Technische Universität München-Weihenstephan
Mathematik und Statistik, Informations- und Dokumentationszentrum

Biometrische und Ökonometrische Methoden

Einführung in die Diskriminanzanalyse

1 Fragestellung

Mit der Diskriminanzanalyse können Gruppenunterschiede anhand von mehreren Merkmalen gleichzeitig untersucht werden. Man kann z.B. landwirtschaftliche Betriebe den drei Gruppen Pflanzenbaubetriebe, Tierhaltungsbetriebe und Gemischtbetriebe aufgrund von erhobenen Betriebskenngrößen wie landwirtschaftliche Nutzfläche, Großvieheinheiten, Betriebseinkomen, etc. zuordnen. Die Merkmale, die dabei in die Untersuchung eingehen, sind die Betriebskenngrößen und die Untersuchungsobjekte sind die Betriebe, wobei für jeden Betrieb bekannt ist, zu welcher Gruppe er gehört.

Man kann nun folgende Fragestellungen untersuchen:

- **Prüfung der Diskriminanz:**
  Unterscheiden sich die Gruppen anhand der Merkmale ihrer zugehörigen Objekte signifikant?
  Beispiel: Existieren signifikante Unterschiede zwischen Standorten aufgrund von Zeigerpflanzen?
- **Prüfung der Merkmalsvariablen:**
  Welche Merkmale sind zur Unterscheidung der Gruppen wichtig?
  Beispiel: Welche Fettsäuren sind zur Gruppierung verschiedener Fettarten geeignet?
- **Klassifikation:**
  In welche Gruppe ist ein weiteres Objekt, dessen Gruppenzugehörigkeit unbekannt ist, einzuordnen?
  Beispiel: Welche Krankheit ist bei bestimmten Symptomen zu diagnostizieren?

2 Abgrenzung zur Clusteranalyse

Der Hauptunterschied der Diskriminanzanalyse (DA) zur Clusteranalyse (CA) ist, daß bei der Clusteranalyse die wirkliche Gruppenzugehörigkeit unbekannt ist, und man versucht, die Objekte in Gruppen einzuordnen (Bild 1).

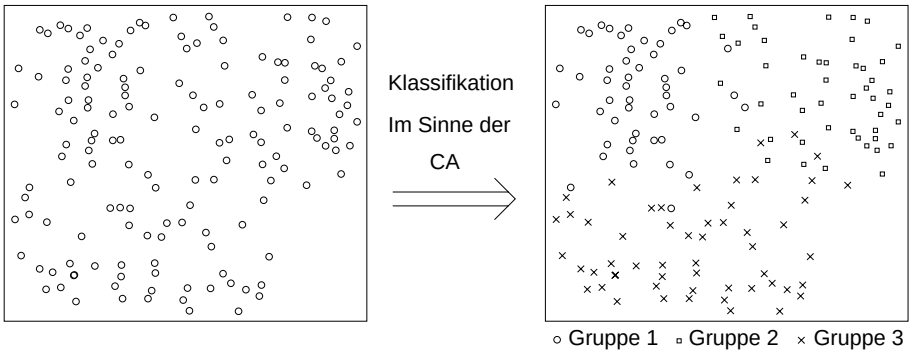

Bild 1: Einteilung von Objekten in unbekannte Gruppen

Bei der Diskriminanzanalyse dagegen sind die Gruppenzugehörigkeiten bekannt, und man will die Gruppenunterschiede untersuchen oder neue Objekte in die bekannten Gruppen einordnen (Bild 2).

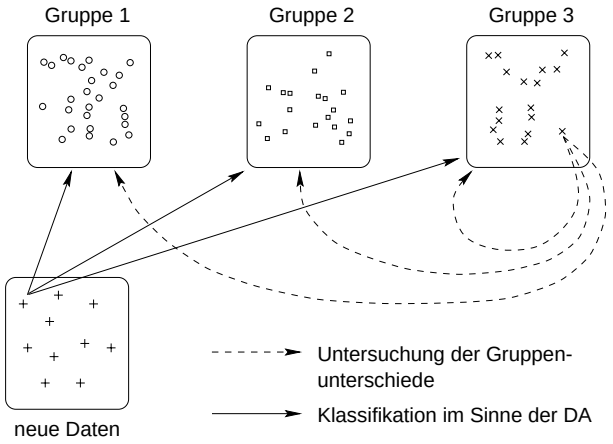

Bild 2: Einteilung neuer Objekte in bekannte Gruppen

Es besteht also ein deutlicher Unterschied zwischen einer Klassifikation im Sinne der Clusteranalyse und einer Klassifikation im Sinne der Diskriminanzanalyse.

Die beiden Methoden ergänzen sich jedoch insofern, als mit Hilfe der Diskriminanzanalyse die von einer vorgeschaltenen Clusteranalyse erzeugten Gruppen näher untersucht werden können. Man versucht also mit der Clusteranalyse diese Gruppen zu erzeugen, in der Diskriminanzanalyse werden die bekannten Gruppen dann untersucht.

3 Grundsätzliche Voraussetzungen

Es werden n Objekte, die zu g bekannten Gruppen gehören, anhand von p Merkmalsvariablen untersucht. Die Anzahl der Objekte in jeder einzelnen Gruppe sei n_j, x_i sei die i-te Merkmalsvariable und x_{ijk} die Ausprägung des i-ten Merkmals des k-ten Objekts der j-ten Gruppe. Es ist natürlich $n_1 + n_2 + \ldots + n_g = n$ und es muß $1 \leq i \leq p$, $1 \leq j \leq g$ und $1 \leq k \leq n_j$ sein.

Beispiel:

Eine Bank führt eine Kreditwürdigkeitsprüfung durch, bei der sie $n = 100$ Kreditkunden aufgrund von $p = 4$ Merkmalen x_1 = Einkommen, x_2 = Alter, x_3 = Beschäftigungsdauer, x_4 = Summe anderer Kredite in $g = 3$ Risikoklassen G_1 = niedrig ($n_1 = 60$ Kunden), G_2 = mittel ($n_2 = 30$ Kunden), G_3 = hoch ($n_3 = 10$ Kunden) einteilt. x_{135} ist dann das Einkommen des 5. Kunden in der Risikogruppe hoch.

Es müssen einige grundsätzliche Voraussetzungen erfüllt sein, um eine Diskriminanzanalyse durchführen zu können:

- $g \geq 2$, d.h. mindestens zwei Gruppen.
- $n_j \geq 2$, d.h. mindestens zwei Objekte pro Gruppe.
- $1 \leq p \leq n - 3$, d.h. mindestens ein Merkmal, maximal jedoch 3 weniger als ingesamt Objekte in der Untersuchung vorhanden sind.
- Die Merkmale müssen mit einer Intervallskala erfaßt worden sein.
- Kein Merkmal darf sich als Linearkombination anderer Merkmale schreiben lassen.
- Die Objekte jeder Gruppe müssen aus einer normalverteilten Grundgesamtheit stammen.

Somit müssen im ersten Schritt einer Diskrimininazanalyse die Gruppen und die Merkmalsvariablen festgelegt werden, die in die Analyse aufgenommen werden sollen. Die Auswahl der Merkmalsvariablen x_i ($i = 1, 2, \ldots, p$) erfolgt auf hypothetischer Basis aufgrund theoretischer Überlegungen oder praktischer Gegebenheiten. So kann im vorhergehenden Beispiel auch noch die Kinderzahl als Merkmal aufgenommen werden. Welche diskriminatorische Bedeutung die einzelnen Variablen haben läßt sich erst nach Durchführung der Diskriminanzanalyse abschätzen. Die Wahl der zu untersuchenden Gruppen beruht entweder auf sachlogischen Überlegungen oder wird mittels einer vorgeschalteten Clusteranalyse ermittelt. Damit steht dann auch die Anzahl g der Gruppen fest. Es kann beispielsweise durchaus sinnvoll sein, im vorhergehenden Beispiel nur die Risikogruppen hoch und niedrig zu bilden.

4 Kanonische Diskriminanzfunktionen

Das Prinzip der Diskriminanzanalyse ist, mehrere Variablen bei minimalem Informationsverlust durch eine Linearkombination zu einer einzigen zusammenzufassen. Die Kombination der Merkmalsvariablen erfogt durch die sog. **Diskriminanzfunktion** (Trennfunktion). Es gibt im wesentlichen zwei Arten von Diskriminanzfunktionen. Die einen sind die **kanonischen** Diskriminanzfunktionen, die in diesem Abschnitt eingeführt werden, die anderen sind die **linearen** bzw. **quadratischen** Diskriminanzfunktionen, die in Abschnitt 5.1 und 6 erklärt werden.

