Multivariate Verfahren

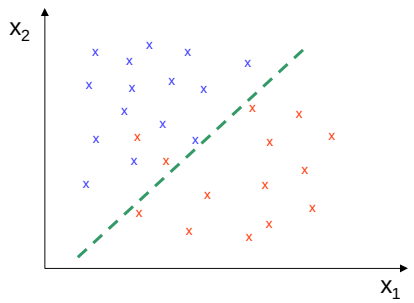
	Lineare Regression	Hauptkomponentenanalyse	Korrespondenzanalyse	Clusteranalyse	Diskriminanzanalyse
Zweck:					
Vorhersage	x				
Dimensionsreduktion		x	x		
Klassifizierung				x	x
Eigenschaften:					
nicht-linear					
verteilungsfrei			x	x	
nominal skalierte Var.			x		x

Diskriminanzanalyse

Prinzip:

- Bildung von Diskriminanzfunktionen (= Linearkombinationen) der Diskriminanzvariablen y mit den Diskriminanzkoeffizienten b_0, b_1, b_2, ... :

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + ...$$



Diskriminanzanalyse

zu untersuchende Fragen:

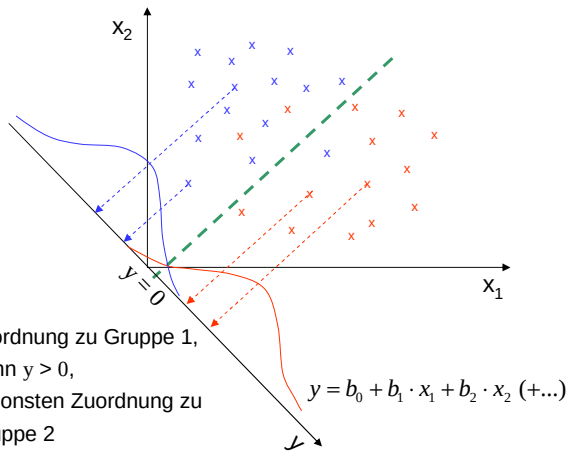
- Unterscheiden sich die ausgewiesenen Gruppen signifikant hinsichtlich der Variablen?
- Welche Variablen sind zur Unterscheidung der Gruppen geeignet?

Vorgehensweise:

→ Definition der Gruppen durch Zuweisung einer nominal skalierten Variablen (z.B. anhand der Ergebnisse einer Clusteranalyse)

→ i.d.R. Anzahl der Gruppen $G <$ Anzahl der Variablen

Diskriminanzanalyse



Zuordnung zu Gruppe 1, wenn $y > 0$, ansonsten Zuordnung zu Gruppe 2

$$y = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 \ (+...)$$

Diskriminanzkriterium

- Diskriminanzkriterium = Maß für die Unterschiedlichkeit der Gruppen

$$\gamma = \frac{\text{Streuung zwischen den Gruppen}}{\text{Streuung innerhalb der Gruppen}} = \frac{\text{erklärte Streuung}}{\text{nicht erklärte Streuung}}$$

$$= \frac{\sum_{g=1}^{G}\left(I_g(\bar{y}_g - \bar{y})^2\right)}{\sum_{g=1}^{G}\sum_{i=1}^{I_g}(y_{gi} - \bar{y}_g)^2} = \max$$

y: Diskriminanzfunktionswert
G: Anzahl der Gruppen
I_g: Anzahl der Fälle in der Gruppe

Diskriminanzkriterium

- Diskriminanzkriterium = Maß für die Unterschiedlichkeit der Gruppen

$$\gamma = \frac{\text{Streuung zwischen den Gruppen}}{\text{Streuung innerhalb der Gruppen}} = \frac{\text{erklärte Streuung}}{\text{nicht erklärte Streuung}} = \textbf{\textit{F}}\text{-Wert}$$

$$= \frac{\sum_{g=1}^{G}\left(I_g(\bar{y}_g - \bar{y})^2\right)}{\sum_{g=1}^{G}\sum_{i=1}^{I_g}(y_{gi} - \bar{y}_g)^2} = \max$$

y: Diskriminanzfunktionswert
G: Anzahl der Gruppen
I_g: Anzahl der Fälle in der Gruppe

$$\left[\frac{\text{Summe der quadrierte Abweichungen der Gruppenzentroide vom Gesamtmittel}}{\text{Summe der quadrierten Abweichungen vom Gruppenzentroid}}\right]$$

