KSFE
KONFERENZ DER SAS®-ANWENDER IN FORSCHUNG UND ENTWICKLUNG

6. Konferenz für SAS-Anwender in Forschung und Entwicklung
28. Februar - 1. März 2002, Universität Dortmund

Bewertung von Diskriminanzanalyseverfahren in SAS bei Nichtnormalität

A. Tuchscherer[*], P. E. Rudolph[*], B. Jäger[**], M.Tuchscherer[*]

* Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere Dummerstorf

** Institut für Biometrie und Medizinische Informatik
Ernst-Moritz-Arndt-Universität Greifswald

Gliederung:

1 Einleitung

2

3

4

5

Einleitung

Diskriminanzanalyse:

Gesamtheit (Menge von Objekten) **aus disjunkten Teilgesamtheiten** (Klassen, Gruppen) :

1

d.h. jedes Element (Objekt) der Gesamtheit gehört zu genau einer Teilgesamtheit.

2

Objekte der Gesamtheit lassen sich nur sinnvoll durch die **gleichzeitige Betrachtung von n Merkmalen**, die an den Objekten messbar seien, unterscheiden.

3

Gegeben:

Messwerte der n Merkmale für Objekte aus allen Teilgesamtheiten, für die bekannt ist, zu welcher Teilgesamtheit sie gehören

4

⇩

Lernstichprobe

Objekt mit beobachteten Merkmalsvektors **x** und unbekannter Zugehörigkeit zu einer der vorliegenden Teilgesamtheiten (Klassen, Gruppen)

5

Einleitung

Ziel der Diskriminanzanalyse:

Objekt mit unbekannter Zugehörigkeit zu einer der Teilgesamtheiten soll genau einer der vorliegenden Teilgesamtheiten (Klassen, Gruppen) auf der Grundlage seines beobachteten Merkmalsvektors **x** und der Lernstichprobe zugeordnet werden.

Die Lösung des Zuordnungsproblems (Klassifizierung):

Geeignete Entscheidungsregeln, die auf der Bestimmung einer oder mehrerer optimaler **Trennfunktionen** (**Diskriminanzfunktionen**) basieren.

PROC DISCRIM

Einleitung

Praxis: Daten mit n + m Merkmalen

1 **PROC STEPDISC**

2 Forward selection | Backward elimination

Stepwise selection

3 Auswahl einer Teilmenge von n Merkmalen, die am besten 'trennen'

4

5 **PROC DISCRIM**

1

2

Diskriminanzanalysen mit der SAS-Prozedur DISCRIM

3

4

5

Diskriminanzanalysen mit der SAS-Prozedur DISCRIM

Mit der **SAS-Prozedur DISCRIM** steht dem Anwender ein umfangreiches Werkzeug zur Durchführung von Diskriminanzanalysen zur Verfügung:

1. METHOD=NORMAL — **parametrische Verfahren**

2. multivariate Normalverteilung innerhalb der Klassen

3. POOL=YES — gemeinsame Klassen-Kovarianzmatrizen: lineare Diskriminanzfunktion

4. POOL=NO — individuelle Klassen-Kovarianzmatrizen: quadratische Diskriminanzfunktion

5.

Diskriminanzanalysen mit der SAS-Prozedur DISCRIM

METHOD=NPAR **nichtparametrische Verfahren**

1 nicht multivariate Normalverteilung oder unbekannte multivariate Verteilung innerhalb der Klassen

(nichtparametrische Schätzung der klassenspezifischen Wahrscheinlichkeitsdichten)

2

'k-nearest-neighbor' - Methode

3 K=k

'k nächste Nachbarn'-Regel:

4 Zuordnung eines Objekts mit Merkmalsvektor **x** zu einer Klasse auf Grundlage der Information der k nächsten Nachbarn von **x**

5 POOL=YES

Diskriminanzanalysen mit der SAS-Prozedur DISCRIM

METHOD=NPAR **nichtparametrische Verfahren**

1 'kernel' - Methode

klassenspezifische Kerndichteschätzung

2 R=r

Zuordnung eines Objekts mit Merkmalsvektor **x** zu einer Klasse auf Grundlage der Information aller **y** aus der Lernstichprobe, die innerhalb einer Umgebung von **x** mit **Radius R** liegen

3 **R:**

4 abhängig von der Anzahl der Merkmale (Dim(**x**)), der Anzahl der Klassen und vom gewählten Kern