4.1 Formulierung der kanonischen Diskriminanzfunktion

Eine **kanonische Diskriminanzfunktion** ist eine Funktion der Form

$$y = b_0 + b_1\, x_1 + b_2\, x_2 + \ldots + b_p\, x_p \tag{1}$$

Dies ist eine lineare Funktion mit der **Diskriminanzvariablen** y und den **Merkmalsvariablen** x_i. Die Koeffizienten $b_1, b_2, \ldots, b_p$ heißen **Diskriminanzkoeffizienten**, während b_0 einfach als **konstantes Glied** bezeichnet wird. Die Diskriminanzfunktion ist also eine Abbildung aus dem p-dimensionalen Merkmalsraum in den eindimensionalen Raum der Diskriminanzvariablen.

Im ersten Schritt der Diskriminanzanalyse werden die unbekannten Koeffizienten b_i so geschätzt, daß ein sog. **Diskriminanzmaß** maximiert wird (näheres siehe Abschnitt 4.3). Dann wird noch ein Schätzer für das konstante Glied b_0 errechnet. Somit steht eine kanonische Diskriminanzfunktion fest.

Sobald die Koeffizienten und der konstante Term bekannt sind, können die n **Diskriminanzwerte** der untersuchten Objekte berechnet werden:

$$y_{jk} = b_0 + b_1\, x_{1jk} + b_2\; x_{2jk} + \ldots + b_p\, x_{pjk} \tag{2}$$

mit $1 \leq j \leq g$ und $1 \leq k \leq n_j$. Dies sind wegen $n_1 + n_2 + \ldots + n_g = n$ genau n Werte.

4.2 Geometrische Veranschaulichung

Die Diskriminanzfunktion (1) bildet geometrisch interpretiert bei nur zwei Merkmalen eine Ebene im Merkmalsraum. Die durch die Diskriminanzfunktion erzeugte Diskriminanzvariable läßt sich dann anschaulich auch als Gerade darstellen und heißt **Diskriminanzachse**. Für einen festen Wert von $y_{jk} = c$ erhält man eine **Isoquante** der Diskriminanzfunktion, also eine Gerade der Form

$$x_2 = \frac{c - b_0}{b_1} - \frac{b_1}{b_2} \cdot x_1 \tag{3}$$

Diese Isoquante ist der geometrische Ort aller Merkmalskombinationen (x_1, x_2), für die die Diskriminanzfunktion den Wert c liefert. Zeichnet man für verschiedene Werte $c_1, c_2, \ldots$ die Isoquanten in das Koordinatensystem der Merkmalsvariablen x_1 und x_2, so erhält man eine zweidimensionale Darstellung der Diskriminanzfunktion. Noch einfacher ist die Darstellung in Form der **Diskriminanzachse**. Dies ist eine Gerade durch den Nullpunkt senkrecht zu den Isoquanten und hat die Gleichung:

$$x_2 = \frac{b_2}{b_1} \cdot x_1 \tag{4}$$

Die Skalierung der Diskriminanzachse erfolgt so, daß die Projektion eines beliebigen Punkts (x_1, x_2) den zugehörigen Diskriminanzwert y liefert.

Beispiel:

Für die Diskriminanzfunktion $y = -2+0.8x_1-0.5x_2$ *sind in Bild 3 die Isoquanten für die* y*-Werte* $-4, -2, 0, 2$ *und* 4 *eingezeichnet. Die Isoquante lautet z.B. für* $y = 0$*:* $x_2 = -4 + 1.6x_1$*. Senkrecht zu den Isoquanten verläuft die Diskriminanzachse.*

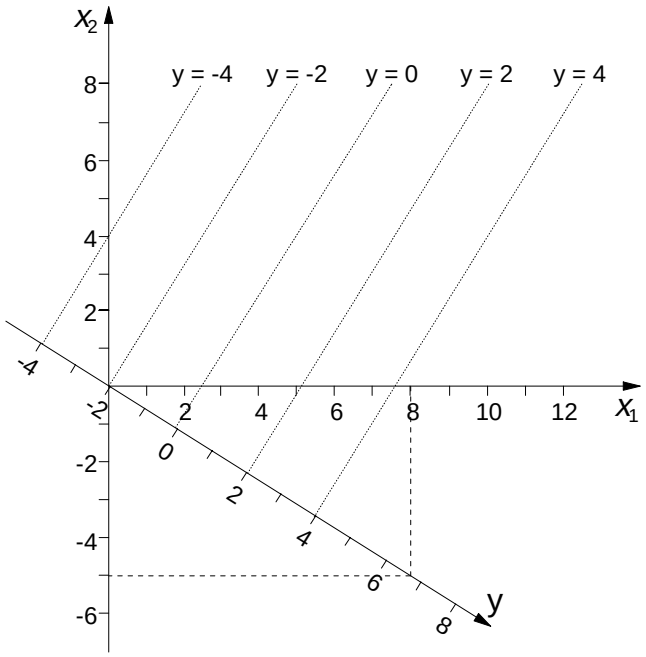

Bild 3: Diskriminanzachse für die Diskriminanzfunktion $y = -2 + 0.8x_1 - 0.5x_2$

Die Diskriminanzachse läßt sich auch ohne Bildung der Isoquanten sehr einfach konstruieren, indem man für einen beliebigen Wert z den Punkt $(b_1 z, b_2 z)$ in das Koordinatensystem einträgt und mit dem Nullpunkt verbindet.

Beispiel:

Für $z = 10$ erhält man für die Diskriminanzfunktion des vorherigen Beispiels den Punkt $(8, -5)$. Die Diskriminanzfunktion liefert für die Merkmalskombination $x_1 = 8$ und $x_2 = 5$ den Wert 6.9, der sich in Bild 3 auf der Diskriminanzachse ablesen läßt. Das konstante Glied $b_0 = -2$ ist die Entfernung des Nullpunkts der Diskriminanzachse zum Nullpunkt des Koordinatensystems der Merkmalsvariablen.

Bild 4 zeigt Punktwolken von zwei Gruppen als Ellipsen dargestellt sowie die Häufigkeitsverteilungen, die durch Projektion der Punktwolken auf die Diskriminanzachse resultieren würde.

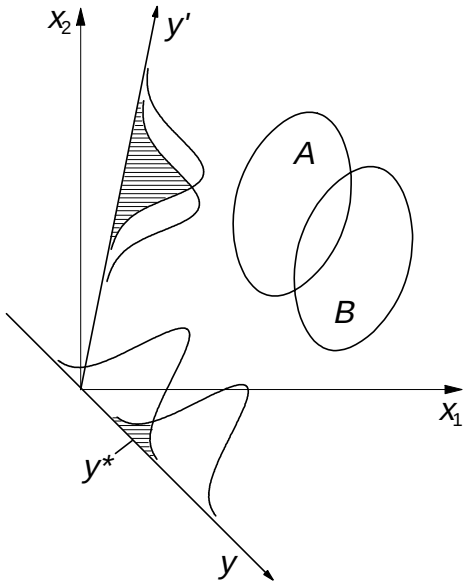

Bild 4: Verteilungen zweier Gruppen und ihre Projektionen auf die Diskriminanzachse

Die Unterschiedlichhkeit von Gruppen läßt sich geometrisch durch die Überschneidung der Häufigkeitsverteilungen erfassen. Je geringer diese Überschneidung ist, desto größer ist die Unterschiedlichkeit. Bild 4 zeigt auch einen kritischen Wert y^*, auf der Diskriminanzachse y. Ein Objekt der Gruppe A wird fehlklassifiziert, wenn sein Diskriminanzwert rechts vom kritischen Wert liegt ($y > y^*$). Umgekehrt wird ein Objekt der Gruppe B fehlklassifiziert, wenn sein Diskriminanzwert links vom kritischen Wert liegt ($y < y^*$).

Bild 4 zeigt eine weitere Diskriminanzachse y', die offensichtlich schlechter als y diskriminiert, da sich die projizierten Verteilungen stärker überschneiden. Durch Drehung um den Ursprung lassen sich unendlich viele Diskriminanzachsen bilden. Optimal ist diejenige Diskriminanzachse, bei der die Überschneidung der projizierten Verteilungen minimal ist.