- Zentroid = arithmetisches Mittel der Diskriminanzwerte einer Gruppe

Diskriminanzkriterium

- Diskriminanzkriterium = Maß für die Unterschiedlichkeit der Gruppen

$$\gamma = \frac{\text{Streuung zwischen den Gruppen}}{\text{Streuung innerhalb der Gruppen}} = \frac{\text{erklärte Streuung}}{\text{nicht erklärte Streuung}} = \textbf{\textit{F}}\text{-Wert}$$

$$= \frac{\sum_{g=1}^{G}\left(I_g(\bar{y}_g - \bar{y})^2\right)}{\sum_{g=1}^{G}\sum_{i=1}^{I_g}(y_{gi} - \bar{y}_g)^2} = \max$$

y: Diskriminanzfunktionswert
G: Anzahl der Gruppen
I_g: Anzahl der Fälle in der Gruppe

- ***F***-Wert: $F = \gamma \cdot \frac{G \cdot (I-1)}{G-1} = \frac{\sum_{g=1}^{G}\left(I_g(\bar{y}_g - \bar{y})^2\right)}{\sum_{g=1}^{G}\sum_{i=1}^{I_g}(y_{gi} - \bar{y}_g)^2} \cdot \frac{G(I-1)}{G-1}$

Diskriminanzfunktionen

- jede Diskriminanzfunktion teilt den Datensatz in zwei Gruppen auf
- für $\boldsymbol{G}$ >2 Gruppen: sukzessive Aufspaltung der Datensatzes durch ($\boldsymbol{G}$ -1) Diskriminanzfunktionen
- Diskriminanzfunktionen sind untereinander unkorreliert
- **Diskriminanzraum** = Menge aller Diskriminanzfunktionen
- **Eigenwertanteil** $\boldsymbol{EA_k}$ = **Diskriminanzanteil** der Diskriminanzfunktion $\boldsymbol{y_k}$:

$$EA_k = \frac{\gamma_k}{\sum_{g=1}^{G}\gamma_g}$$

k: Index der Diskriminanzfunktion

- Bestimmung der Diskriminanzfunktionen durch Eigenwertzerlegung

Kanonischer Korrelationskoeffizient

- Normierung durch Setzen der nicht-erklärten Streuung auf 1:

$$\gamma = \frac{\gamma}{1} = \frac{\text{erklärte Streuung}}{\text{nicht erklärte Streuung}} \qquad (\gamma : \text{Eigenwert der Disk.funktion}) \Rightarrow$$

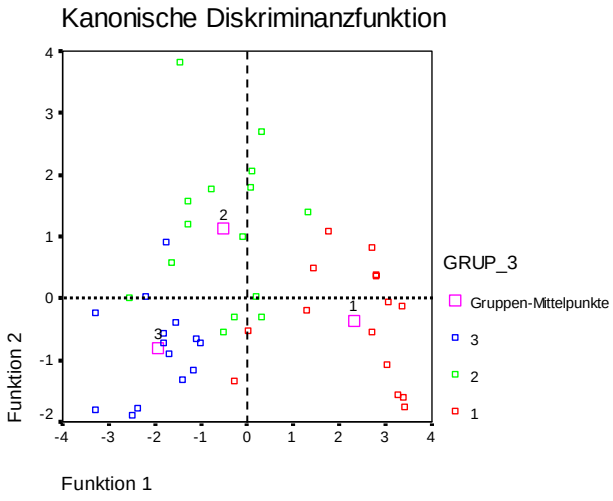
$$\frac{\gamma}{1+\gamma} = \frac{\text{erklärte Streuung}}{\text{Gesamtstreuung}} \quad \text{und} \quad \frac{1}{1+\gamma} = \frac{\text{nicht erklärte Streuung}}{\text{Gesamtstreuung}}$$

- Kanonischer Korrelationskoeffizient:

$$c = \sqrt{\frac{\gamma}{1+\gamma}}$$

→ für Anzahl der Gruppen G **= 2** identisch mit Korrelation zwischen Diskriminanzwerten und Gruppierungsvariable

Diskriminanzfunktionswerte



Klassifikationsfunktion

- nicht zu verwechseln mit der Diskriminanzfunktion!
- beide: Linearkombinationen der Werte der einzelnen Variablen

Klassifikationsfunktion:

- spezifisch für jede Gruppe (Anzahl = ***G***)
- einzelne Fälle werden jeweils der Gruppe zugeordnet, für die die jeweilige Klassifikationsfunktion den höchsten Wert ergibt

Diskriminanzfunktion:

- trennt jeweils zwischen verschiedenen Gruppen
- Anzahl der Diskriminanzfunktionen = ***G-1***

