TECHNISCHE UNIVERSITÄT MÜNCHEN-WEIHENSTEPHAN
MATHEMATIK UND STATISTIK
INFORMATIONS- UND DOKUMENTATIONSZENTRUM R. KRAFT

Biometrische und Ökonometrische Methoden II SS 00

Diskriminanzanalyse

Diskriminanzanalyse

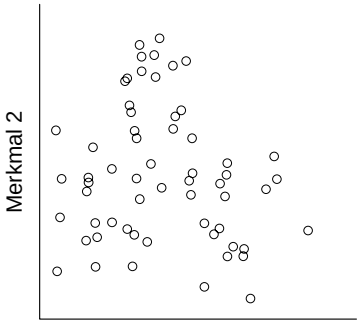
Fragestellung

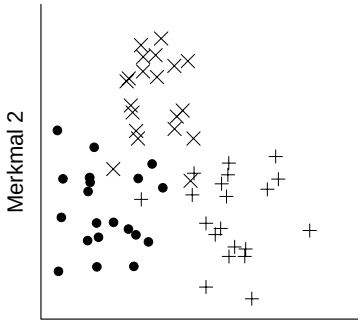
Untersuchung von Gruppenunterschieden anhand mehrerer Merkmale

Beispiel:

Untersuchungsobjekte:	Landwirtschaftliche Betriebe
Betriebskenngrößen:	Landwirtschaftliche Nutzfläche Großvieheinheiten Betriebseinkommen etc.
Einteilung:	Pflanzenbaubetriebe Tierhaltungsbetriebe Gemischtbetriebe

n	Objekte	($n_1 + n_2 + ... + n_g = n$)
g	bekannte Gruppen	
p	Merkmalsvariablen x_i	
x_{ijk}	Ausprägung des i-ten Merkmals des k-ten Objekts der j-ten Gruppe (i=1,...,p, j=1,...,g, k=1,...,n_j)	





Aufgaben

Prüfung der Diskriminanz

Unterscheiden sich die Gruppen anhand der Merkmale ihrer zugehörigen Objekte signifikant?

Beispiel: Existieren signifikante Unterschiede zwischen Standorten anhand von Zeigerpflanzen?

Prüfung der Merkmalsvariablen

Welche Merkmale sind zur Unterscheidung der Gruppen wichtig?

Beispiel: Welche Fettsäuren sind zur Gruppierung von verschieden Fettarten geeignet?

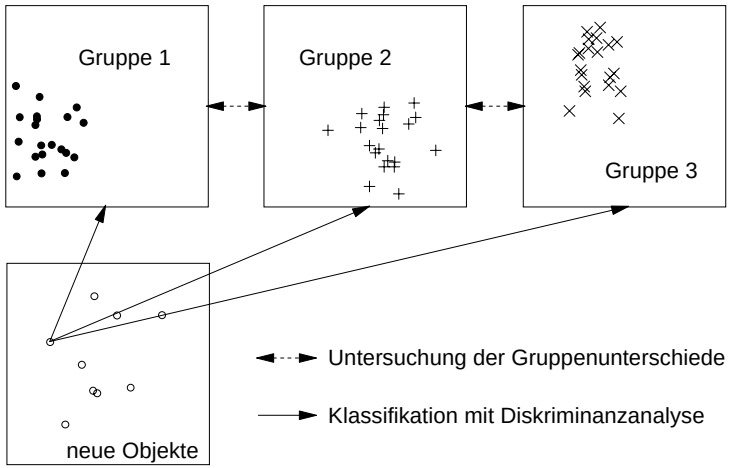
Klassifikation

In welche Gruppe ist ein Objekt, dessen Gruppenzugehörigkeit unbekannt ist, einzuordnen?

Beispiel: Welche Krankheit ist bei bestimmten Symptomen zu diagnostizieren?