4.3 Schätzung der Koeffizienten

Die Koeffizienten $b_1, b_2 \ldots, b_p$ der Diskriminanzfunktion (1) werden derart bestimmt, daß ein Diskriminanzmaß maximal wird. Das dabei verwendete Diskriminanzmaß D' lautet:

$$D' = \frac{\text{Varianz zwischen den Gruppen}}{\text{Varianz innerhalb der Gruppen}} = \frac{\text{SQ}_{\text{zwischen}}/(g-1)}{\text{SQ}_{\text{innerhalb}}/(n-g)} \tag{5}$$

Dabei ist

$$\text{SQ}_{\text{zwischen}} = \frac{1}{g-1} \cdot \sum_{j=1}^{g} n_j \, (\overline{y}_{j.} - \overline{y}_{..})^2 \tag{6}$$

$$\text{SQ}_{\text{innerhalb}} = \frac{1}{n-g} \cdot \sum_{j=1}^{g} \sum_{k=1}^{n_j} (y_{jk} - \overline{y}_{j.})^2 \tag{7}$$

mit $\overline{y}_{j.}$ als Mittelwert der Diskriminanzwerte $\overline{y}_{jk}$ innerhalb der j-ten Gruppe und $\overline{y}_{..}$ als dem Gesamtmittelwert aller Diskriminanzvariablen.

Der Faktor $\dfrac{n-g}{g-1}$ ist konstant, und somit genügt es, als Diskriminanzmaß

$$D = \frac{\sum_{j=1}^{g} n_j \, (\overline{y}_{j.} - \overline{y}_{..})^2}{\sum_{j=1}^{g} \sum_{k=1}^{n_j} (\overline{y}_{jk} - \overline{y}_{j.})^2} \tag{8}$$

zu maximieren. Der dabei entstehende Maximalwert $\gamma = \max(D)$ heißt **Eigenwert** der Diskriminanzfunktion.

Multipliziert man nun die b_i mit einem Faktor, daß $\text{SQ}_{\text{innerhalb}}/(n-g) = 1$ ist, und bestimmt dann b_0 so, daß der Gesamtmittelwert der Diskriminanzvariablen $\overline{y}_{..} = 0$ ist, so erhält man die **normierte Diskriminanzfunktion**.

Das oben beschriebene Verfahren erzeugt die Diskriminanzfunktion, die die größte diskriminatorische Bedeutung besitzt. Man kann aber noch weitere Diskriminanzfunktionen erstellen, indem man zur Berechnung dieser weiteren Funktionen die Maximierungsvorschrift zur Bestimmung der Koeffizienten modifiziert. Die Anzahl ist jedoch begrenzt durch die Anzahl der Gruppen und der Merkmalsvariablen. Falls mehr Variablen als Gruppen vorhanden sind, gibt es $g-1$ und sonst p Diskriminanzfunktionen. Kurz gesagt gibt es $q = \min(g-1, p)$ lineare Diskriminanzfunktionen, und diese stehen senkrecht aufeinander. Wie sie im einzelnen berechnet werden, soll in diesem Umdruck nicht behandelt werden. Bei der Berechnung entstehen neben den Diskriminanzfunktionen auch noch q zugehörige

Eigenwerte $\gamma = \gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_q$, und je größer ein Eigenwert ist, desto höher ist die relative Wichtigkeit der zugehörigen Diskriminanzfunktion. Ein Maß dafür ist der sog. **Eigenwertanteil**:

$$\mathrm{A}_i = \frac{\gamma_i}{\gamma_1 + \gamma_2 + \ldots + \gamma_q} \quad (1 \leq i \leq q) \tag{9}$$

Dieser gibt den Anteil der durch die i-te Diskriminanzfunktion erklärten Streuung an.

Beispiel:

Tab. 1 zeigt den Gehalt an Ölsäure und Linolsäure in g/100 g *Öl bei Sojaöl und Maisöl von jeweils* 25 *Proben.*

Es liegen also $g = 2$ *Gruppen (Sojaöl und Maisöl) und* $p = 2$ *Merkmale (Ölsäuregehalt* (x_1) *und Linolsäuregehalt* (x_2)) *vor. Die beiden Gruppen sollen nun aufgrund ihres unterschiedlichen Fettsäuremusters optimal getrennt (diskriminiert) werden. Das Streudiagramm in Bild 5 zeigt offensichtlich die beiden Gruppen, die sich jedoch etwas überlappen.*

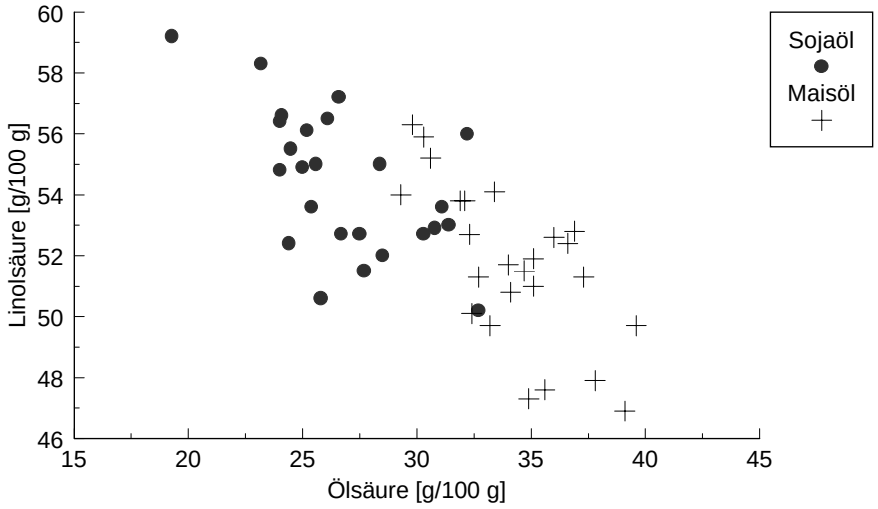

Bild 5: Streudiagramm des Ölsäure- und Linolsäuregehalts von Sojaöl und Maisöl

Die Diskriminanzfunktion[1] *lautet:* $y = -18.1492 + 0.3835\, x_1 + 0.1216\, x_2$

Mit $\overline{y}_{..} = 0$ *sowie* $\overline{y}_{1.} = -1.25$ *und* $\overline{y}_{2.} = 1.25$ *folgt für den Zähler von Gleichung (8):*

$$\sum_{j=1}^{g} n_j\,(\overline{y}_{j.} - \overline{y}_{..})^2 = \sum_{j=1}^{2} 25\,\overline{y}_{j.}^2 = 25 \cdot \overline{y}_{1.} + 25 \cdot \overline{y}_{2.} = 25 \cdot (-1.25)^2 + 25 \cdot 1.25^2 = 78.125$$

[1] Die Diskriminanzfunktion wurde mit dem Statistikpaket SPSS berechnet.

Sojaöl [g/100 g]		Maisöl [g/100 g]	
Ölsäure	Linolsäure	Ölsäure	Linolsäure
25.4	53.6	30.3	55.9
25.8	50.6	32.4	50.1
30.8	52.9	29.3	54.0
31.1	53.6	33.2	49.7
26.1	56.5	37.8	47.9
32.7	50.2	31.9	53.8
31.4	53.0	35.1	51.0
30.3	52.7	34.1	50.8
25.6	55.0	39.1	46.9
25.0	54.9	39.6	49.7
25.2	56.1	36.6	52.4
23.2	58.3	32.7	51.3
19.3	59.2	34.9	47.3
32.2	56.0	29.8	56.3
28.4	55.0	32.3	52.7
26.7	52.7	32.1	53.8
24.0	56.4	34.0	51.7
24.5	55.5	30.6	55.2
28.5	52.0	35.1	51.9
24.4	52.4	36.0	52.6
27.5	52.7	34.7	51.5
24.0	54.8	33.4	54.1
26.6	57.2	35.6	47.6
27.7	51.5	36.9	52.8
24.1	56.6	37.3	51.3

Tabelle 1: Ölsäuregehalt (x_1) und Linolsäuregehalt (x_2) von Sojaöl und Maisöl ($n_1 = 25$, $n_2 = 25$, $n = 50$)

Die Berechnung des Nenners von (8) liefert:

$$\sum_{j=1}^{g}\sum_{k=1}^{n_j}(\overline{y}_{jk}-\overline{y}_{j.})^2 = \sum_{k=1}^{25}(\overline{y}_{1k}+1.25)^2 + \sum_{k=1}^{25}(\overline{y}_{2k}-1.25)^2 = 29.148+18.851 = 47.999$$

Somit ist das Diskriminanzmaß $D = \dfrac{78.125}{47.999} = 1.628$. *Da es im Zweigruppenfall nur eine Diskriminanzfunktion gibt, existiert auch nur ein Eigenwert der Diskriminanzfunktion* $\gamma_1 = \gamma = 1.628$ *und der Eigenwertanteil ist natürlich* $\mathrm{A} = 1$.