Signifikante Diskriminanz zwischen zwei Gruppen

- Nullhypothese: die beiden Gruppen unterscheiden sich nicht
- Wilks' Lambda: $\Lambda = \frac{1}{1+\gamma} = \frac{\text{nicht erklärte Streuung}}{\text{Gesamtstreuung}}$ und $\Lambda = 1 - c^2$
- Prüfgröße: $\chi^2 = \left(l - \frac{n+G}{2} - 1 \right) \cdot \ln \Lambda$ mit $n \cdot (G\text{-}1)$ Freiheitsgraden

  (l: Anzahl der Fälle, n: Anzahl der Variablen, G: Anzahl der Gruppen)

Signifikante Diskriminanz zwischen mehreren Gruppen

=> multivariates Wilks' Lambda: $\Lambda = \prod_{k=1}^{K} \frac{1}{1+\gamma_k}$ bzw. $\ln \frac{1}{\Lambda} = -\ln \Lambda$

(k: Anzahl der Diskriminanzfunktionen)

- residuelle Diskriminanz: $\Lambda = \prod_{q=k+1}^{K} \frac{1}{1+\gamma_q}$

  (nach Bestimmung von k Diskriminanzfunktionen)

- Prüfgröße: $\chi^2 = \left(l - \frac{n+G}{2} - 1 \right) \cdot \ln \Lambda$ mit $(n\text{-}k)\cdot(G\text{-}k\text{-}1)$ Freiheitsgraden

  (l: Anzahl der Fälle, n: Anzahl der Variablen, G: Anzahl der Gruppen)

Beitrag der einzelnen Variablen

- **standardisierter Diskriminanzkoeffizient**: $b_i^* = b_i \cdot s_i$

  mit s_i = Innergruppen-Varianz von x_i :

  (SS_{in}: *Sum of Squares within groups*) $s_i = \sqrt{\frac{SS_{in}}{I-G}}$

- **mittlerer Diskriminanzkoeffizient** einer Variablen x_i (für k Disk.-Funktionen):

  $$\overline{b_i} = \sum_{k=1}^{K} |b^*_{ik}| \cdot EA_k$$

- **Ladung** = Korrelation der Variablen mit einzelnen Diskriminanzfunktionen

Kreuzvalidierung

Kreuzvalidierung = **leave-one-out-Validierung** (<100 Beobachtungen):

- Bestimmung der Diskriminanzfunktionen mit allen außer einer Beobachtung, anschließend Überprüfung der Klassifizierung an dieser ausgelassenen Beobachtung

k-fache Kreuzvalidierung (>100 Beobachtungen):

- Aufteilung des Datensatzes in k verschiedene Untergruppen
- Bestimmung der Diskriminanzfunktionen mit allen außer einer Untergruppe, anschließend Überprüfung der Klassifizierung an dieser ausgelassenen Untergruppe

Vergleich der Gruppenzentroide

Gleichheitstest der Gruppenmittelwerte

	Wilks-Lambda	F	df1	df2	Signifikanz
CA	.889	1.751	2	28	.192
CL	.947	.783	2	28	.467
DOC	.969	.447	2	28	.644
FE	.956	.641	2	28	.535
K	.992	.110	2	28	.896
MG	.473	15.579	2	28	.000
MN	.565	10.758	2	28	.000
NA	.956	.652	2	28	.529
NO3	.868	2.123	2	28	.139
PH	.985	.213	2	28	.810
SI	.817	3.135	2	28	.059
SO4	.902	1.516	2	28	.237
El. Leitfähigkeit	.784	3.864	2	28	.033

Erklärte Varianz

Eigenwerte

Funktion	Eigenwert	% der Varianz	Kumulierte %	Kanonische Korrelation
1	7.424[a]	84.6	84.6	.939
2	1.354[a]	15.4	100.0	.758

a. Die ersten 2 kanonischen Diskriminanzfunktionen werden in dieser Analyse verwendet.

Wilks' Lambda

Test der Funktion(en)	Wilks-Lambda	Chi-Quadrat	df	Signifikanz
1 bis 2	.050	65.719	26	.000
2	.425	18.836	12	.093

Beiträge der Variablen

Struktur-Matrix

	Funktion	
	1	2
MG	.383*	-.140
MN	.309*	.207
K	-.030*	-.029
NO3	.047	-.316*
El. Leitfähigkeit	.153	.273*
SI	.141	.238*
FE	-.013	.181*
SO4	.101	.156*
CA	.111	.156*
NA	.049	.145*
CL	.061	.144*
DOC	-.033	.133*
PH	.031	.078*

Gemeinsame Korrelationen innerhalb der Gruppen zwischen Diskriminanzvariablen und standardisierten kanonischen Diskriminanzfunktionen
Variablen sind nach ihrer absoluten Korrelationsgröße innerhalb der Funktion geordnet.