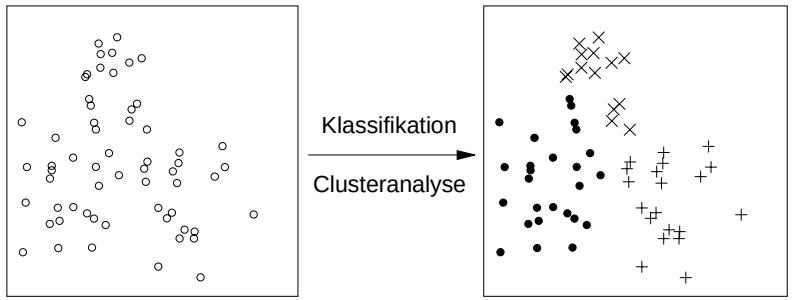
Diskriminanzanalyse

Gruppenzugehörigkeit bekannt
Einordnung neuer Objekte in bekannte Gruppen
überwachte Klassifikation (supervised classification)



Clusteranalyse

Gruppenzugehörigkeit unbekannt
Einordnung von Objekten in ähnliche Gruppen
unüberwachte Klassifikation (unsupervised classification)



Kanonische Diskriminanzfunktion

Prinzip: Mehrere Merkmalsvariablen bei minimalem Informationsverlust zu einer einzigen Variablen zusammenfassen

Lineare kanonische Diskriminanzfunktion (Trennfunktion)

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

y	Diskriminanzvariable
x_i	Merkmalsvariablen
$b_1\ b_2, \ldots, b_p$	Diskriminanzkoeffizienten
b_0	konstantes Glied

Abbildung aus dem *p*-dimensionalen Merkmalsraum in den eindimensionalen Raum der Diskriminanzvariablen

Bestimmung der Dikriminanzkoeffizienten durch Maximierung eines Diskriminanzmaßes

Diskriminanzwerte

Für jedes Objekt existiert genau ein Diskriminanzwert

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

$j = 1, 2, \ldots, g$
$k = 1, 2, \ldots, n_j$ $\quad (n_1 + n_2 + \ldots + n_g = n)$

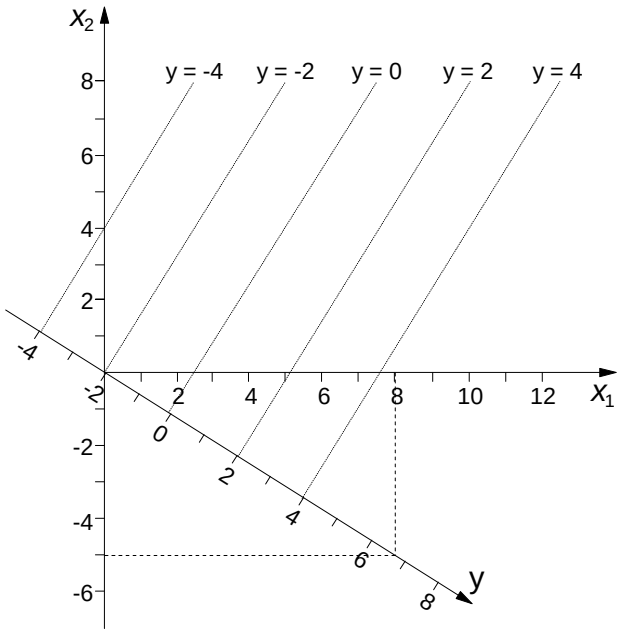
Geometrische Veranschaulichung

Isoquanten: Diskriminanzfunktion $y = b_0 + b_1 x_1 + b_2 x_2$ bildet bei 2 Merkmalen eine Ebene im Merkmalsraum und lässt sich für festen Wert von $y = c$ als Gerade darstellen

$$c = b_0 + b_1 x_1 + b_2 x_2 \quad \Rightarrow \quad x_2 = \frac{c - b_0}{b_1} - \frac{b_1}{b_2} \cdot x_1$$