4.4 Prüfung der Diskriminanz

Wie bereits erwähnt, kann man i.a. q Diskriminanzfunktionen berechnen, deren diskriminatorische Bedeutung von der Größe der zugehörigen Eigenwerte abhängt. Damit stellt sich sofort die Frage, wie groß ein Eigenwert mindestens sein muß, damit seine Diskriminanzfunktion überhaupt Trennkraft besitzt. Oder grundlegender: Wie groß ist die Wahrscheinlichkeit, daß sich für eine Stichprobe ein bestimmter Diskriminanzgrad errechnen läßt, obwohl überhaupt keine Gruppenunterschiede vorhanden sind?

Der gebräuchlichste Test zur Prüfung der Signifikanz der Diskriminanzfunktionen (Signifikanztest der Unterschiede zwischen den Gruppen) geht indirekt vor. Man nimmt an, die Diskriminanzfunktionen werden sequentiell hergeleitet, und zwar nach ihrer diskriminatorischen Bedeutung, zuerst also die mit dem höchsten Eigenwert, dann die mit dem zweithöchsten, usw. Dazu definiert man das sog. **Wilks' Lambda**[2]

$$\Lambda_k = \prod_{i=k+1}^{q} \frac{1}{1+\gamma_i} = \frac{1}{1+\gamma_{k+1}} \cdot \frac{1}{1+\gamma_{k+1}} \cdot \ldots \cdot \frac{1}{1+\gamma_q}, \tag{10}$$

wobei die γ_i die Eigenwerte der Diskriminanzfunktionen sind und k die Anzahl der Funktionen, die bereits hergeleitet sind (im ersten Schritt ist $k = 0$).

Der Test untersucht vor Erstellung der k-ten Diskriminanzfunktion, ob überhaupt noch genügend Information in den Merkmalsvariablen zur Unterscheidung der Gruppen vorhanden ist, nachdem die bereits errechneten Funktionen schon einen Teil dieser Information ausgeschöpft haben. Falls dies nicht der Fall ist, sind die k-te und alle folgenden Diskriminanzfunktionen nicht mehr signifikant.

Der Vorteil der Größe Λ_k besteht darin, daß sie durch die Transformation

$$\chi^2 = -\left(n - \frac{p+g}{2} - 1\right) \cdot \ln \Lambda_k \tag{11}$$

in eine Testgröße χ_0^2 umgewandelt werden kann, die unter Berücksichtigung der Voraussetzungen unter Abschnitt 3 χ^2-verteilt ist mit $(p-k) \cdot (g-k-1)$ Freiheitsgraden.

Der Signifikanztest über Wilks' Lambda gibt nur eine Aussage über die mit den hergeleiteten Diskriminanzfunktionen erzielte Trennung **insgesamt**, bzw. über die relative Diskriminanzkraft der einzelnen Funktionen. Er beinhaltet auch den Test der Nullhypothese, daß sich die Gruppen nicht unterscheiden. Es wird jedoch **keine** Aussage darüber gemacht, wie gut die Funktionen die einzelnen Gruppen trennen.

[2]Dieses Wilks' Lambda stellt ein Maß für die Unterschiedlichkeit der Gruppen dar. Je geringer sein Wert ist, desto homogener sind die Gruppen und desto größer ist der Unterschied zwischen den Gruppen.

Beispiel:

Im voherigen Beispiel existierte nur ein Eigenwert $\gamma = 1.628$. *Es gibt daher auch nur ein Wilk's Lambda zur Signifikanzprüfung der Diskriminanzfunktion.*

$$\Lambda_k = \prod_{i=k+1}^{q} \frac{1}{1+\gamma_i} \;\Rightarrow\; \Lambda_0 = \prod_{i=1}^{1} \frac{1}{1+\gamma_i} = \frac{1}{1+\gamma_1} = \frac{1}{1+\gamma} = \frac{1}{1+1.628} = 0.381$$

Die χ^2*-Testgröße ist dann*

$$\chi_0^2 = -\left(n - \frac{p+g}{2} - 1\right) \cdot \ln \Lambda_k = -\left(50 - \frac{2+2}{2} - 1\right) \cdot \ln 0.381 = 45.353$$

bei $(p-k)\cdot(g-k-1) = (2-0)\cdot(2-0-1) = 2$ *Freiheitsgraden.*

Es ist $\chi_0^2 = 45.353 > 9.210 = \chi^2_{2;0,99}$. *Damit kann eine hochsignifikante* ($\alpha = 1\%$) *diskriminisatorische Bedeutung der Diskriminanzfunktion gesichert werden. Der p-Wert ist fast* 0.

4.5 Prüfung der Merkmalsvariablen

Neben den im letzten Abschnitt untersuchten Fragen (Signifikanz der Gruppenunterschiede, relative Diskriminanzkraft der einzelnen Diskriminanzfunktionen) ist auch die Frage von Interesse, welche Bedeutung die Merkmalsvariablen für eine bestimmte Diskriminanzfunktion besitzen. Dies läßt sich prinzipiell an den Diskriminanzkoeffizienten b_i ablesen. Da die Merkmalsvariablen i.a. auf verschiedenen Skalen gemessen sind, verwendet man zur Bestimmung der Bedeutung der Merkmalsvariablen die **standardisierten Diskriminanzkoeffizienten** b_i^*

$$b_i^* = b_i \cdot s_i, \tag{12}$$

wobei s_i die Standardabweichung der i-ten Merkmalsvariable ist. Die Variablen mit den betragsmäßig größeren standardisierten Diskriminanzkoeffizieneten haben größere Bedeutung für die Unterscheidung der Gruppen.

Beispiel:

Im Beispiel der Gruppierung von Soja- und Maisöl anhand ihres Fettsäuremusters waren die Diskriminanzkoeffizienten $b_1 = 0.3835$ *(Ölsäure* x_1*) und* $b_2 = 0.1216$ *(Linolsäure* x_2*). Die Standardabweichungen in den Merkmalsvariablen sind* $s_1 = 4.7820$ *und* $s_2 = 2.7918$. *Damit folgt für die standardisierten Diskriminanzkoeffizienten* $b_1^* = b_1 \cdot s_1 = 0.3835 \cdot 4.7815 = 1.8337$ *und* $b_2^* = b_2 \cdot s_2 = 0.1216 \cdot 2.7918 = 0.3395$.

5 Klassifikation

Für die Klassifikation von neuen Elementen existieren grundsätzlich drei verschiedene Konzepte: Klassifizierung mittels Distanzen, mittels Klassifizierungsfunktionen und über Wahrscheinlichkeitsberechnungen.

In diesem Abschnitt wird davon ausgegangen, daß q Diskriminanzfunktionen verwendet werden. Folgende Bezeichnungen sollen gelten:

- $a = (a_1, a_2, \ldots, a_p)'$ ist der Vektor der gemessenen oder erhobenen Merkmalsausprägungen $a_1, a_2, \ldots, a_p$ eines neuen Untersuchungsobjekts.
- $y^{(l)}(x) = y^{(l)}\Big((x_1, x_2, \ldots, x_p)\Big) = b_0^{(l)} + b_1^{(l)}\, x_1 + b_1^{(l)}\, x_1 \ldots + b_p^{(l)}\, x_p$ ist die l-te Diskriminanzfunktion mit den zugehörigen Diskriminanzkoeffizienten $b_i^{(l)}$ und $y^{(l)}(a)$ der zugehörige Diskriminanzwert des Objekts a.
- $\overline{x}_{ij.}$ ist der empirische Mittelwert der i-ten Merkmalsvariablen über die Elemente der j-ten Gruppe.
- $\overline{x}_j = (\overline{x}_{1j.}, \overline{x}_{2j.} \ldots, \overline{x}_{pj.})$ ist der Mittelpunkt oder **Centroid** der j-ten Gruppe im Merkmalsraum.
- $\overline{y}_j^{(l)} = y^{(l)}(\overline{x}_j)$ ist der Diskriminanzwert des Centroids der j-ten Gruppe bezüglich der l-ten Diskriminanzfunktion.