KERNEL=NORMAL: gesamte Lernstichprobe

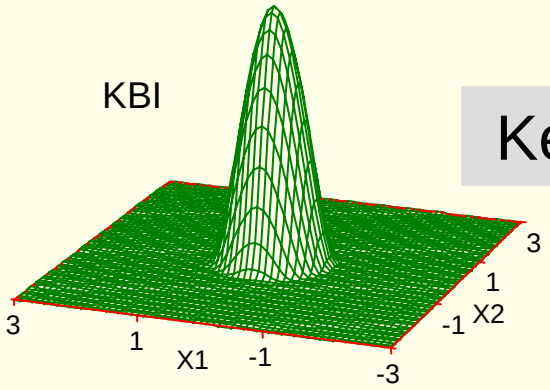
5 POOL=NO oder POOL=YES

Diskriminanzanalysen mit der SAS-Prozedur DISCRIM

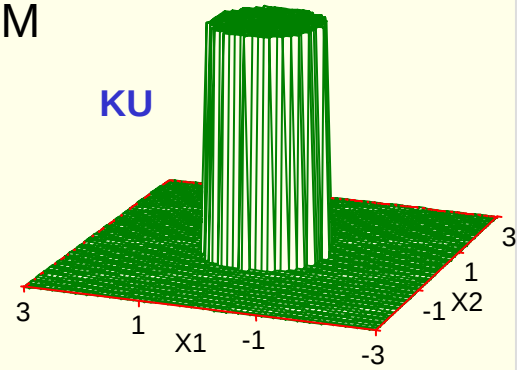
1
2
3
4
5

Kerndichten

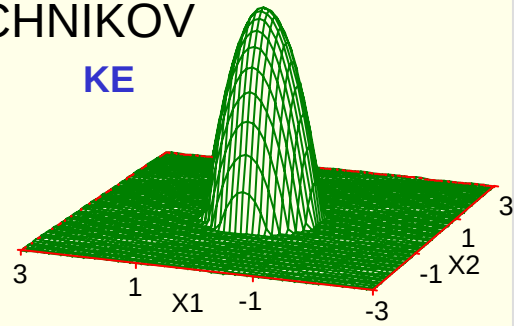
KERNEL=BIWEIGHT



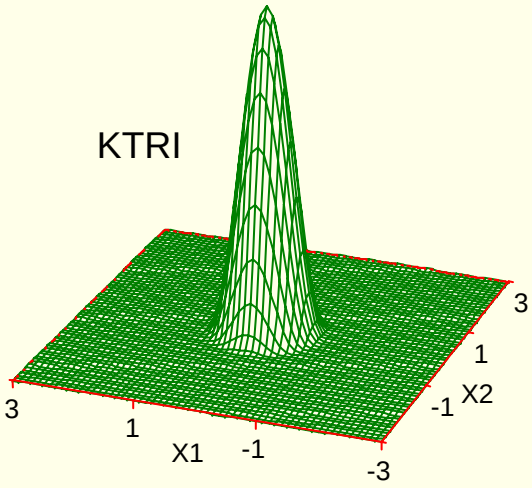
KERNEL=UNIFORM



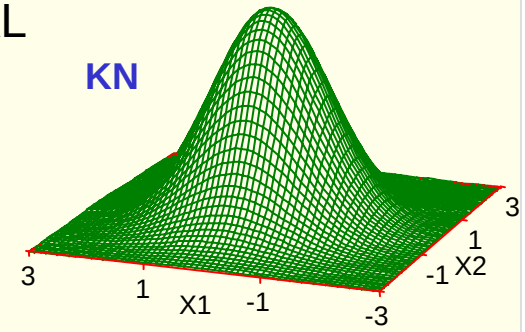
KERNEL=EPANECHNIKOV



KERNEL=TRIWEIGHT



KERNEL=NORMAL



1

2

3 Zufallsvariablenerzeugung

4

5

Zufallsvariablenerzeugung

1	Multivariate Normalverteilungen	Rudolph u.a. (1999)
2		
3	Mischung von Normalverteilungen	
4	Lognormalverteilungen aus Johnsons Translationssystem	
5		

Zufallsvariablenerzeugung

Mischung von Normalverteilungen:

$$\underline{X} \sim p \cdot N_n(\mu_1, \Sigma_1) + (1-p) \cdot N_n(\mu_2, \Sigma_2) \qquad 0 < p < 1$$

Erwartungswertvektor:

$$\mu^* = E(\underline{X}) = p\mu_1 + (1-p)\mu_2$$

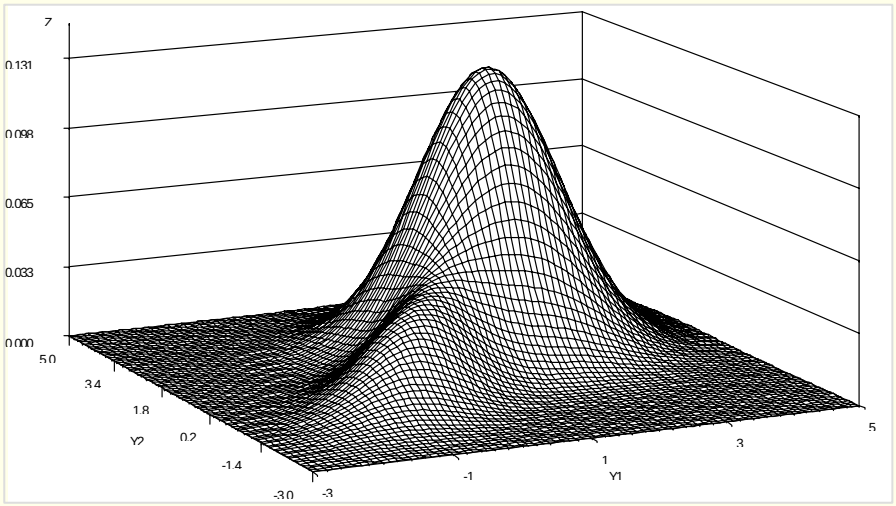
Kovarianzmatrix:

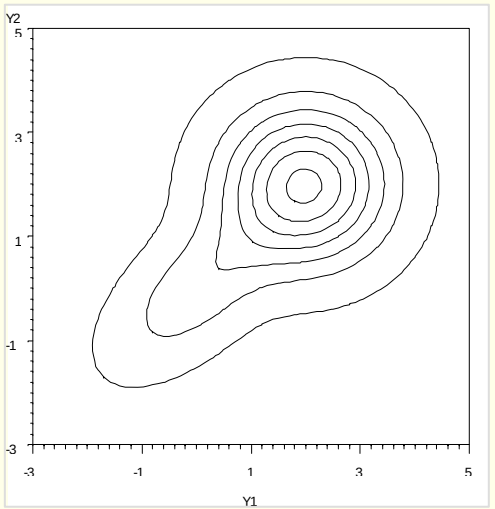
$$\Sigma^* = Cov(\underline{X}) = p\Sigma_1 + (1-p)\Sigma_2 + p(1-p)(\mu_1 - \mu_2)(\mu_1 - \mu_2)'$$

Zufallsvariablenerzeugung

Mischung von Normalverteilungen:

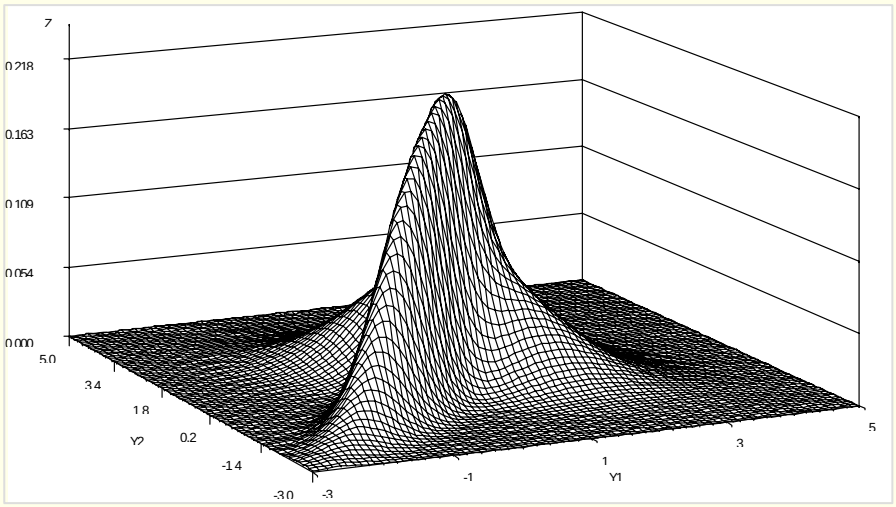
1



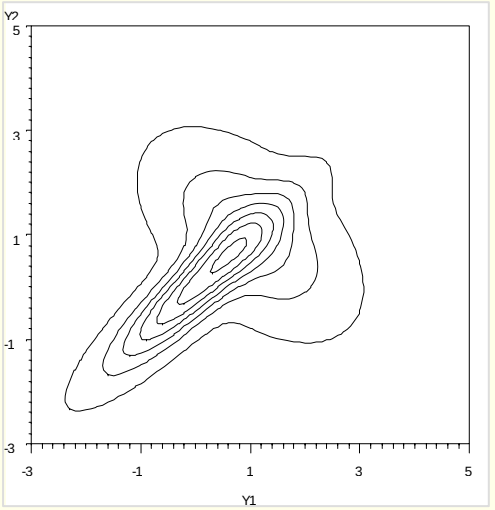


2

3



4



5

Zufallsvariablenerzeugung

Johnsons Translationssystem:

$$\underline{Y} = (\underline{y}_1, \ldots, \underline{y}_n)' \sim N_n(\mu, \Sigma)$$

