrisch}) \tag{159}$$

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A} \tag{160}$$

$$(\mathbf{A}^{1/2})^{-1} = \sum_{k=1}^{n} \frac{1}{\sqrt{\lambda_k}} \mathbf{p}_k \mathbf{p}_k' = \mathbf{P}\Lambda^{-1/2}\mathbf{P}' \tag{161}$$

$$\mathbf{A}^{1/2}\mathbf{A}^{-1/2} = \mathbf{I}, \quad \mathbf{A}^{-1/2}\mathbf{A}^{-1/2} = \mathbf{A}^{-1}. \tag{162}$$

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

$$(\mathbf{a}'\mathbf{b}) \leq (\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b}), \tag{163}$$

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

$$\begin{aligned} 0 < |\mathbf{a} - x\mathbf{b}|^2 = (\mathbf{a} - x\mathbf{b})'(\mathbf{a} - x\mathbf{b}) &= \mathbf{a}'\mathbf{a} - x\mathbf{b}'\mathbf{a} + \mathbf{a}'x\mathbf{b} + x\mathbf{b}'x\mathbf{b} \\ &= \mathbf{a}'\mathbf{a} - 2x(x\mathbf{b}'\mathbf{a}) + x^2(\mathbf{b}'x\mathbf{b}). \end{aligned} \tag{164}$$

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

$$(\mathbf{a}'\mathbf{x})' \leq (\mathbf{a}'\mathbf{B}\mathbf{a})(\mathbf{b}'\mathbf{B}^{-1}\mathbf{b}), \tag{165}$$

und $(\mathbf{a}'\mathbf{b})' = (\mathbf{a}'\mathbf{A}\mathbf{a})(\mathbf{b}'\mathbf{A}^{-1}\mathbf{b})$ gilt genau dann, wenn $\mathbf{a} = c\mathbf{A}^{-1}\mathbf{b}$ oder $\mathbf{b} = c\mathbf{A}\mathbf{a}$, für ein $c \in \mathbb{R}$.

Beweis: Es sei $\mathbf{B}^{1/2} = \sum_{k=1}^{n} \sqrt{\lambda_k}\mathbf{p}_k\mathbf{p}_k'$, λ_k und $\mathbf{p}_k$ die Eigenwerte und Eigenvektoren von B. Dann ist

$$\mathbf{B}^{-1/2} = \sum_{k=1}^{n} \frac{1}{\sqrt{\lambda_k}}\mathbf{p}_k\mathbf{p}_k'.$$

Dann ist

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

Wendet man jetzt die Cauchy-Schwarzsche Ungleichung auf die Vektoren $\mathbf{B}^{1/2}\mathbf{a}$ und $\mathbf{B}^{-1/2}\mathbf{a}$ an, so folgt die Behauptung. □

Es sei nun $\mathbf{B}$ eine positiv definite $(n\times n)$-Matrix und $\mathbf{a}$ sei ein gegebener n-dimensionaler Vektor. $\mathbf{x}$ sei ein beliebiger n-dimensionaler Vektor. Dann gilt

$$\max_{\mathbf{x}\neq 0} \frac{\mathbf{x}'\mathbf{a}}{\mathbf{x}'\mathbf{B}\mathbf{x}} = \mathbf{a}'\mathbf{B}^{-1}\mathbf{a}, \tag{166}$$

und das Maximum wird angneommen für $\mathbf{x} = c\mathbf{B}^{-1}\mathbf{a}$, für beliebige reelle Konstante c.

Beweis: Nach der verallgemeinerten 406Cauchy-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 $\mathbf{x} \neq 0$. 412Nimmt 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 (166). □

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 auf die Länge 1 korrespondierenden Eigenvektoren $\mathbf{p}_1, \ldots, \mathbf{p}_n$. Dann gilt

$$\max_{\mathbf{x}\neq 0} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_1 \quad \text{für } \mathbf{x} = \mathbf{p}_1 \tag{167}$$

$$\min_{\mathbf{x}\neq 0} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_n, \quad \text{für } \mathbf{x} = \mathbf{p}_n \tag{168}$$

$$\max_{\mathbf{x}\perp \mathbf{p}_1,\ldots,\mathbf{p}_k} \frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \lambda_{k+1}, \quad \text{für } 440\mathbf{x} = \mathbf{p}_{k+1}, \quad k = 1, \ldots n-1, \tag{169}$$