Das zugrundeliegende Klassifikationsprinzip ist einsichtig. Ein Objekt a wird anhand seiner gemessenen bzw. erhobenen Merkmalsausprägungen $a_1, a_2, \ldots, a_p$ in die Gruppe klassifiziert, zu deren Gruppenmittelpunkt (Centroid) es den kleinsten Abstand hat. Existiert lediglich eine Diskriminanzfunktion, so wird ein Objekt derjenigen Gruppe zugeordnet, für die die euklidische Distanz

$$|y(a) - \overline{y}_j| = \sqrt{y(a)^2 - \overline{y}_j^2} \tag{13}$$

minimal ist.

Bei nur 2 Gruppen und gleicher Gruppengröße kann man einen **kritischen Distanzwert** y^* als Mittelwert der beiden Gruppencentroide bestimmen:

$$y^* = \frac{\overline{y}_1 + \overline{y}_2}{2} \tag{14}$$

Durch die Normierung ist $y^* = 0$ und eine Objekt a wird in die Gruppe 1 klassifiziert, wenn $y(a) < y^* = 0$ und in die Gruppe 2, wenn $y(a) > y^* = 0$. Im Falle $y(a) = y^* = 0$ ist keine Klassifizierung möglich.

Beispiel:

Eine Speiseölprobe mit $a_1 = 30\%$ Ölsäure- und $a_2 = 50\%$ Linolsäuregehalt ($a = (30, 50)$) soll in die Gruppen Sojaöl (G_1) oder Maisöl (G_2) klassifiziert werden. Die Diskriminanzwerte der Centroide sind $\overline{y}_1 = -1.25$ und $\overline{y}_2 = 1.25$, also $y^* = 0$. Aufgrund der Diskriminanzfunktion $y = -18.1492 + 0.3835\,x1 + 0.1216\,x_2$ ergibt sich ein $y(a) = -0.5642 < 0 = y^*$, sodaß eine Gruppierung als Sojaöl erfolgt. Auch mit Hilfe des euklidischen Abstand erfolgt diese Gruppierung, denn $|-0.5642 - (-1.25)| = 0.6858 < 1.8142 = |-0.5642 - 1.25|$.

5.1 Distanzkonzept und Klassifizierungsfunktionen

Man kann im allgemeinen Fall das Klassifikationsziel auf zweierlei Arten verfolgen: Entweder untersucht man die euklidischen Abstände im q-dimensionalen Raum der Diskriminanzvariablen y_i oder die sog. **Mahalanobis-Distanzen** im p-dimensionalen Raum der Mermalsvariablen x_i.

Im ersten Fall benötigt man die q kanonischen Diskriminanzfunktionen und definiert den quadratischen Abstand[3] $DY_j^2(a)$ des zu klassifierenden Objekts a zur Gruppe j.

$$DY_j^2(a) = \sum_{l=1}^{q} \left(y^{(l)}(a) - \overline{y}_j^{(l)} \right)^2, \quad (j = 1, 2, \ldots, g) \tag{15}$$

Die Klassifizierungsregel lautet also:

- Ordne a der Gruppe j zu, für die $DY_j^2(a)$ am kleinsten ist.

Beispiel:

Im vorherigen Beispiel wurden die euklidischen Distanzen zu den Gruppencentroiden berechnet. Die quadratischen Distanzen sind:

$DY_1^2 = (-0.5642 - (-1.25))^2 = 0.4703 = 0.6858^2$

$DY_2^2 = (-0.5642 - 1.25)^2 = 3.2913 = 1.8142^2$

Das Klassifikationsergebnis ist natürlich wie erwartet gleich dem obigen Ergebnis, d.h. die Probe wird als Sojaöl klassifiziert.

Im zweiten Fall ist das Abstandsmaß definiert als:

$$DX_j^2(a) = (a - \overline{x}_j)'\, S^{-1}\, (a - \overline{x}_j) \tag{16}$$

Die Matrix S stellt in dieser Formel eine Kovarianzmatrix der Merkmalsvariablen dar. Für die Wahl von S gibt es zwei Möglichkeiten.

[3] Das Weglassen der Wurzel hat keinen Einfluß auf das Minimierungskriterium.

Falls die Streuungen in den Gruppen als gleich betrachtet werden können, verwendet man für S die gepoolte Innergruppen-Kovarianzmatrix S_{p}, deren (r,c)-tes Element sich als

$$s_{r,c} = \sum_{i=1}^{g} \sum_{j=1}^{n_j} (x_{rij} - \overline{x}_{ri.}) \cdot (x_{cij} - \overline{x}_{ci.}) \tag{17}$$

berechnet. In diesem Fall läßt sich das Abstandsmaß (16) umformen zu

$$\overline{DX}_j^2(a) = -2 \left(\overline{x}_j' \, S_{\mathrm{p}}^{-1} \, a - 0.5 \, \overline{x}_j' \, S_{\mathrm{p}}^{-1} \overline{x}_j \right) + a' \, S_{\mathrm{p}}^{-1} a \tag{18}$$

Die p in a linearen Funktionen innerhalb der Klammern heißen **lineare Diskriminanzfunktionen** bzw. **lineare Klassifizierungsfunktionen**. Da $a' \, \overline{C}^{-1} a$ unabhängig von der Gruppe ist, sind folgende Entscheidungsregeln äquivalent:

- Ordne a der Gruppe j zu, für die $\overline{DX}_j^2(a)$ am kleinsten ist (Entscheidung mit Distanzen).
- Ordne a der Gruppe j zu, für die die lineare Diskriminanzfunktion am größten ist (Entscheidung mit linearen Klassifizierungsfunktionen).

Im Falle ungleicher Streuungen bestimmt man das Abstandsmaß mit der Kovarianzmatrix S_j der Merkmalsvariablen in der jeweiligen Gruppe j:

$$DX_j^2(a) = (a - \overline{x}_j)' \, S_j^{-1} \, (a - \overline{x}_j). \tag{19}$$

Dieses Maß läßt sich nicht mehr auf eine lineare Funktionen zurückführen. Man spricht deshalb von p **quadratischen Diskriminanzfunktionen** bzw. **quadratischen Klassifizierungsfunktionen**.

Beispiel:

Bei zwei Merkmalsvariablen ist die quadrierte Mahalanobis-Distanz:

$$DX_j^2(a) = \frac{(a_1 - \overline{x}_{1j})^2 \, s_2^2 + (a_2 - \overline{x}_{2j})^2 \, s_1^2 - 2 \, (a_1 - \overline{x}_{1j})(a_2 - \overline{x}_{2j}) \, s_{12}}{s_1^2 \, s_2^2 - s_{12}}$$

Dabei sind s_1^2 und s_2^2 die Varianzen und s_{12} die Kovarianz der beiden Variablen. Im Beispiel der Gruppierung von Soja- und Maisöl anhand ihres Fettsäuremusters sind die Varianzen $s_1^2 = 22.9$ und $s_2^2 = 7.8$, die Kovarianz ist $s_{12} = -9.7$, die Centroide $\overline{x}_1 = (26.8, 54.4)$ und $\overline{x}_2 = (34.2, 51.7)$. Damit folgen für die Mahalanobis-Distanzen der Probe $a = (30, 50)$ für $DX_1^2(a)$

$$\frac{(30 - 26.8)^2 \cdot 7.8 + (50 - 54.4)^2 \cdot 22.9 + 2 \cdot (30 - 26.8) \cdot (50 - 54.4) \cdot 9.7}{22.9 \cdot 7.8 + 9.7} = 1.3$$

und für $DX_2^2(a)$

$$\frac{(30 - 34.2)^2 \cdot 7.8 + (50 - 51.7)^2 \cdot 22.9 + 2 \cdot (30 - 34.2) \cdot (50 - 51.7) \cdot 9.7}{22.9 \cdot 7.8 + 9.7} = 1.8.$$

Die Gruppierung erfolgt also wie erwartet in Gruppe G_1 (Sojaöl).