*. Größte absolute Korrelation zwischen jeder Variablen und einer Diskriminanzfunktion

Standardisierte kanonische Diskriminanzfunktionskoeffizienten

	Funktion	
	1	2
CA	-.013	-.185
CL	2.385	1.701
DOC	-.051	1.182
FE	-.511	.123
K	-2.336	1.312
MG	1.398	-1.083
MN	.477	.256
NA	-.931	-1.920
NO3	.650	.077
PH	-.419	.938
SI	1.082	1.400
SO4	1.825	.763
El. Leitfähigkeit	-1.393	.787

Mittlere Diskriminanzfunktionskoeffizienten: Gewichtung der Werte für die einzelnen Disk.-Funktionen mit dem Eigenwert der jeweiligen Disk. Funktion

Klassifikationsmatrix

Klassifizierungsergebnisse[b,c]

			Vorhergesagte Gruppenzugehörigkeit			
		GRUP_3	1	2	3	Gesamt
Original	Anzahl	1	13	1	1	15
		2	0	13	2	15
		3	0	1	13	14
	%	1	86.7	6.7	6.7	100.0
		2	.0	86.7	13.3	100.0
		3	.0	7.1	92.9	100.0
Kreuzvalidiert[a]	Anzahl	1	9	5	1	15
		2	5	5	5	15
		3	1	2	11	14
	%	1	60.0	33.3	6.7	100.0
		2	33.3	33.3	33.3	100.0
		3	7.1	14.3	78.6	100.0

a. Die Kreuzvalidierung wird nur für Fälle in dieser Analyse vorgenommen. In der Kreuzvalidierung ist jeder Fall durch die Funktionen klassifiziert, die von allen anderen Fällen außer diesem Fall abgeleitet werden.

b. 88.6% der ursprünglich gruppierten Fälle wurden korrekt klassifiziert.

c. 56.8% der kreuzvalidierten gruppierten Fälle wurden korrekt klassifiziert.

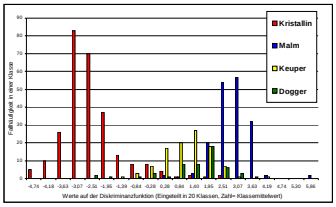
Beispiel

Diplomarbeit P. Klaas (2003):

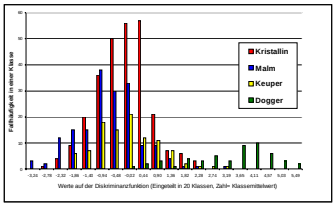
- Grundwassermessstellen aus dem Bereich des Kristallins, Dogger, Malm, Keuper in Nordbayern
- 597 Probenahmestellen, 38 Variablen
- Hauptkomponentenanalyse, Clusteranalyse, Diskriminanzanalyse, Korrespondenzanalyse

Diskriminanzanalyse

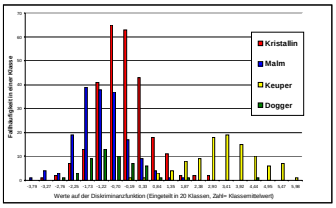
1. Disk.-Funktion (Kan. Korr. 0,92)