1 Transformation $\underline{x}_i = a1_i * T_j(\underline{y}_i) + a2_i$

$$\underline{X} = (\underline{x}_1, \ldots, \underline{x}_n)' \qquad a1_i, a2_i \in R \qquad i = 1, \ldots, n$$

2

Transformationen: Kontrolle als Skalen- und Lokationsparameter

3

$$\underline{x}_i = T_N(\underline{y}_i) = \underline{y}_i$$ Normalverteilung (N)

4

$$\underline{x}_i = T_L(\underline{y}_i) = a1_i \exp(\underline{y}_i) + a2_i$$ **Lognormalverteilung (L)**

$$\underline{x}_i = T_U(\underline{y}_i) = a1_i \sinh(\underline{y}_i) + a2_i$$ Sinh^{-1}-Normalverteilung (U)

5

$$\underline{x}_i = T_B(\underline{y}_i) = a1_i (1 + \exp(\underline{y}_i))^{-1} + a2_i$$ Logit-Normalverteilung (B)

Zufallsvariablenerzeugung

Lognormalverteilungen aus Johnsons Translationssystem:

$$\underline{Y} = (\underline{y}_1, \ldots, \underline{y}_n)' \sim N_n(0, \Sigma)$$

1 $\quad \underline{z}_i = \exp(\underline{y}_i) \qquad i = 1, \ldots, n$

2 $\quad E(\underline{z}_i) = \mu_i^z = \exp(\sigma_i^2 / 2) \Rightarrow \mu = 0$

3 $\quad Var(\underline{z}_i) = \sigma_i^{z2} = (\exp(2\sigma_i^2) - \exp(\sigma_i^2))$

4 $\quad \underline{x}_i = a1_i \cdot \dfrac{\underline{z}_i - \mu_i^z}{\sigma_i^z} + a2_i \qquad a1_i, a2_i \in R \qquad i = 1, \ldots, n$

5 $\quad \underline{X} = (\underline{x}_1, \ldots, \underline{x}_n)'; \quad E(\underline{X}) = \mu^*; \quad Cov(\underline{X}) = \Sigma^*$

$$\mu_i^* = a2_i$$
$$\Sigma_{ii}^* = a1_i^2$$

Zufallsvariablenerzeugung

Lognormalverteilungen aus Johnsons Translationssystem:

1 $$\underline{x}_i = a1_i \cdot \frac{\underline{z}_i - \mu_i^z}{\sigma_i^z} + a2_i \qquad a1_i, a2_i \in R \qquad i = 1,\ldots,n$$

2 $$Cor(\underline{x}_i, \underline{x}_j) = Cor(\underline{z}_i, \underline{z}_j)$$

3 $$Cor(\underline{z}_i, \underline{z}_j) = \rho_{ij}^* = \frac{\exp(\rho_{ij}\sigma_i\sigma_j) - 1}{(\exp(\sigma_i^2) - 1)^{1/2} \cdot (\exp(\sigma_j^2) - 1)^{1/2}}$$

4 $$Cor(\underline{y}_i, \underline{y}_j) = \rho_{ij} = \frac{1}{\sigma_i\sigma_j} \cdot \ln\left[1 + \rho_{ij}^* \cdot (\exp(\sigma_i^2) - 1)^{1/2} \cdot (\exp(\sigma_j^2) - 1)^{1/2}\right]$$

5 Σ_{ij}

Zufallsvariablenerzeugung

Lognormalverteilungen aus Johnsons Translationssystem:

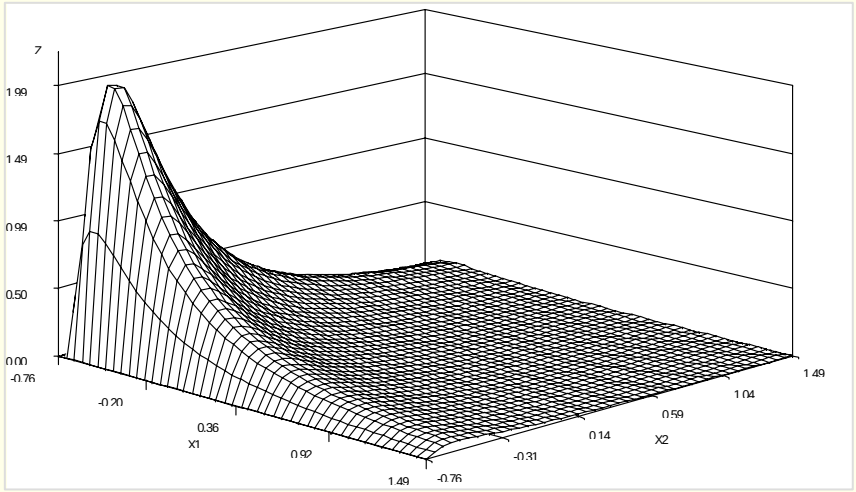
1

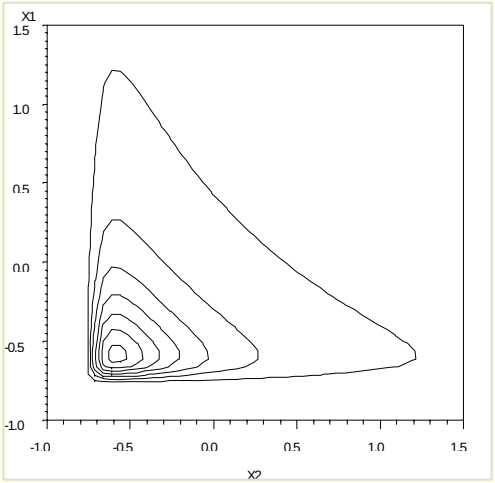
2

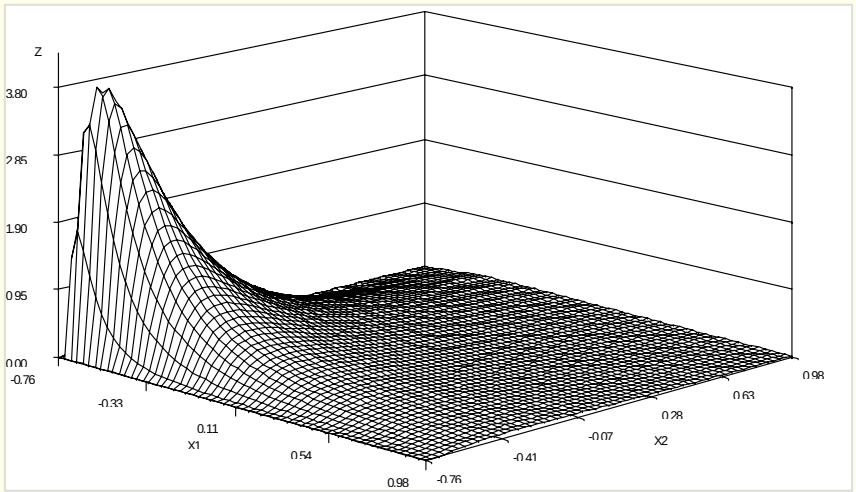
3

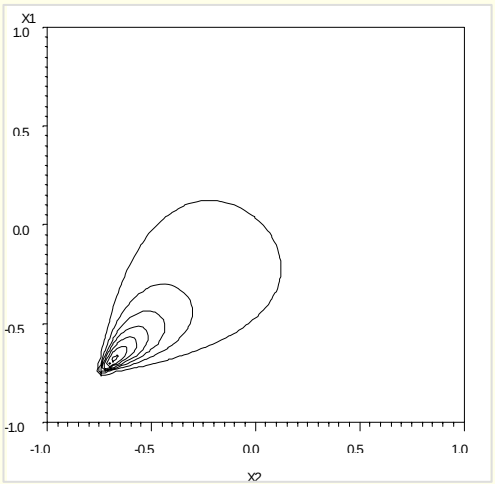
4

5









1

2

3

4 Simulationsexperiment

5

Simulationsexperiment

Bewertung von Diskriminanzanalyseverfahren der SAS-Prozedur DISCRIM bei Nichtnormalität

1

2

3

4

5

Datenerzeugung (Lernstichprobe, Testdaten) für alle Nsim Wiederholungen

PROC DISCRIM mit Nsim als BY-Variable für alle ausgewählten Diskriminanzanalyseverfahren

Zusammenfassung und Auswertung der Fehlklassifikationen

Simulationsexperiment

Diskriminanzanalyseverfahren:

1

Maximum-Likelihood Diskriminanzanalyse (**ML**)
mit der Voraussetzung gleicher Kovarinzmatrizen
in beiden Klassen und gleichen a-priori-Wahrscheinlichkeiten

2

Methode der 'k nächsten Nachbarn' für k=3 (**NN3**)
mit gleicher Kovarinzmatrix in beiden Klassen und gleichen
a-priori-Wahrscheinlichkeiten

3

Kerndichteschätzung
mit gleicher Kovarinzmatrix in beiden Klassen und gleichen
a-priori-Wahrscheinlichkeiten sowie Bandbreite R

4

Gleichverteilungskern: KERNEL=UNI (**KU**)
Normalverteilungskern: KERNEL=NOR (**KN**)

5

Epanechnikow-Kern: KERNEL=EPA (**KE**)

Simulationsexperiment

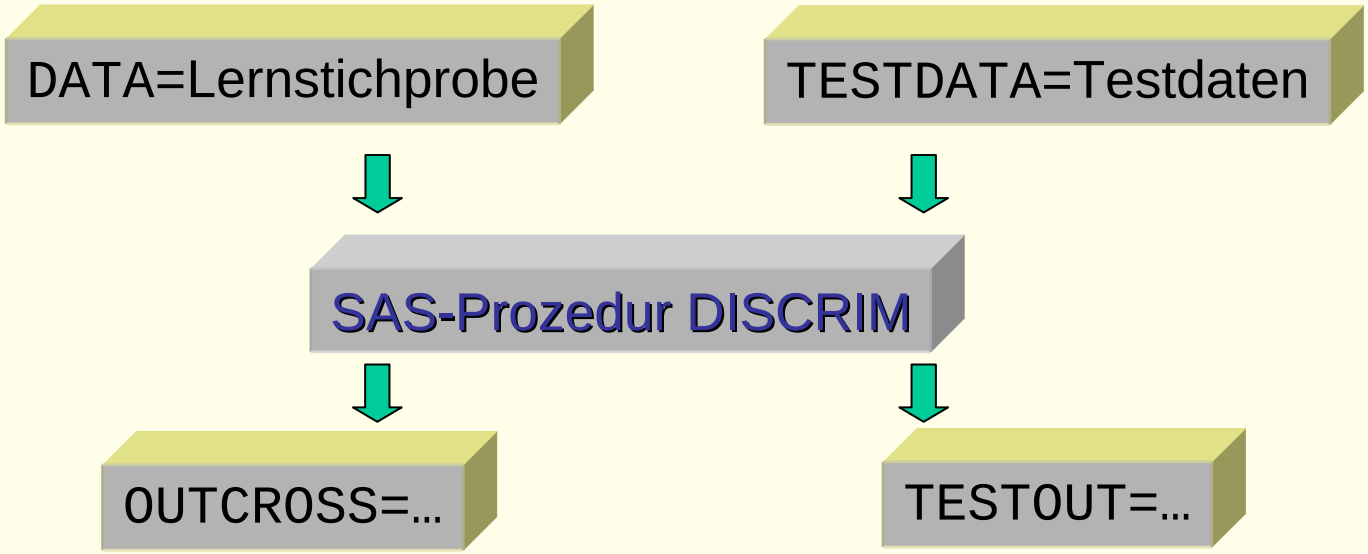
Kerndichteschätzung:

		Bandbreite R		
n	K[i]	UNIFORM	NORMAL	EPANECHNIKOW
3	10	1.130730	0.664390	1.588514
3	20	1.024129	0.601753	1.438754
3	50	0.898475	0.527922	1.262229
3	100	0.813770	0.478152	1.143230

Simulationsexperiment

Beurteilung der Diskriminanzanalyseverfahren anhand ihrer Fehlklassifikationswahrscheinlichkeiten:

1
2
3
4
5



Fehlklassifikation bei Kreuzvalidierung der Lernstichprobe

Fehlklassifikation bei Testdaten

Simulationsexperiment

1

2

3

4

5

Fehlklassifikationswahrscheinlichkeiten

'crossvalidation' der Lernstichproben

$K[i] \in \{10,\ 20,\ 50,\ 100\}$

Klassifizierung aufgrund von Testdaten

$n_T = 1000$

Nsim Wiederholungen

'Daten':

Rudolph u.a. (1999)

$$\mu_1^* = (3 \quad 7 \quad 5)' \qquad \mu_2^* = (2 \quad 8 \quad 4)'$$

$$\Sigma^* = \Sigma_1^* = \Sigma_2^* = \begin{pmatrix} 1.8 & 1.3 & 0.5 \\ 1.3 & 2.1 & 1.2 \\ 0.5 & 1.2 & 2.7 \end{pmatrix}$$

Simulationsexperiment

Mischnormalverteilungen

$$\underline{X}_i \sim p_i \cdot N_3(\mu_{i1}, \Sigma_{i1}) + (1 - p_i) \cdot N_3(\mu_{i2}, \Sigma_{i2}) = VN_3(\mu_i^*, \Sigma^*)$$

1

$$\mu_{11} = \begin{pmatrix} 2.70 \\ 6.85 \\ 4.85 \end{pmatrix} \quad \Sigma_{11} = \begin{pmatrix} 3.15 & 2.58 & 0.43 \\ 2.58 & 3.86 & 2.81 \\ 0.43 & 2.81 & 5.21 \end{pmatrix} \quad \mu_{21} = \begin{pmatrix} -0.70 \\ 7.10 \\ 2.20 \end{pmatrix} \quad \Sigma_{21} = \begin{pmatrix} 0.90 & 2.20 & 2.30 \\ 2.20 & 11.10 & 9.30 \\ 2.30 & 9.30 & 12.60 \end{pmatrix}$$