Diskriminanzachse: Gerade durch den Nullpunkt senkrecht zu den Isoquanten

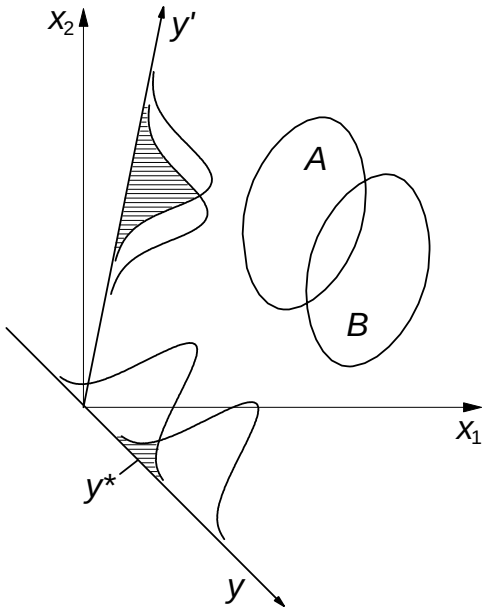
$$x_2 = \frac{b_2}{b_1} \cdot x_1$$



Isoquanten und Dikriminanzachse für $y = -2 + 0.8x_1 - 0.5x_2$

Diskriminanzachsen

Projektion der Verteilungen zweier Gruppen auf Diskriminanzachsen



je geringer Überschneidung, desto größer die Unterschiedlichkeit
Optimum: Diskriminanzachse, bei der Überschneidung minimal

y^* kritischer Wert auf Diskriminanzachse

Fehlklassifikation eines Objekts der Gruppe *A*, wenn sein Diskriminanzwert rechts vom kritischen Wert y^* liegt ($y < y^*$)

Fehlklassifikation eines Objekts der Gruppe *B*, wenn sein Diskriminanzwert links vom kritischen Wert y^* liegt ($y > y^*$)

Schätzung der Diskriminanzkoeffizienten

Diskriminanzmaß

$$\Gamma = \frac{\text{Variation zwischen den Gruppen}}{\text{Variation innerhalb der Gruppen}} = \frac{SQ_{zwischen}}{SQ_{innerhalb}} \rightarrow \max$$

$$SQ_{zwischen} = \sum_{j=1}^{g} n_j (\bar{y}_{j.} - \bar{y}_{..})^2$$

$$SQ_{innerhalb} = \sum_{j=1}^{g} \sum_{k=1}^{n_j} (y_{jk} - \bar{y}_{j.})^2$$

$$\Gamma = \frac{\sum_{j=1}^{g} n_j (\bar{y}_{j.} - \bar{y}_{..})^2}{\sum_{j=1}^{g} \sum_{k=1}^{n_j} (y_{jk} - \bar{y}_{j.})^2} \rightarrow \max$$

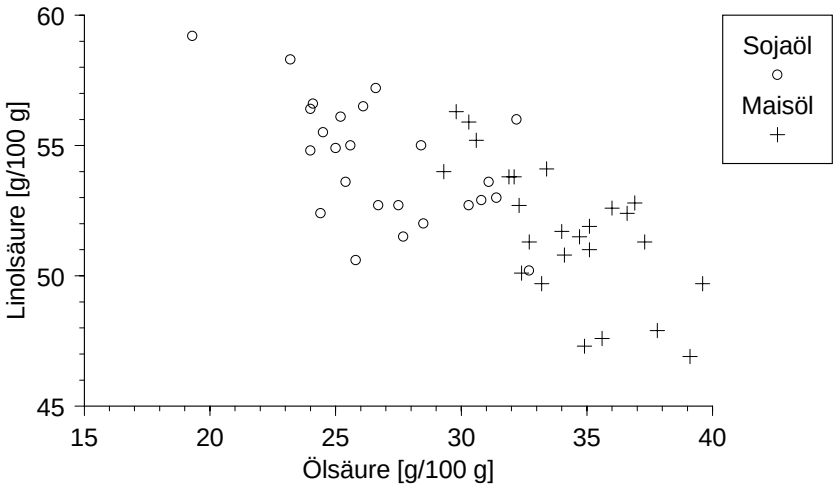
$\gamma = \max(\Gamma)$: **Eigenwert** der Diskriminanzfunktion
größte diskriminatorischer Bedeutung

Weitere Diskriminanzfunktionen

Bestimmung des nächst größeren Eigenwerts
ortogonale Diskriminanzfunktionen
Höchstzahl: $q = \min(g-1, p)$

