Bewertung von Diskriminanzanalyseverfahren in SAS bei Nichtnormalität

Armin Tuchscherer, Paul Eberhard Rudolph, Margret Tuchscherer
Forschungsinstitut für die Biologie
landwirtschaftlicher Nutztiere Dummerstorf
Wilhelm-Stahl-Allee 2
18196 Dummerstorf
atuchsch@fbn-dummerstorf.de
rudolph@fbn-dummerstorf.de
mtuchsch@fbn-dummerstorf.de

Bernd Jäger
Institut für Biometrie und Medizinische Informatik
Ernst-Moritz-Arndt-Universität Greifswald
Rathenaustr. 48
17489 Greifswald
bjaeger@biometrie.uni-greifswald.de

Zusammenfassung

Das Ziel einer Diskriminanzanalyse ist die Zuordnung von Objekten mit unbekannter Klassenzugehörigkeit zu vorgegebenen Klassen, für die eine Lernstichprobe existiert mit Objekten, deren Klassenzugehörigkeit bekannt ist. Die SAS Prozedur DISCRIM bietet dazu eine Reihe von parametrischen und nichtparametrischen Verfahren an. Für den Anwender ist es jedoch immer wieder schwierig, aus dem angebotenen Spektrum parametrischer und nichtparametrischer Verfahren das für seine Daten passende Verfahren herauszufinden. Eine solche Auswahl wird insbesondere auch dadurch erschwert, dass eine Überprüfung der für die jeweiligen Verfahren unterstellten Voraussetzungen häufig schwierig ist. Eine Möglichkeit zur Auswahl eines geeigneten Verfahrens ist die mittels Simulation durchführbare Ermittlung der Fehlklassifikati-

onswahrscheinlichkeiten bei einer Neuzuordnung (Reklassifizierung) der Objekte der Lernstichprobe, das die Prozedur DISCRIM anbietet. Im Beitrag werden die Verfahren unter Verwendung eines SAS-Makros zur Erzeugung multivariater nichtnormalverteilter Zufallsgrößen (Tuchscherer u.a., 2000, [8]) mittels Simulation im Hinblick auf die bestimmten Fehlklassifikationswahrscheinlichkeiten bewertet. Daraus werden Empfehlungen zur Nutzung der untersuchten Diskriminanzanalyseverfahren abgeleitet.

Keywords: Diskriminanzanalyse, SAS Prozedur DISCRIM, Nichtnormalität, Fehlklassifikationswahrscheinlichkeit, Simulation.

1 Einleitung

Das SAS-System enthält in SAS/STAT mit den Prozeduren CANDISC, DISCRIM und STEPDISC eine Reihe von Werkzeugen zur praktischen Durchführung von Diskriminanzanalysen. Für den Anwender ist es jedoch immer wieder schwierig, aus dem angebotenen Spektrum parametrischer und nichtparametrischer Verfahren das für seine Daten passende Verfahren herauszufinden. Insbesondere für kleine Stichprobenumfänge in der Lernstichprobe ist häufig nicht bekannt, ob parametrische oder nichtparametrische Verfahren die besseren Klassifizierungsergebnisse liefern.

In der Regel ist das Ziel einer Diskriminanzanalyse die Zuordnung von Objekten mit unbekannter Klassenzugehörigkeit zu vorgegebenen Klassen, für die eine Lernstichprobe existiert mit Objekten, deren Klassenzugehörigkeit bekannt ist.

Die SAS-Prozedur DISCRIM bietet als einzige der drei in SAS verfügbaren Prozeduren zur Diskriminanzanalyse die Möglichkeit, aufgrund einer Lernstichprobe Objekte mit unbekannter Klassenzugehörigkeit den bekannten Klassen zuzuordnen. Dabei stehen parametrische und nichtparametrische Verfahren zur Verfügung. Eine Möglichkeit zur Auswahl eines geeigneten Verfahrens ist die Betrachtung der Fehlklassifikationswahrscheinlichkeiten bei einer Neuzuordnung der Objekte der Lernstichprobe z. B. durch CROSSVALIDATION, das die Prozedur DISCRIM anbietet.