5.2 Wahrscheinlichkeitskonzept

Das Wahrscheinlichkeitskonzept ermöglicht es, sog. **A-priori-Wahrscheinlichkeiten**, also Vorwissen über die Klassenzugehörigkeiten) mit in die Diskriminanzanalyse einzubeziehen. Es bezeichne

- $P_a(j)$ die gegebene oder geschätzte Wahrscheinlichkeit der Zugehörigkeit eines Objekts a zur j-ten Gruppe **vor** Berechnung der Diskriminanzfunktionen (A-priori-Wahrscheinlichkeit),
- $P(y(a)|j)$ die Wahrscheinlichkeit, daß das Objekt a den Diskriminanzwert $y(a)$ hat bei gegebener Gruppenzugehörigkeit zur j-ten Gruppe (bedingte Wahrscheinlichkeit),
- $P(j|y(a))$ die Wahrscheinlichkeit, daß das Objekt a zu Gruppe j gehört bei gegebenem Diskriminanzwert $y(a)$ (A-posteriori-Wahrscheinlichkeit).

Mit den A-priori-Wahrscheinlichkeiten läßt sich z.B. berücksichtigen, daß die betrachteten Gruppen in der Realität mit unterschiedlicher Häufigkeit vorkommen; sie müssen sich über die Gruppen zu 1 addieren:

$$\sum_{j=1}^{g} P_a(j) = 1 \tag{20}$$

Die Klassifizierungsregel bei diesem Konzept lautet:

- Ordne a der Gruppe j zu, für die die A-posteriori-Wahrscheinlichkeit $P(j|y(a))$ maximal ist.

Die A-posteriori-Wahrscheinlichkeit erhält man über den Satz von Bayes:

$$P(j|y(a)) = \frac{P(y(a)|j) \cdot P_a(j)}{\sum_{k=1}^{g} P(y(a)|k) \cdot P_a(k)} \tag{21}$$

Die noch fehlenden bedingten Wahrscheinlichkeiten lassen sich aus den Distanzen $DY_j^2(a)$ durch Transformation ermitteln. Die A-posteriori-Wahrscheinlichkeiten lauten dann:

$$P(j|y(a)) = \frac{\mathrm{e}^{-DY_j^2(a)/2} \cdot P_a(j)}{\sum_{k=1}^{g} \left(\mathrm{e}^{-DY_j^2(a)/2} \cdot P_a(k)\right)} \tag{22}$$

Falls $P(j|y(a))$ maximal ist, muß dies auch

$$\ln(P(j|y(a))) = \frac{-DY_j^2(a)}{2} + \ln P_a(j) - \ln \sum_{k=1}^{g} \left(\mathrm{e}^{-DY_j^2(a)/2} \cdot P_a(k)\right) \tag{23}$$

sein, und wegen der Unabhängigkeit von $\sum_{k=1}^{g} \mathrm{e}^{-DY_j^2(a)/2 \cdot P_a(k)}$ von j kann man als einfachere Klassifikationsregel folgendes aufstellen:

- Ordne a der Gruppe j zu, für die $-0.5 \left(DY_j^2(a) - 2 \ln P_a(j) \right)$ maximal ist.

Beispiel:

Bei der Auswertung von Satellitenaufnahmen zur Klassifikation von Feldfrüchten kann zur Unterscheidung von Weizen und Gerste der bekannte Flächenanteil als A-priori-Wahrscheinlichkeit in die Diskriminanzanalyse mit einbezogen werden.

5.3 Prüfung der Klassifikation

Die Prüfung der Klassifikation kann prinzipiell auf drei Arten erfolgen.

Prüfung mit den Ausgangsdaten

Nach Erstellen der Diskriminanzfunktionen werden alle Objekte mit diesen Funktionen klassifiziert. Da man für diese Objekte die wahre Gruppenzugehörigkeit kennt, kann man in Form einer **Klassifikationsmatrix** die Anzahl der richtig und falsch klassifizierten Objekte jeder Gruppe ausgeben:

	Wahre Gruppenzugehörigkeit					
klassifiziert in	Gruppe 1	$\cdots$	Gruppe j	$\cdots$	Gruppe g	
Gruppe 1	m_{11}	$\cdots$	m_{1j}	$\cdots$	m_{1g}	
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
Gruppe j	m_{j1}	$\cdots$	m_{jj}	$\cdots$	m_{jg}	
$\vdots$	$\vdots$		$\vdots$		$\vdots$	
Gruppe g	m_{g1}	$\cdots$	m_{gj}	$\cdots$	m_{gg}	
Gesamt	n_1	$\cdots$	n_j	$\cdots$	n_g	n

Das heißt m_{jj} Objekte von den n_j Objekten aus Gruppe j sind richtig in Gruppe j klassifiziert worden. Der Rest aus Gruppe j ist mit den erstellten Diskriminanzfunktionen fehlklassifiziert worden, z.B. m_{1j} Objekte in die Gruppe 1. Relativ gesehen sind also m_{jj}/n_j aus Gruppe j richtig klassifiziert worden.

Wichtig ist, daß hier die Objekte zur Prüfung der Klassifikation verwendet werden, die in die Erstellung der Diskriminanz- bzw. Klassifikationsfunktionen mit eingeflossen sind.

Prüfung mit Kreuzvalidierung

Bei der Diskriminanzanalyse mit Kreuzvalidierung werden n Sätze von Diskriminanz- bzw. Klassifizierungsfunktionen erstellt. Im ersten Durchgang wird das erste Objekte aus den Daten entfernt, die Diskriminanzfunktionen aus den restlichen Daten erstellt, und mit diesen Funktionen dann die erste Beobachtung klassifiziert. Im zweiten Schritt wird die erste Beobachtung wieder in die Daten aufgenommen, die zweite Beobachtung aus den Daten entfernt, aus den restlichen Daten die Diskriminanzfunktionen erstellt und die zweite Beobachtung klassifiziert. Dies wird solange fortgesetzt, bis jede Beobachtung einmal weggelassen worden ist. Wiederum kann man die Anzahl der richtig und der falsch klassifizierten Objekte in Form einer Klassifikationsmatrix ausgeben.

Da die Objekte, die zur Kontrolle klassifiziert werden, nicht an der Erstellung der Diskriminanzfunktionen beteiligt sind, steht in der Regel zu erwarten, daß mit Kreuzvalidierung weniger Objekte richtig klassifiziert werden als ohne Kreuzvalidierung.

Prüfung mit Kontrolldaten

Falls in einer Untersuchung genügend Objekte vorhanden sind, kann man diese aufteilen in eine Menge von Objekten, mit denen die Diskriminanz- bzw. Klassifikationsfunktionen erstellt werden, und eine zweite Menge (Kontrolldaten) die mit diesen Funktionen dann klassifiziert werden. Da von den Kontrolldaten die wahren Gruppenzugehörigkeiten wiederum bekannt sind, kann auch in diesem Fall die Anzahl der richtig und falsch klassifizierten Objekte in Form der Klassifikationsmatrix ausgegeben werden.

Hier ist es nun tatsächlich so, daß die Kontrolldaten nichts mit der Erstellung der Diskriminanzfunktionen zu tun haben. Dies ist der idealste der drei Fälle. Es müssen aber genügend Objekte vorhanden sein, damit eine Aufteilung möglich ist.

Beispiel:

Mit der Diskriminanzfunktion $y = -18.1492 + 0.3835\,x_1 + 0.1216\,x_2$ *werden die Proben* 3, 4, 6, 7 *und* 14 *der Gruppe Sojaöl aus Tab. 1 in die Gruppe Maisöl sowie die Probe* 3 *aus der Gruppe Maisöl in die Gruppe Sojaöl klassifiziert. Dies kann man leicht durch Einsetzen der Fettsäuregehalte in die Diskriminanzfunktion zeigen, z.B. ist für die Probe* 3 *von Sojaöl der Diskriminazwert*

$y = -18.1492 + 0.3835 \cdot 30.8 + 0.1216 \cdot 52.9 = 0.0952 \;\Rightarrow\;$ *Gruppe* G_2 *(Maisöl)*

und für die Probe 3 *von Maisöl*

$y = -18.1492 + 0.3835 \cdot 29.3 + 0.1216 \cdot 54.0 = -0.3463 \;\Rightarrow\;$ *Gruppe* G_1 *(Sojaöl).*

Die Prüfung mit den Ausgangsdaten liefert also folgende Klassifikationsmatrix:

	Wahre Gruppenzugehörigkeit		
klassifiziert in	*Gruppe* 1	*Gruppe* 2	
Gruppe 1	20	5	
Gruppe 2	1	24	
Gesamt	25	25	50
Korrekt	80%	96%	88%

Die Prüfung mit Kreuzvalidierung liefert in diesem Fall die gleiche Klassifikationsmatrix. Eine Prüfung mit Kontrolldaten ist aufgrund des geringen Stichprobenumfangs nicht sinnvoll.