$\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_q$
Eigenwertanteil: $A_i = \frac{\gamma_i}{\gamma_1 + \gamma_2 + \ldots + \gamma_q}$

Sojaöl und Maisöl



g = 2 Gruppen: Sojaöl, Maisöl
p = 2 Merkmale: Ölsäuregehalt (x_1), Linolsäuregehalt (x_2)
n = 50 Proben: $n_1 = 25$, $n_2 = 25$

Diskriminanzfunktion: $y = -18.1492 + 0.3835\, x_1 + 0.1216\, x_2$

Mittelwerte: $\bar{y}_{..} = 0,\ \bar{y}_{1.} = -1.25, \bar{y}_{2.} = 1.25$

$SQ_{zwischen}$: $\sum_{j=1}^{g} n_j (\bar{y}_{j.} - \bar{y}_{..})^2 = \sum_{j=1}^{2} 25 \cdot 1.25^2 = 78.125$

$SQ_{innerhalb}$: $\sum_{j=1}^{g} \sum_{k=1}^{n_j} (y_{jk} - \bar{y}_{j.})^2 = \sum_{j=1}^{2} \sum_{k=1}^{25} (y_{jk} - \bar{y}_{j.})^2 = 47.999$

Diskriminanzmaß: $\Gamma = \frac{SQ_{zwischen}}{SQ_{innerhalb}} = \frac{78.125}{47.999} = 1.628$

Eigenwertanteil: $A = 1 = 100\%$

Prüfung der Diskriminanz

Wilks' Lambda

$$\Lambda_k = \prod_{i=k+1}^{q} \frac{1}{1+\gamma_i} = \frac{1}{1+\gamma_{k+1}} \cdot \frac{1}{1+\gamma_{k+2}} \cdot \ldots \cdot \frac{1}{1+\gamma_q}$$

Transformation

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

χ^2-verteilt mit $(p - k) \cdot (g - k - 1)$ Freiheitsgraden

Sojaöl und Maisöl

$$\Lambda_0 = \prod_{i=1}^{1} \frac{1}{1+\gamma_i} = \frac{1}{1+\gamma} = \frac{1}{1+1.628} = 0.381$$

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

Freiheitsgrade: $(2 - 0) \cdot (2 - 0 - 1) = 2$

$$\chi_0^2 = 45.353 > 9.210 = \chi_{2;0.99}^2$$

Hochsignifikante (α = 1%) diskriminatorische Bedeutung der Diskriminanzfunktion

Prüfung der Merkmalsvariablen

Standardisierte Diskriminanzkoeffizienten

$$b_i^* = b_i \cdot s_i$$

s_i: Standardabweichung der i-ten Merkmalvariablen

Soja- und Maisöl

$b_1 = 0.3835,\ s_1 = 4.8720$
$b_2 = 0.1216,\ s_2 = 2.7918$

$b_1^* = 0.3835 \cdot 4.7820 = 1.8337$
$b_2^* = 0.1216 \cdot 2.7918 = 0.3395$

Ölsäure (x_1) hat größere Bedeutung für die Unterscheidung der Gruppen als Linolsäure (x_2)

Klassifikation

Welcher Gruppe wird ein Objekt *o* unbekannter Gruppenzugehörigkeit anhand seiner Merkmale $o_1,o_2,...,o_p$ zugeordnet?

Klassifikation mittels Distanzen (Distanzkonzept)

Ein Objekt *o* wird in die Gruppe klassifiziert, zu deren Centroid (Gruppenmittelpunkt) es den kleinsten Abstand hat.