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

$$\begin{aligned}
\frac{\mathbf{x}'\mathbf{B}\mathbf{x}}{\mathbf{x}'\mathbf{x}} &= \frac{\mathbf{x}'\mathbf{B}^{1/2}\mathbf{B}^{1/2}\mathbf{x}}{\mathbf{x}'\mathbf{P}'\mathbf{P}\mathbf{x}} \\
&= \frac{\mathbf{x}'\mathbf{P}\Lambda^{1/2}\mathbf{P}'\mathbf{x}}{\mathbf{y}'\mathbf{y}} \\
&= \frac{\mathbf{y}'\Lambda\mathbf{y}}{\mathbf{y}'\mathbf{y}} \\
&= \frac{\sum_{k=1}^n \lambda_k y_k^2}{\sum_{k=1}^n y_k^2} \leq \lambda_1 \frac{\sum_{k=1}^n y_k^2}{\sum_{k=1}^n y_k^2} = \lambda_1.
\end{aligned} \tag{170}$$

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.

$$\frac{\mathbf{p}_1'\mathbf{B}\mathbf{p}_1}{\mathbf{p}_1'\mathbf{p}_1} = \mathbf{p}_1'\mathbf{B}\mathbf{p}_1 = \lambda_1. \tag{171}$$

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 (171) 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 1.3

Beweis: Der Vektor $\vec{u} = (u_1, \ldots, u_j, \ldots, u_p)'$ soll so bestimmt werden, dass λ maximal wird. Dies bedeutet, dass λ als Funktion der u_j, $1 \leq j \leq p$ aufgefasst wird, und diese Funktion wird maximiert, indem man die partiellen Ableitungen $\partial\lambda/\partial u_j$ bildet und gleich Null setzt; die Lösungen des entstehenden Gleichungssystems liefern die $\hat{u}_j$, für die λ maximal wird. Die Bestimmung von $\partial\lambda/\partial u_j$ bedeutet die Bildung der partiellen Ableitungen $\partial\vec{u}/\partial u_j$; man findet $\partial\vec{u}/\partial u_j = (0, \ldots, 0, 1, 0, \ldots, 0)'$, wobei die 1 an der j-ten Stelle steht, denn $\partial u_j/\partial u_j = 1$, und $\partial u_k/\partial u_j = 0$ für $k \neq j$. Also ist $\partial u/\partial u_j = \epsilon_j$, ϵ_j der j-te Einheitsvektor.

Die Anwendung der Produkt- und Quotientenregel auf (100) liefert dann

$$\frac{\partial\lambda}{\partial u_j} = \frac{(\epsilon_j' B\vec{u} + \vec{u}\,' B\epsilon_j)\vec{u}\,' W\vec{u} - \vec{u}\,' Bu(\epsilon_j' W\vec{u} + \vec{u}\,' W\epsilon_j)}{(\vec{u}\,' W\vec{u})^2}, \quad \text{für alle } j.$$

Hierbei sind aber $\epsilon_j' B\vec{u}$, $\vec{u}' B\epsilon_j$ Skalare, die den gleichen Wert haben, ebenso $\epsilon_j' W\vec{u}$ und $\vec{u}' W\epsilon_j$, so daß man $\epsilon_j' B\vec{u} + \vec{u}' B\epsilon_j = 2\epsilon_j' B\vec{u}$ schreiben kann; analog ist $\epsilon_j' W\vec{u} + u' W\epsilon_j = 2\epsilon_j' W\vec{u}$. Die $\epsilon_1' Bu, \ldots, \epsilon_p' Bu$ bilden aber gerade die Komponenten des Vektors $B\vec{u}$, und die Komponenten $\epsilon_1' W\vec{u}, \ldots, \epsilon_p' W\vec{u}$ bilden den Vektor $W\vec{u}$, so dass

$$\frac{\partial\lambda}{\partial\vec{u}} = \frac{2((B\vec{u})(\vec{u}' W\vec{u}) - (\vec{u}' B\vec{u})W\vec{u})}{(\vec{u}' W\vec{u})^2} \tag{172}$$

Für $\partial\lambda/\partial u = 0$ folgt

$$B\vec{u}(\vec{u}' W\vec{u}) = (\vec{u}' B\vec{u})W\vec{u}),$$

d.h.

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

also

$$B\vec{u} = \lambda W\vec{u},$$

und Multiplikation von links mit W^{-1} liefert schließlich

$$W^{-1}B\vec{u} = \lambda\vec{u}.$$

Literatur

[1] Bartlett, M.S. (1938) Further aspects of the theory of multiple regression. *Proc. Camb. Philos. Soc.*, 34, 33 – 40

[2] Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*}, 7, 179 – 188