2

$$p_1 = 0.4 \qquad p_2 = 0.1$$

3

$$\mu_{12} = \begin{pmatrix} 3.20 \\ 7.10 \\ 5.10 \end{pmatrix} \quad \Sigma_{12} = \begin{pmatrix} 0.80 & 0.40 & 0.50 \\ 0.40 & 0.90 & 0.10 \\ 0.43 & 0.10 & 1.00 \end{pmatrix} \quad \mu_{22} = \begin{pmatrix} 2.30 \\ 8.10 \\ 4.20 \end{pmatrix} \quad \Sigma_{22} = \begin{pmatrix} 1.00 & 0.90 & -0.30 \\ 0.90 & 1.00 & 0.10 \\ -0.30 & 0.10 & 1.20 \end{pmatrix}$$

4

$$\mu_1^* = \begin{pmatrix} 3 \\ 7 \\ 5 \end{pmatrix} \quad \Sigma^* = \begin{pmatrix} 1.8 & 1.3 & 0.5 \\ 1.3 & 2.1 & 1.2 \\ 0.5 & 1.2 & 2.7 \end{pmatrix} \quad \mu_2^* = \begin{pmatrix} 2 \\ 8 \\ 4 \end{pmatrix}$$

5

Simulationsexperiment

1

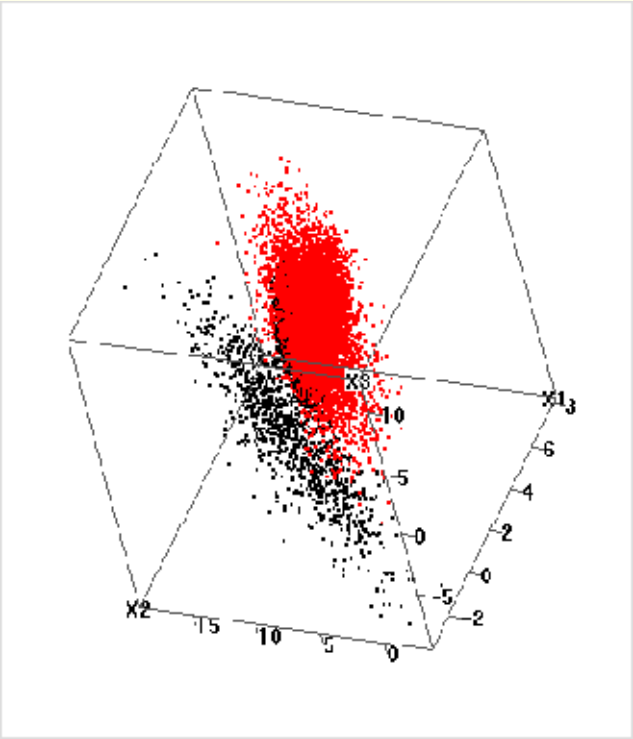
2

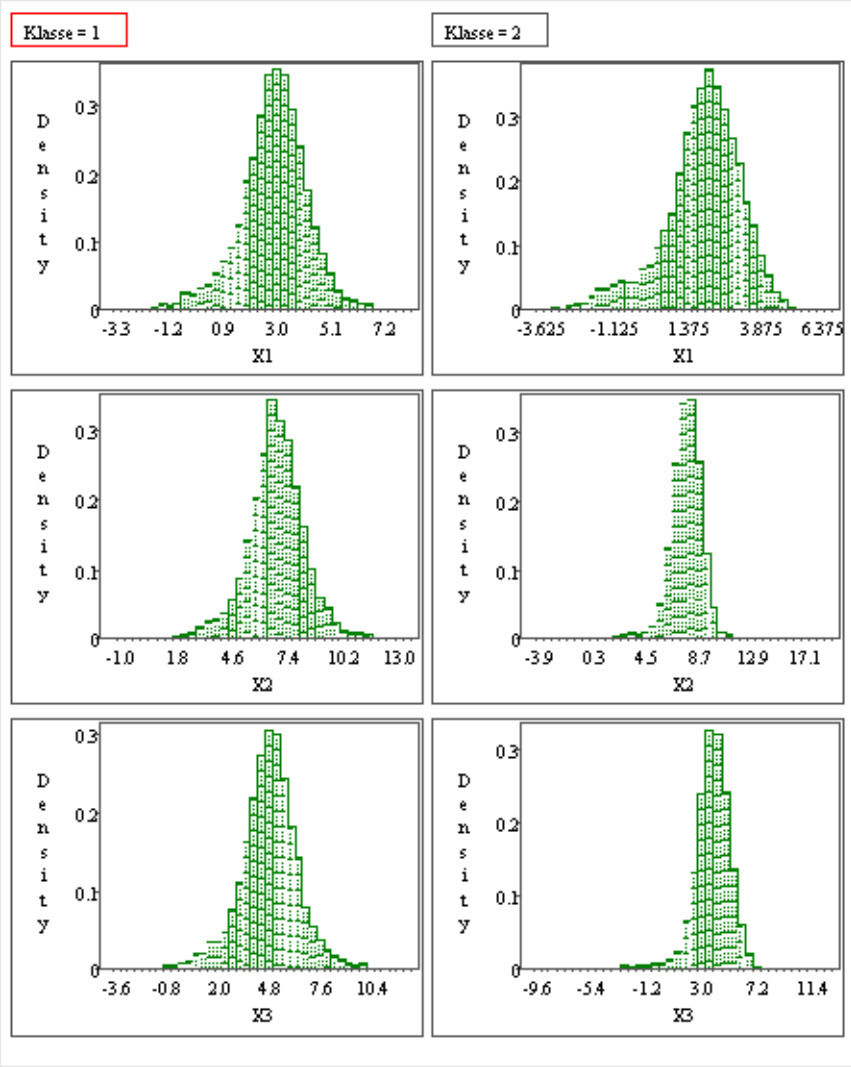
3

4

5

Mischnormalverteilung





SAS/INSIGHT

Simulationsexperiment

Lognormalverteiungen (Johnsons Translationssystem)

1

$$\underline{X}_i \sim LGN_3(\mu_i^*, \Sigma^*) \qquad \underline{x}_{ij} = a1^{1/2} \cdot \frac{\exp(\underline{y}_{ij}) - \exp(\sigma_j^2/2)}{(\exp(2\sigma_j^2) - \exp(\sigma_j^2))^{1/2}} + a2_i \,, \quad \underline{Y}_i \sim N_3(0, \Sigma)$$

2

$$\Sigma = \begin{pmatrix} 1.0 & 1.3 & 0.0 \\ 1.3 & 1.0 & 1.2 \\ 0.0 & 1.2 & 1.0 \end{pmatrix}$$