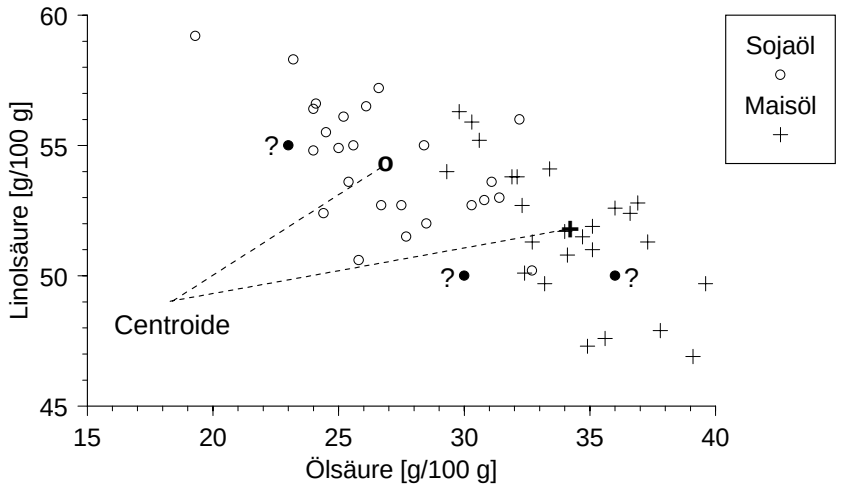
Klassifikation mittels Klassifizierungsfunktionen

Ein Objekt *o* wird in die Gruppe klassifiziert, für die die lineare Diskriminanzfunktion am größten ist.

Klassifikation mittels Wahrscheinlichkeiten (Wahrscheinlichkeitskonzept)

Ein Objekt *o* wird in die Gruppe klassifiziert, für die die Wahrscheinlichkeit, dass es zu dieser Gruppe gehört, bei gegebenem Diskriminanzwert am größten ist.

Soja- und Maisöl



Objekt *o* mit o_1 = 30% Öl- und o_2 = 50% Linolsäure,*o* = (30,50)

Diskriminanzfunktion: $y = -18.1492 + 0.3835\,x_1 + 0.1216\,x_2$

Mittelwerte: $\bar{y}_{..} = 0,\ \bar{y}_{1.} = -1.25, \bar{y}_{2.} = 1.25$

Euklidische Distanz: $\left|y(o) - \bar{y}_{j.}\right| = \sqrt{y(o)^2 - \bar{y}_{j.}^2}$

Kritischer Wert: $y^* = (\bar{y}_{1.} + \bar{y}_{2.})/2 = 0$

Diskriminanzwert: $y(o) = -0.5642 < 0$

Distanzen: $\left|y(o) - \bar{y}_{1.}\right| = |-0.5642 - (-1.25)| = 0.6858$

$\left|y(o) - \bar{y}_{2.}\right| = |-0.5642 - 1.25| = 1.8142$

Klassifizierung der Probe als Sojaöl

Distanzkonzept und Klassifizierungsfunktionen

Quadrierte euklidische Distanz

$$DY_j^2(o) = \sum_{l=1}^{q} \left(y^{(l)}(o) - \bar{y}_j^{(l)} \right)^2 \qquad (j = 1,2,\ldots,g)$$

Mahalanobis-Distanz

$$DX_j^2(o) = \left(o - \bar{x}_j\right)' S^{-1} \left(o - \bar{x}_j\right) \qquad (j = 1,2,\ldots,g)$$

S: Kovarianzmatrix der Merkmalsvariablen

Gleiche Streuungen in den Gruppen:

$$S = S_p = \left(s_{rc} = \sum_{j=1}^{g} \sum_{k=1}^{n_j} \left(x_{rjk} - \bar{x}_{rj.}\right)\left(x_{cjk} - \bar{x}_{cj.}\right) \right)_{p \times p}$$

S_p: Innergruppen-Kovarianzmatrix

$$\overline{DX}_j^2(o) = -2\left(\bar{x}_j' S_p^{-1} o - 0.5\, \bar{x}_j' S_p^{-1} \bar{x}_j\right) + o' S_p^{-1} o$$

Lineare Diskriminanz- bzw. Klassifizierungsfunktionen

Ungleiche Streuungen in den Gruppen:

$$DX_j^2(o) = \left(o - \bar{x}_j\right)' S_j^{-1} \left(o - \bar{x}_j\right)$$

S_j: Kovarianzmatrix der Merkmalsvariablen in Gruppe j

Quadratische Diskriminanz- bzw. Klassifizierungsfunktionen

Wahrscheinlichkeitskonzept

Mit A-priori-Wahrscheinlichkeiten lässt sich berücksichtigen, dass in der Realität Objekte in den betrachteten Gruppen mit unterschiedlicher Häufigkeit vorkommen.