3. Disk.-Funktion (Kan. Korr. 0,72)

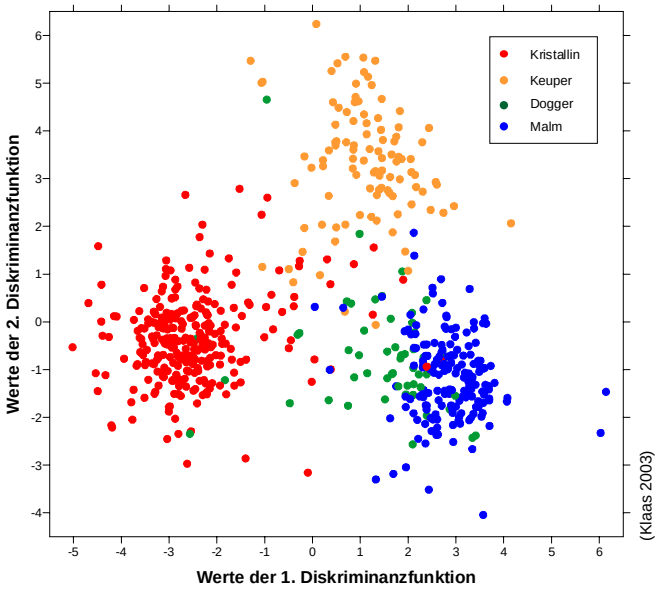


2. Disk.-Funktion (Kan. Korr. 0,83)

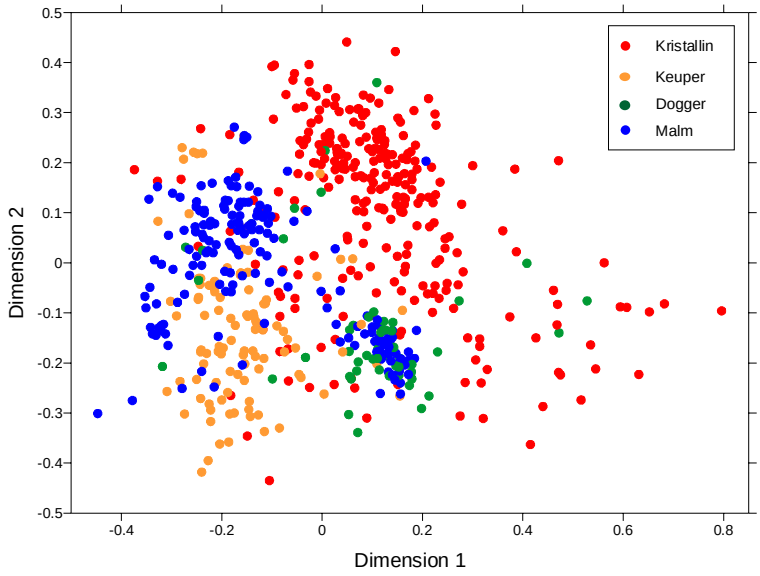


(Klaas 2003)

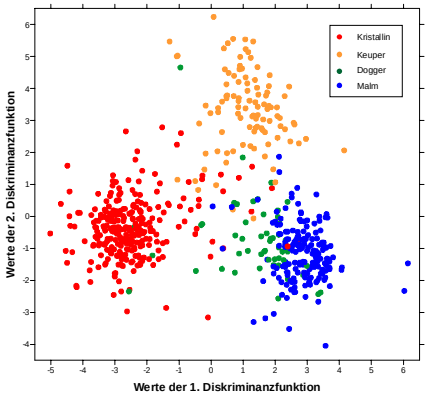
Diskriminanzfunktionen



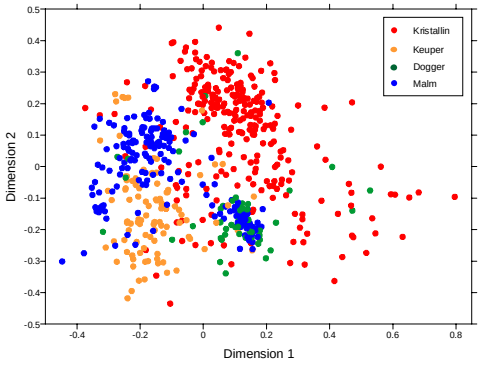
Korrespondenzanalyse



Diskriminanzanalyse



Korrespondenzanalyse



Aufgabe

1. Weisen Sie, basierend auf den Ergebnissen der Clusteranalyse, Gruppen aus.

2. Führen Sie eine Diskriminanzanalyse durch, indem Sie
   a) alle Variablen gleichzeitig („*Standard*“),
   b) die Variablen schrittweise nacheinander („*schrittweise vorwärts*“) aufnehmen.
   Stellen Sie die Ergebnisse grafisch dar. Durch welche Funktionen werden die einzelnen Gruppen abgetrennt?

3. Unterscheiden sich die ausgewiesenen Gruppen signifikant? Welche Variablen sind zur Unterscheidung der Gruppen besonders gut geeignet? Erstellen Sie Box-Whisker-Plots für ausgesuchte Variablen.

4. Überprüfen Sie die Robustheit der Diskriminanzanalyse, indem sie eine Kreuzvalidierung der Klassifizierung vornehmen. Schließen Sie dazu für die Diskriminanzanalyse einzelne Fälle aus, und überprüfen Sie anschließend mittels der Klassifikationsfunktion die Zuordnung dieses Falles. Stellen Sie die Ergebnisse in einer Tabelle dar.