3

$$a2_1 = \begin{pmatrix} 3 \\ 7 \\ 5 \end{pmatrix} \qquad a1 = \begin{pmatrix} 1.8 \\ 2.1 \\ 2.7 \end{pmatrix} \qquad a2_2 = \begin{pmatrix} 2 \\ 8 \\ 4 \end{pmatrix}$$

4

$$\mu_1^* = \begin{pmatrix} 3 \\ 7 \\ 5 \end{pmatrix} \qquad \Sigma^* = \begin{pmatrix} 1.8 & 1.3 & 0.5 \\ 1.3 & 2.1 & 1.2 \\ 0.5 & 1.2 & 2.7 \end{pmatrix} \qquad \mu_2^* = \begin{pmatrix} 2 \\ 8 \\ 4 \end{pmatrix}$$

5

Simulationsexperiment

1

2

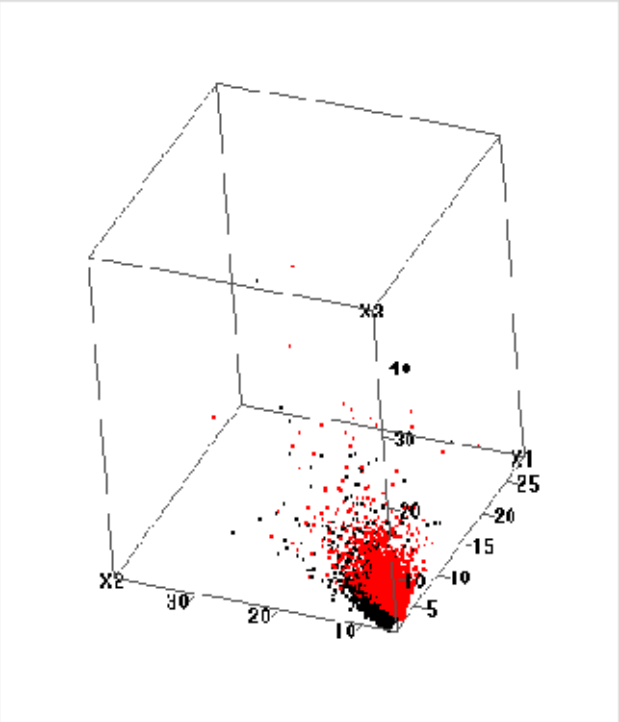
3

4

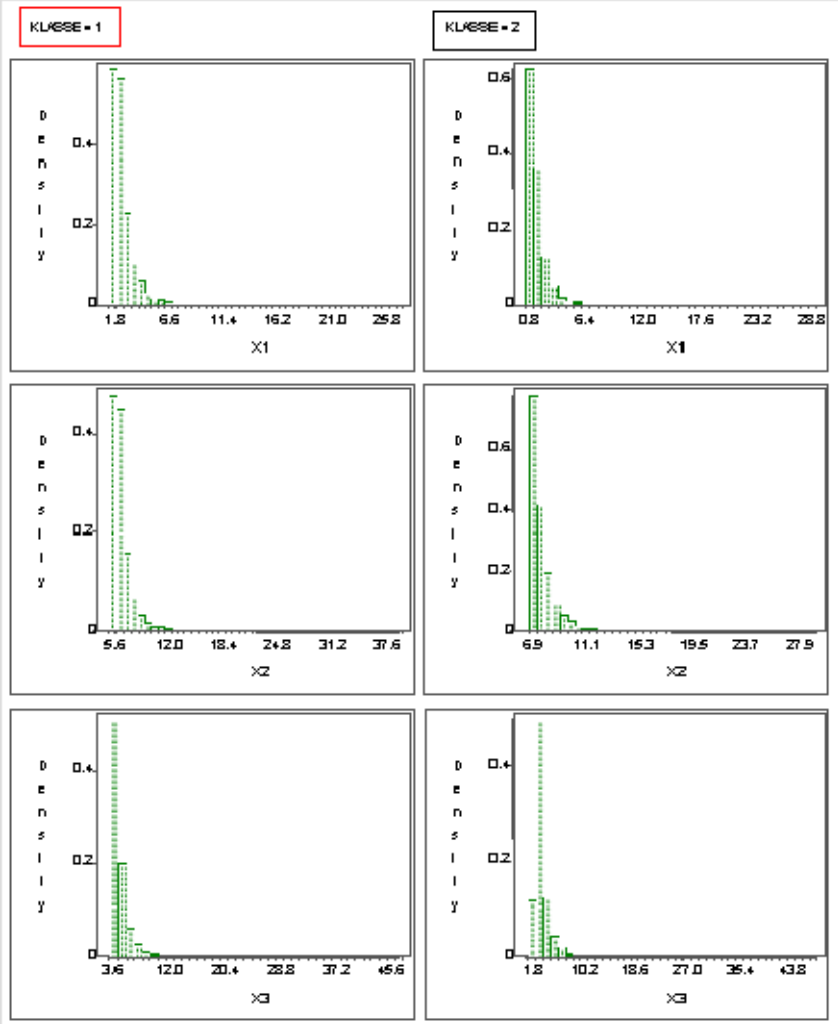
5

Lognormalverteiung

(Johnsons Translationssystem)



SAS/INSIGHT



Simulationsexperiment

Datenerzeugung mit SAS (z.B. Lognormalverteilung):

1
```
Aufruf des Makros JOHNSON_LGN zur Erzeugung der
Datei LSP3 mit jeweils 10 Datensätzen je Klasse mit
Nsim = 10 Wiederholungen
```

2
```
%JOHNSON_LGN(10,Klassen,Mittel,KOVAR,NCOVAR,
             LSP3,{'NSIM' 'KLASSE' 'X1' 'X2' 'X3'});
```

3

4
```
Aufruf des Makros JOHNSON_LGN zur Erzeugung der
Datei SIM mit jeweils 1000 Testdaten je Klasse mit
Nsim = 10 Wiederholungen
```

5
```
%JOHNSON_LGN(10000,Klassen,Mittel,KOVAR,NCOVAR,
             SIM,{'NSIM' 'KLASSE' 'X1' 'X2' 'X3'});
```

Simulationsexperiment

Datenerzeugung mit SAS (z.B. Lognormalverteilung):