A-priori-Wahrscheinlichkeit: $P_o(j) \qquad \left(\sum_{j=1}^{g} P_o(j) = 1 \right)$

A-posteriori-Wahrscheinlichkeit: $$P(j|y(o)) = \frac{P(y(o)|j) \cdot P_o(j)}{\sum_{k=1}^{g} \left(P(y(o)|k) \cdot P_o(k) \right)}$$

Bedingte Wahrscheinlichkeit: $P(y(o)|j) = e^{-DY_j^2(o)/2}$

$$P(j|y(o)) = \frac{e^{-DY_j^2(o)/2} \cdot P_o(j)}{\sum_{k=1}^{g} \left(e^{-DY_k^2(o)/2} \cdot P_o(k) \right)}$$

Falls $P(j|y(o))$ maximal ist, dann auch

$$\ln P(j|y(o)) = \frac{-DY_j^2(o)}{2} + \ln P_o(j) - \ln \sum_{k=1}^{g} \left(e^{-DY_k^2(o)/2} \cdot P_o(k) \right)$$

Da $\sum_{k=1}^{g} \left(e^{-DY_k^2(o)/2} \cdot P_o(k) \right)$ unabhängig von j ist, folgt:

Ein Objekt o wird in die Gruppe klassifiziert, für die der Ausdruck $-\frac{1}{2}\left(DY_j^2(o) - 2\ln P_o(j)\right)$ maximal ist.

Prüfung der Klassifikation mit Ausgangsdaten

Klassifikationsmatrix

klassifiziert in	Wahre Gruppenzugehörigkeit					
	Gruppe 1	...	Gruppe j	...	Gruppe g	
Gruppe 1	m_{11}	...	m_{1j}	...	m_{1g}	
⋮	⋮		⋮		⋮	
Gruppe j	m_{j1}	...	m_{jj}	...	m_{jg}	
⋮	⋮		⋮		⋮	
Gruppe g	m_{g1}	...	m_{gj}	...	m_{gg}	
gesamt	n_1	...	n_j	...	n_g	n

Gesamtanteil richtig klassifizierter Objekte: $\dfrac{\sum_{j=1}^{g} m_{jj}}{n}$

Anteil richtig klassifizierter Objekte pro Gruppe: $\dfrac{m_{jj}}{n_j}$

Soja- und Maisöl

klassifiziert als	Wahre Fettart		
	Sojaöl	Maisöl	
Sojaöl	20	1	
Maisöl	5	24	
gesamt	25	25	50
korrekt	80%	96%	88%

Prüfung der Klassifikation mit Kreuzvalidierung

1. Schritt: 1. Objekt aus Gesamtdatensatz entfernt
Diskriminanzanalyse mit übrigem Datensatz
Klassifikation des 1. Objekts

2. Schritt: 2. Objekt aus Gesamtdatensatz entfernt
Diskriminanzanalyse mit übrigem Datensatz
Klassifikation des 2. Objekts

usw.

n-ter Schritt: *n*-tes Objekt aus Gesamtdatensatz entfernt
Diskriminanzanalyse mit übrigem Datensatz
Klassifikation des *n*-ten Objekts

n Sätze von Diskriminanz- bzw. Klassifizierungsfunktionen
i.a. weniger richtig klassifizierte Objekte in Klassifikationsmatrix

Prüfung der Klassifikation mit Kontrolldaten

Entfernung von Kontrolldaten aus dem Gesamtdatensatz

Diskriminanzanalyse mit restlichem Datensatz

Klassifikation des Kontrolldatensatzes

Klassifikationsmatrix des Kontrolldatensatzes

Soja- und Maisöl - Diskriminanzanalyse

```
MTB > Discriminant 'Fett' 'Oel' 'Linol';
SUBC>   Predict 30 50;
SUBC>   Brief 3.
```