6 Diskriminanzanalyse in MINITAB

Die Routine `DISCRIMINANT` von MINITAB führt lineare oder quadratische Diskriminanzanalysen nur über Klassifizierung durch, so daß weder die kanonischen Diskriminanzfunktionen errechnet werden können (siehe Abschnitt 4.3), noch eine Prüfung der Diskriminanzfunktionen (siehe Abschnitt 4.4) bzw. der Merkmalsvariablen (siehe Abschnitt 4.5) durchgeführt werden kann.

Die allgemeine Syntax auf Kommandozeilenebene lautet:

```
DISCRIMINANT Gruppen in C, Merkmale in C,...,C
   QUADRATIC
   PRIORS      K,...,K
   LDF         C,...,C
   FITS        C, [C]
   XVAL
   PREDICT     E,...,E
   BRIEF       K
```

Im `DISCRIMINANT`-Befehl muß als erster Parameter die Spalte mit den Gruppencodes angegeben werden, danach die Spalten mit den Merkmalsvariablen.

Soll eine quadratische Diskriminanzanalyse bei unterschiedlichen Streuungen in den Gruppen durchgeführt werden, dann ist das Subkommando `QUADRATIC` anzugeben. Es kann nicht gleichzeitig mit `LDF` verwendet werden.

Mit dem `PRIORS`-Subkommando kann man eigene A-priori-Wahrscheinlichkeiten für die Gruppen eingegeben werden. Die Anzahl der Wahrscheinlichkeiten muß gleich der Anzahl der Gruppen sein.

Das Subkommando `LDF` (*linear discriminant function*) speichert die Koeffizienten der linearen Diskriminanzfunktionen[4] in den Spalten `C,...,C`. `LDF` kann nicht gleichzeitig mit `QUADRATIC` verwendet werden.

`FITS` speichert die für jede Beobachtung geschätzte Gruppe in Spalte `C`. In der optionalen zweiten Spalte wird die für jede Beobachtung durch Kreuzvalidierung ermittelte Gruppe gespeichert, falls `XVAL` angegeben wurde.

`XVAL` führt eine Kreuzvalidierung durch.

Mit dem `PREDICT`-Subkommando kann für neue Objekte mit den Merkmalen `E,...,E` aufgrund der Klassifikationsfunktion ermittelt werden. Die Anzahl der Spalten oder Konstanten `E` muß gleich der Anzahl der Merkmale sein.

`BRIEF` bestimmt den Umfang des MINITAB-Outputs. Der Parameter `K` geht von `1` bis `4`.

Beispiel:

Es soll der Datensatz aus Tab. 1 analysiert sowie ein Objekt mit 30% *Olsäure und* 50% *Linolsäure klassifiziert werden.*

Das Einlesen der Daten erfolgt aus einer ASCII-Datei `PFLOEL.DAT`. *In Spalte* `C1` *stehen die Codes für die Gruppen (*`1`*: Sojaöl,* `2`*: Maisöl), in Spalte* `C2` *der Gehalt an Ölsäure und in* `C3` *der Gehalt an Linolsäure.*

```
MTB > name c1 'PflOel' c2 'Oel' c3 'Linol'
MTB > Read "PFLOEL.DAT" 'PflOel'-'Linol'.
Entering data from file: PFLOEL.DAT
     50 rows read.
MTB > Print 'PflOel'-'Linol'.

Data Display

 Row  PflOel    Oel  Linol

   1       1   25.4   53.6
   2       1   25.8   50.6
   :       :    :      :
   :       :    :      :
  25       1   24.1   56.6
  26       2   30.3   55.9
  27       2   32.4   50.1
   :       :    :      :
   :       :    :      :
  50       2   37.3   51.3
```

[4]Die linearen Diskriminanzfunktionen für jede Gruppe sind nicht zu verwechseln mit der kanonischen Diskriminanzfunktion

Die lineare Diskriminanzanalyse mit anschließender Klassifikation liefert folgenden Output[5]*:*

```
MTB > Discriminant 'PflOel' 'Oel' 'Linol';
SUBC>   Predict 30 50.

Discriminant Analysis

Summary of Classification

Put into       ....True Group....
Group              1        2
1                 20        1
2                  5       24
Total N           25       25
N Correct         20       24
Proportion     0.800    0.960

N =   50     N Correct =   44     Proportion Correct = 0.880

Summary of Misclassified Observations

Observation     True     Pred
               Group    Group
    3 **           1        2
    4 **           1        2
    6 **           1        2
    7 **           1        2
   14 **           1        2
   28 **           2        1

Prediction for Test Observations

 Observation Pred  Group
           1           1
```

Die Klassifikationsmatrix gibt die Anzahl der richtig und falsch klassifizierten Objekte für die beiden Gruppen aus. In Gruppe 1 (Sojaöl) wurden von 25 *Objekten* 20 *richtig und* 5 *falsch klassifiziert. In Gruppe (Maisaöl) wurden von* 25 *Objekten* 24 *richtig und eines falsch klassifiziert. Dies entspricht jeweils einem Anteil von*

[5]Im Output werden zunächst nur die wesentlichen Teile ausgegeben. Die anderen Teile wurden entfernt.

80% *bzw.* 96%. *Die Gesamtzahl an richtig klassifizierten Objekten beträgt* 44 *von* 50. *Dies entspricht* 88%. *Die anschließende Tabelle zeigt eine Liste der falsch klassifizierten Objekte. Es wurden also die Objekte* 3, 4, 6, 7 *und* 14 *aus Gruppe* 1 *in Gruppe* 2 *sowie das Objekt* 28 *aus Gruppe* 2 *in Gruppe* 1 *fehlklassifiziert. Die Klassifizierung des neuen Objekts* $(30, 50)$ *erfolgt in die Gruppe Sojaöl* (1).

Im folgenden abschließenden Beispiel soll noch ein vollständiger MINITAB-Output für einen Standardbeispielsdatensatz aus der statistischen Literatur angegeben werden. Es handelt sich um die von R.A. Fisher 1936 veröffentlichten Iris-Daten mit 3 Gruppen und 4 Merkmalen.

Beispiel:

Bei drei Irisarten (Iris setosa, Iris versicolor und Iris virginica) wurden an jeweils 50 *Exemplaren die Breite und Länge der Kelchblätter (engl. sepal) und Blütenbläter (engl. petal) in Millimetern gemessen. Die Merkmalsvariablen sind im folgenden mit* `SepalLen`, `SepalWid`, `PetalLen` *und* `PetalWid`, *die Gruppenvariable mit* `Species` *bezeichnet. Die Daten werden aus einer ASCII-Datei* `IRIS.DAT` *eingelesen und eine quadratische Diskriminanzanalyse durchgeführt. Beim Subkommando* `BRIEF 3` *erfolgt zusätzlich die Ausgabe der Kovarianzmatrizen für die einzelnen Gruppen.*

```
MTB > name c1 'SepalLen' c2 'SepalWid' c3 'PetalLen' c4 'PetalWid'
MTB > name c5 'Species'
MTB > Read "IRIS.DAT" 'SepalLen'-'Species'.
Entering data from file: IRIS.DAT
    150 rows read.
```

Das MINITAB-Worksheet hat nun folgende Gestalt:

```
MTB > Print 'SepalLen'-'Species'.

Data Display


 Row  SepalLen  SepalWid  PetalLen  PetalWid  Species

   1        50        33        14         2        1
   2        64        28        56        22        3
   3        65        28        46        15        2
   4        67        31        56        24        3
   5        63        28        51        15        3
   :         :         :         :         :        :
   :         :         :         :         :        :
 150        53        37        15         2        1
```

```
MTB > Discriminant 'Species' 'SepalLen'-'PetalWid';
SUBC>   Quadratic;
SUBC>   XVal;
SUBC>   Brief 3.

Discriminant Analysis

Quadratic Method for Response:   Species
Predictors:  SepalLen  SepalWid  PetalLen  PetalWid

Group        1        2        3
Count       50       50       50

Summary of Classification

Put into       ....True Group....
Group              1        2        3
1                 50        0        0
2                  0       48        1
3                  0        2       49
Total N           50       50       50
N Correct         50       48       49
Proportion     1.000    0.960    0.980

N =  150     N Correct =  147     Proportion Correct = 0.980

Summary of Classification with Cross-validation

Put into       ....True Group....
Group              1        2        3
1                 50        0        0
2                  0       47        1
3                  0        3       49
Total N           50       50       50
N Correct         50       47       49
Proportion     1.000    0.940    0.980

N =  150     N Correct =  146     Proportion Correct = 0.973
```