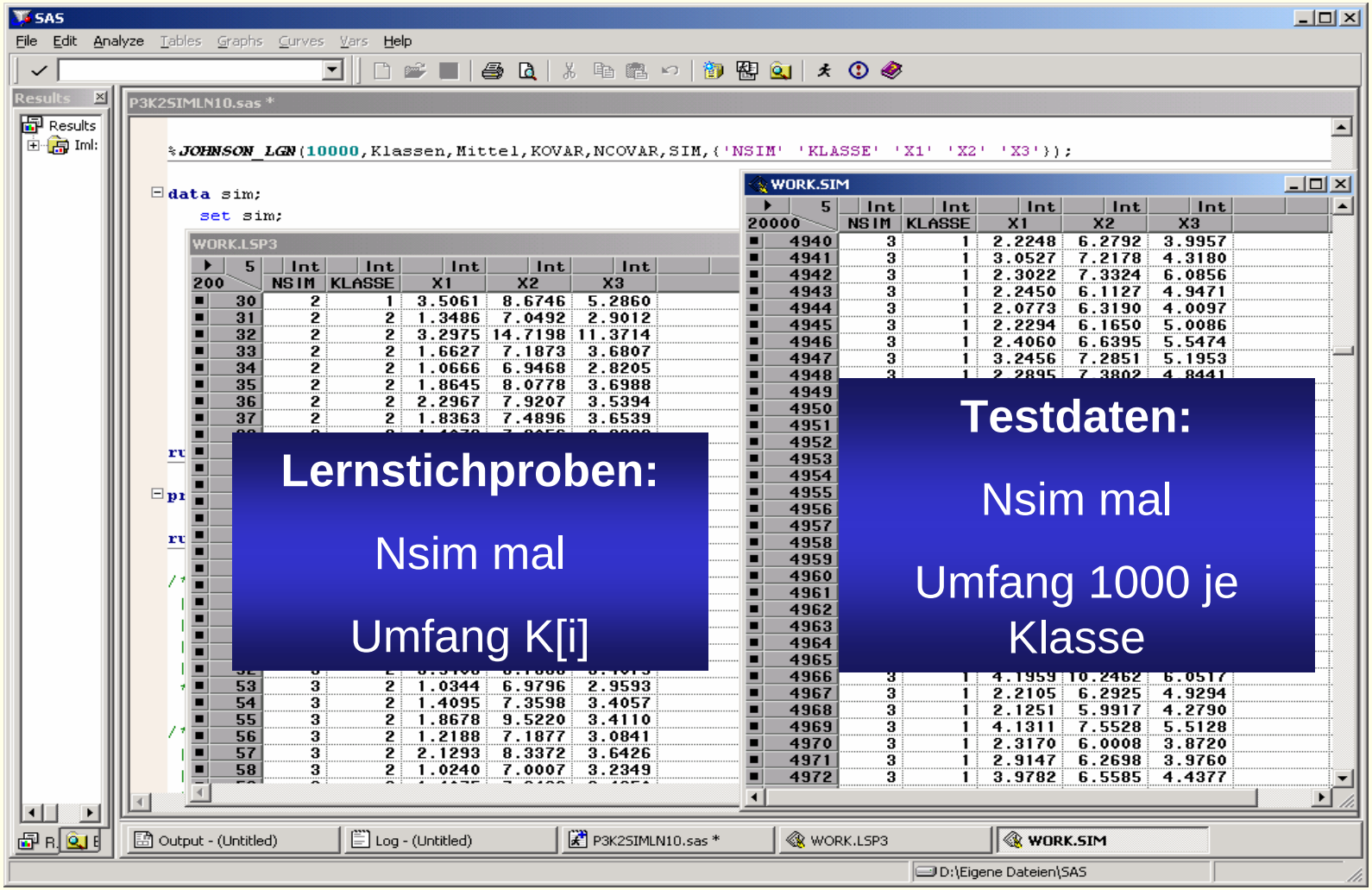
1

2

3

4

5



Simulationsexperiment

SAS-Statements (Beispiel):

```
1  /*-------------------------------------------------------*
   |                                                       |
2  |  Diskriminanzanalyse mit Kerndichteschätzung          |
   |  Kernel=normal ;K[i]=10                               |
   |                                                       |
   *-------------------------------------------------------*/
3  proc discrim data=lsp3  testdata=sim
               outcross=crosknrp testout=sknrp
               method=npar r=0.664390 kernel=nor
               pool=yes
               noprint;
4     class  Klasse;
      var    x1-x3;
      priors equal;
      by Nsim;
5  run;
```

1

2

3

4

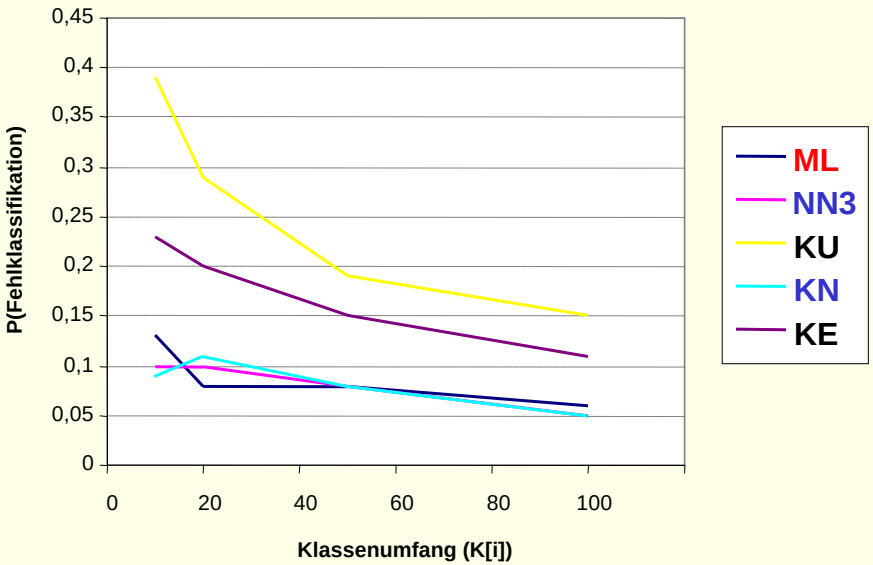
5 Simulationsergebnisse und Diskussion

Simulationsergebnisse und Diskussion

Mischnormalverteilung - Fehlklassifikationswahrscheinlichkeiten

1
2
3
4
5

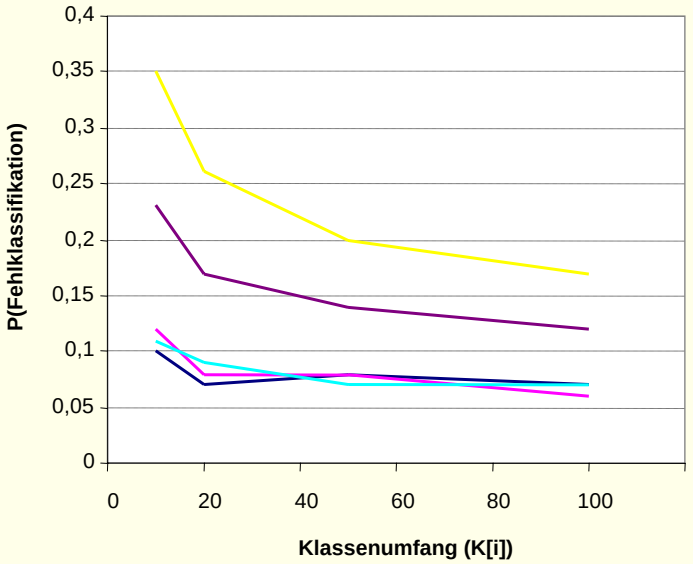
Mittlere Fehlklassifikationsrate Lernstichprobe