Discriminant Analysis

Summary of Classification

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

N = 50 N Correct = 44 Prop. Correct = 0.880

Summary of Misclassified Observations

Obs.	True Group	Pred Group	Group	Squared Distance	Prob.
3 **	Sojaoel	Maisoel	Maisoel	1.332	0.560
			Sojaoel	1.814	0.440
4 **	Sojaoel	Maisoel	Maisoel	1.068	0.677
			Sojaoel	2.552	0.323
6 **	Sojaoel	Maisoel	Maisoel	1.642	0.776
			Sojaoel	4.126	0.224
7 **	Sojaoel	Maisoel	Maisoel	0.8531	0.700
			Sojaoel	2.5471	0.300
14 **	Sojaoel	Maisoel	Maisoel	3.386	0.926
			Sojaoel	8.440	0.074
28 **	Maisoel	Sojaoel	Maisoel	2.6171	0.297
			Sojaoel	0.8916	0.703

Prediction for Test Observations

Obs.	Pred Group	From Group	Sq Dist	Prob.
1	Sojaoel			
		Maisoel	5.974	0.197
		Sojaoel	3.158	0.803

Soja- und Maisöl - Diskriminanzanalyse (Forts.)

Squared Distance Between Groups

	Maisoel	Sojaoel
Maisoel	0.00000	6.25326
Sojaoel	6.25326	0.00000

Linear Discriminant Function for Group

	Maisoel	Sojaoel
Constant	-721.63	-676.24
Oel	13.50	12.54
Linol	18.99	18.69

Variable	Pooled Mean	Means for Group Maisoel	Sojaoel
Oel	30.506	34.192	26.820
Linol	53.034	51.692	54.376

Variable	Pooled StDev	StDev for Group Maisoel	Sojaoel
Oel	3.031	2.802	3.244
Linol	2.466	2.585	2.341

Pooled Covariance Matrix

	Oel	Linol
Oel	9.187	
Linol	-4.727	6.080

Covariance Matrix for Group Maisoel

	Oel	Linol
Oel	7.849	
Linol	-5.013	6.681

Covariance Matrix for Group Sojaoel

	Oel	Linol
Oel	10.524	
Linol	-4.441	5.480

Soja- und Maisöl - Datensatz

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

Notation

$o = (o_1,o_2,...,o_p)'$ ist der Vektor der gemessenen Merkmalsausprägungen eines Untersuchungsobjekts.

$y^{(l)} = y^{(l)}((x_1,x_2,...,x_p)) = b_0^{(l)} + b_1^{(l)}x_1 + b_2^{(l)}x_2 + ... + b_p^{(l)}x_p$ ist die l-te Diskriminanzfunktion mit Diskriminanzkoeffizienten b_i und $y^{(l)}(o)$ der zugehörige Diskriminanzwert des Objekts o.

$\bar{x}_{ij.}$ ist der empirische Mittelwert der i-ten Merkmalsvariablen der j-ten Gruppe.

$\bar{x}_j = (\bar{x}_{1j.},\bar{x}_{2j.},...,\bar{x}_{pj.})$ ist Mittelpunkt oder **Centroid** der j-ten Gruppe.

$\bar{y}_j^{(l)} = y^{(l)}(\bar{x}_j)$ ist der Diskriminanzwert des Centroids der j-ten Gruppe bezüglich der l-ten Diskriminanzfunktion.

$P_o(j)$ Ist die gegebene oder geschätzte Wahrscheinlichkeit der Zugehörigkeit eines Objekts o zur j-ten Gruppe vor Berechnung der Diskriminanzfunktionen (**A- priori-Wahrscheinlichkeit**).

$P(j|y(o))$ ist die Wahrscheinlichkeit, dass das Objekt o zur Gruppe j gehört bei gegebenem Diskriminanzwert $y(o)$ (**A-posteriori-Wahrscheinlichkeit**).

$P(y(o)|j)$ ist die Wahrscheinlichkeit, dass das Objekt o den Diskriminanzwert $y(o)$ hat beigegebener Zugehörigkeit zur j-ten Gruppe (**bedingte Wahrscheinlichkeit**).