From Group	Generalized Squared Distance to Group 1	2	3
1	5.353	110.740	178.261
2	328.415	7.546	23.332
3	711.438	25.413	9.494

Variable	Pooled Mean	Means for Group 1	2	3
SepalLen	58.433	50.060	59.360	65.880
SepalWid	30.573	34.280	27.700	29.740
PetalLen	37.580	14.620	42.600	55.520
PetalWid	11.993	2.460	13.260	20.260

Variable	Pooled StDev	StDev for Group 1	2	3
SepalLen	5.148	3.525	5.162	6.359
SepalWid	3.397	3.791	3.138	3.225
PetalLen	4.303	1.737	4.699	5.519
PetalWid	2.047	1.054	1.978	2.747

Pooled Covariance Matrix

	SepalLen	SepalWid	PetalLen	PetalWid
SepalLen	26.501			
SepalWid	9.272	11.539		
PetalLen	16.751	5.524	18.519	
PetalWid	3.840	3.271	4.267	4.188

Covariance Matrix for Group 1

	SepalLen	SepalWid	PetalLen	PetalWid
SepalLen	12.4249			
SepalWid	9.9216	14.3690		
PetalLen	1.6355	1.1698	3.0159	
PetalWid	1.0331	0.9298	0.6069	1.1106

Covariance Matrix for Group 2

	SepalLen	SepalWid	PetalLen	PetalWid
SepalLen	26.6433			
SepalWid	8.5184	9.8469		
PetalLen	18.2898	8.2653	22.0816	
PetalWid	5.5780	4.1204	7.3102	3.9106

```
Covariance Matrix for Group 3
         SepalLen SepalWid PetalLen PetalWid
SepalLen  40.4343
SepalWid   9.3763  10.4004
PetalLen  30.3290   7.1380  30.4588
PetalWid   4.9094   4.7629   4.8824   7.5433

Summary of Misclassified Observations

Observation  True  Pred X-val Group  Squared Dist.   Probab.
             Group Group Group          Pred   X-val  Pred X-val
    5 **         3     2     2      1  520.06  520.06  0.00  0.00
                                    2   12.93   12.93  0.60  0.66
                                    3   13.78   14.28  0.40  0.34
    8 **         2     2     3      1  431.45  431.45  0.00  0.00
                                    2   20.04   24.54  0.81  0.31
                                    3   22.98   22.98  0.19  0.69
    9 **         2     3     3      1  488.11  488.11  0.00  0.00
                                    2   16.06   17.99  0.34  0.16
                                    3   14.70   14.70  0.66  0.84
   12 **         2     3     3      1  534.06  534.06  0.00  0.00
                                    2   15.64   17.37  0.15  0.07
                                    3   12.23   12.23  0.85  0.93
```

Die Klassifikationsmatrizen ohne und mit Kreuzvalidierung sind in diesem Fall nicht identisch. Mit Kreuzvalidierung wird ein zusätzliches Objekt aus Gruppe 2 in Gruppe 3 fehlklassifiziert. Ebenfalls als Matrix werden die quadrierten Abstände zwischen den Gruppencentroiden ausgegeben. Danach folgen die Mittelwerte und Standardabweichungen jeweils für alle Gruppen (gepoolt) und für jede einzelne Gruppe. Ein Vergleich der gepoolten Kovarianzmatrix mit den Kovarianzmatrizen der einzelnen Gruppen zeigt, daß es durchaus gerechtfertigt ist, ungleiche Varianzen zwischen den Gruppen anzunehmen und die quadratische Diskriminanzanalyse zu wählen. Die Zusammenfassung am Ende des Outputs enthält die Reihennummer und wahre Gruppenzugehörigkeit der fehlklassifizierten Objekte sowie jeweils getrennt für ohne (`Pred`*) und mit (*`X-val`*) Kreuzvalidierung die Klassifikationsgruppe, die generalisierten quadrierten Distanzen zum Gruppencentroid*[6] *und in den letzten beiden Spalten die A-posteriori-Wahrscheinlichkeit, also die Wahrscheinlichkeit, daß die Beobachtung bei gegebenem Diskriminanzwert zu Gruppe j gehört. Bei Beobachtung 5 beträgt die Wahrscheinlich-*

[6] Aufgrund der unterschiedlichen Kovarianzmatrizen S_i in Gleichung (19) ist die Distanzmatrix nicht mehr symmetrisch. Die Distanzen zwischen gleichen Gruppen sind verschieden von 0, da im quadratischen Fall zu den Distanzen der natürliche Logarithmus der Determinante der Kovarianzmatrix addiert wird, also: $DX_j^2(a) = (a - \overline{x}_j)' \, S_j^{-1} \, (a - \overline{x}_j) + \ln |S_j|$

keit für Gruppe 1 0%, *die Wahrscheinlichkeit für Gruppe* 2 *ohne Kreuzvalidierung* 60%, *mit Kreuzvalidierung* 66%, *und die Wahrscheinlichkeit für Gruppe* 3 *ohne Kreuzvalidierung* 40%, *mit Kreuzvalidierung* 34%.

Bild 6 zeigt abschließend den Matrixplot der Irisdaten. Eine eindeutige Gruppe bildet Iris setosa, was auch dem Klassifikationsergebnis von 100% *richtig klassifizierten Objekten in dieser Gruppe entspricht.*

```
MTB > MatrixPlot 'SepalLen'-'PetalWid';
SUBC>   Symbol 'Species'.
```

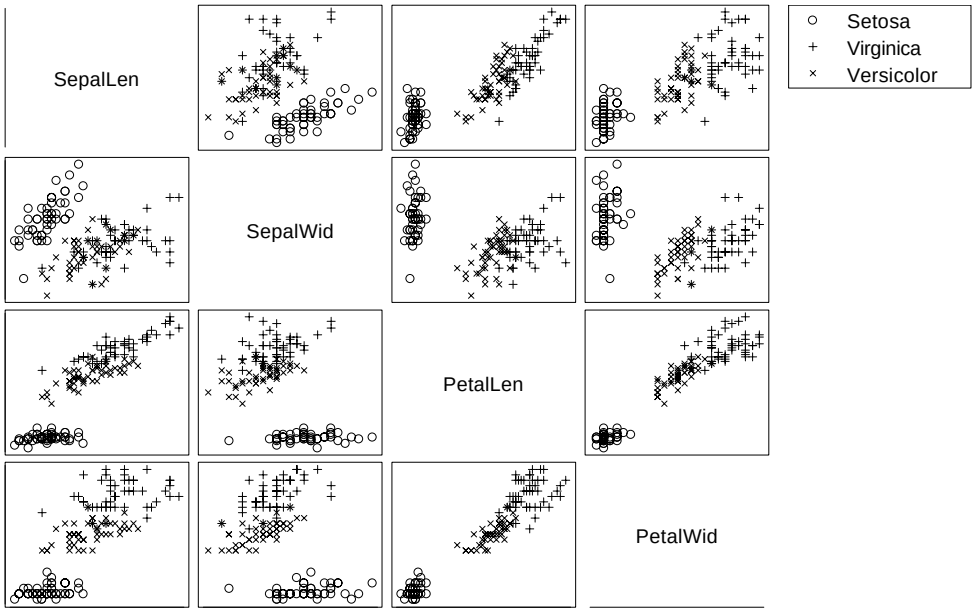

Bild 6: Matrixplot der Irisdaten

7 Literatur

Backhaus K., Erichson B., Plinke W., Schuchard-Ficher C., Weiber R. 1996: Multivariate Analysemethoden: Eine anwendungsorientierte Einführung. 8. Auflage, Springer.

Johnson R., Wichern D. 1982: Applied Multivariate Statistical Methods. 2. edition, Prentice Hall.

Klecka W.R. 1980: Discriminant Analysis. Sage publications.

Minitab Reference Manual.

Morrison D. 1982: Multivariate Statistical Methods. McGraw Hill.

SAS/STAT User's Guide.