K[i]	ML	NN3	KU	KN	KE
10	0.40	0.40	1.00	0.40	0.60
20	0.25	0.35	0.65	0.35	0.50
50	0.22	0.20	0.40	0.24	0.32
100	0.15	0.11	0.27	0.13	0.19

MAX

Mittlere Fehlklassifikationsrate Testdaten



K[i]	ML	NN3	KU	KN	KE
10	0.21	0.22	0.63	0.28	0.49
20	0.21	0.18	0.43	0.20	0.28
50	0.19	0.16	0.33	0.19	0.25
100	0.16	0.13	0.30	0.15	0.21

Simulationsergebnisse und Diskussion

Lognormalverteilung - Fehlklassifikationswahrscheinlichkeiten

1

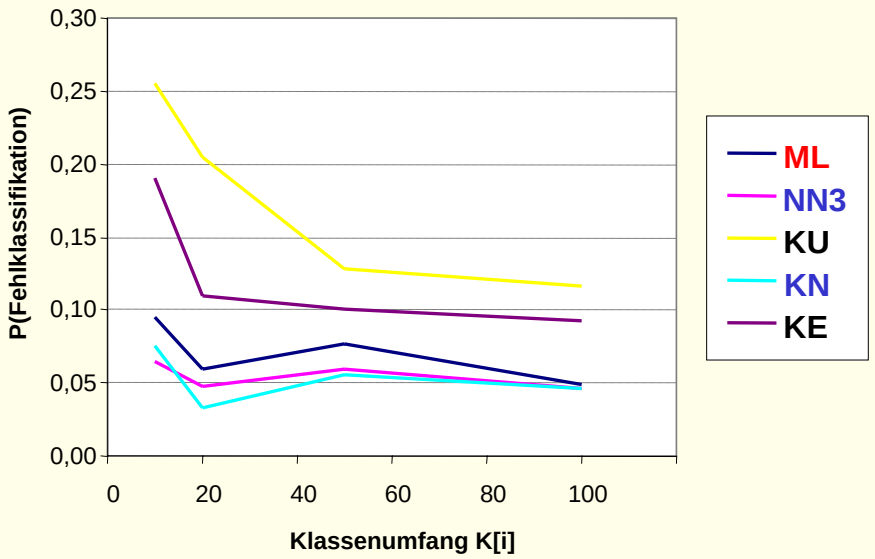
2

3

4

5

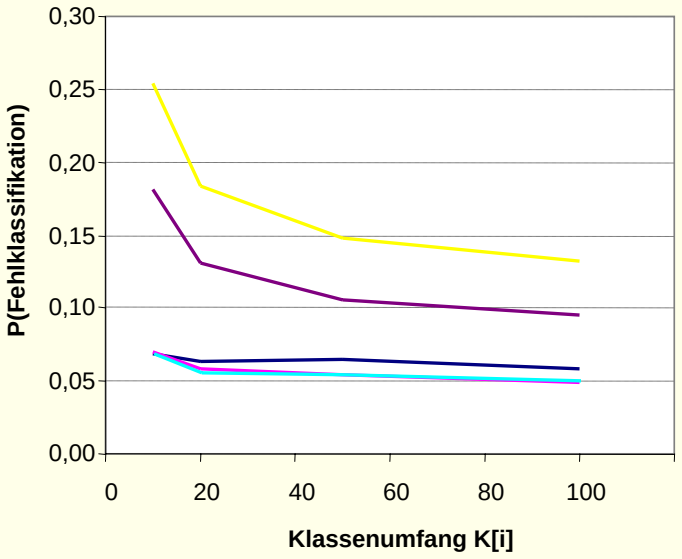
Mittlere Fehlklassifikationsrate Lernstichprobe



K[i]	ML	NN3	KU	KN	KE
10	0.20	0.30	0.60	0.20	0.60
20	0.25	0.15	0.40	0.10	0.25
50	0.18	0.12	0.24	0.12	0.16
100	0.08	0.08	0.16	0.08	0.14

MAX

Mittlere Fehlklassifikationsrate Testdaten



K[i]	ML	NN3	KU	KN	KE
10	0.15	0.17	0.39	0.17	0.30
20	0.13	0.13	0.26	0.12	0.19
50	0.13	0.12	0.19	0.13	0.16
100	0.09	0.07	0.18	0.08	0.14

Simulationsergebnisse und Diskussion

Bei näherungsweise multivariater Normalverteilung:

ML

Rudolph u.a. (1999)

1

Bei geringer Abweichung von der multivariaten Normalverteilung:

ML, **NN3** und **KN**

2

Bei stärkerer Abweichung von der multivariaten Normalverteilung:

NN3 und **KN**

3

Tuchscherer u.a. (2002)

Ein 'gleichmäßig bestes' Verfahren gibt es nicht.

4

Sehr kleine Lernstichproben sind problematisch.

Bei Unsicherheit zur multivariaten Verteilung empfehlen wir:

5

NN3

Literatur:

1. Huberty, C. J. (1994). *Applied Discriminant Analysis.* John Wiley & Sons. Inc., New York.

2. Johnson, M. E. (1987). *Multivariate Statistical Simulation.* J. Wiley, New York.

3. Johnson, N. L.; Kotz, S. (1972). *Distributions in Statistics: Continuous Multivariate Distributions.* J. Wiley, New York.

4. Kleijnen; J.; van Groenendaal, W. (1992). *Simulation: A Statistical Perspective.* J. Wiley, Chichester.

5. SAS Institute Inc. (1999). *SAS/STAT User's Guide, Version 8,* Cary, NC: SAS Institute Inc.

6. Sumpf, D.; Rudolph, P. E.; Biebler, K.-E.; Jäger, B. (1997). *Faktoren- und Diskriminanzanalyse mit SAS.* GinkgoPark Mediengesellschaft, Gützkow.

7. Tuchscherer, A.; Rudolph, P. E.; Jäger, B.; Tuchscherer, M. (1999). Ein SAS-Makro zur Erzeugung multivariat normalverteilter Zufallsgrößen. In *Proceedings der 3. Konferenz der SAS-Anwender in Forschung und Entwicklung,* Ed. Ortseifen, Heidelberg, S. 293-306.

8. Tuchscherer, A.; Rudolph, P. E.; Jäger, B.; Tuchscherer, M. (2000). Erzeugung nichtnormaler multivariater Zufallsgrößen mit SAS. In *Proceedings der 4. Konferenz der SAS-Anwender in Forschung und Entwicklung*, Eds. Bödecker, R.-H.; Hollenhorst, M. S., Gießen, 235-265.

9. Rudolph, P. E.; Tuchscherer, A.; Jäger, B.; Biebler, K.-E. (1999). Beurteilung von Diskriminanzanalyseverfahren in und mit SAS. In *Proceedings der 3. Konferenz der SAS-Anwender in Forschung und Entwicklung,* Ed. Ortseifen, Heidelberg, S. 245-258.