In Rudolph u.a. [9] wurden mittels Simulation multivariat normalverteilter Zufallsgrößen bereits die Eigenschaften einiger in SAS verfügbarer Diskriminanzanalyseverfahren für verschiedene Lernstichprobenumfänge untersucht. In diesem Beitrag werden diese Untersuchungen auf nichtnormalverteilte Zufallsvariablen ausgeweitet. Dabei kommen die Makros von Tuchscherer u.a. [8] zur Erzeugung von gemischten multivariaten Normalverteilungen und multivariaten Verteilungen des Translationssystems von Johnson zur Anwendung.

2 Diskriminanzanalysen mit der SAS-Prozedur DISCRIM

In der **Diskriminanzanalyse** geht man davon aus, dass eine Gesamtheit (Menge von Objekten) aus $n_K \geq 2$ Teilgesamtheiten (Klassen, Gruppen) besteht, wobei jedes Element (Objekt) der Gesamtheit genau einer Teilgesamtheit angehört. Über die Objekte der Gesamtheit sei bekannt, dass sie sich nur sinnvoll durch die gleichzeitige Betrachtung von $n \geq 2$ Merkmalen, die an den Objekten messbar seien, unterscheiden lassen.

Es sollen nun Messwerte der n Merkmale für Objekte, für die bekannt ist, zu welcher Teilgesamtheit sie gehören, aus allen Teilgesamtheiten vorliegen. Das bedeutet, für jedes derartige Objekt sind sein n-dimensionaler Merkmalsvektor $x \in \mathbb{R}^n$ sowie seine Klassenzugehörigkeit k bekannt. Die Menge der Merkmalsvektoren von Objekten mit bekannter Gruppenzugehörigkeit wird häufig auch als **Lernstichprobe** bezeichnet. Der statistische Hintergrund besteht darin, dass man x als Realisierung einer n-dimensionalen Zufallsgröße $\underline{x}' = (\underline{x}_1, \ldots, \underline{x}_n)$ auffassen kann.

Das Ziel der Diskriminanzanalyse besteht dann in folgender Aufgabe:
Ein Objekt mit dem Beobachtungsvektor x, dessen Klassenzugehörigkeit unbekannt ist, soll einer der vorliegenden Gruppen zugeordnet werden.

Die Lösung des Zuordnungsproblems (Klassifizierung) erfolgt mit geeigneten Entscheidungsregeln, die auf der Bestimmung einer oder mehrerer optimaler **Trennfunktionen (Diskriminanzfunktionen)** basieren. (Die Anzahl der benötigten Trennfunktionen richtet sich nach der Anzahl der Teilgesamtheiten.) Unterschiedliche Vorgehensweisen führen zu unterschiedlichen Diskriminanzfunktionen und daraus abgeleiteten Zuordnungsregeln. Auf Einzelheiten soll hier nicht eingegangen werden.

Mit der SAS-Prozedur DISCRIM steht dem Anwender ein umfangreiches Werkzeug zur Durchführung von Diskriminanzanalysen zur Verfügung, insbesondere zur Klassifizierung von Objekten mit unbekannter Klassenzugehörigkeit mittels einer Lernstichprobe und der Beobachtungsvektoren der zu klassifizierenden Objekte.

Neben parametrischen Verfahren, die eine multivariate Normalverteilung unterstellen, gibt es in der Prozedur DISCRIM auch parameterfreie Verfahren, die keine spezielle Verteilung der Beobachtungsvektoren voraussetzen. Diese parameterfreien Verfahren basieren auf nichtparametrischen Schätzungen der klassenspezifischen Wahrscheinlichkeitsdichten. Diese Dichteschätzungen können auf der Basis der Bestimmung der „k nächsten Nachbarn“ oder mit Kerndichteschätzungen gewonnen werden. Während bei der „k nächste Nachbarn“-Methode die Anzahl der Nachbarn k wählbar ist, kann bei den Kernschätzungen neben 5 verschiedenen Kernen auch ein Parameter R gewählt werden.

Der mit den Details und den Voraussetzungen der Diskriminanzanalyseverfahren nicht so vertraute Anwender wird häufig vor dem Problem stehen, welches Verfahren für seine Daten, die häufig nicht normal verteilt sind, zu wählen ist.

Bevor wir das Simulationsexperiment beschreiben, ist es zunächst erforderlich, kurz auf die Erzeugung der nichtnormalen n-dimensionalen Zufallsvariablen mit abhängigen Komponenten einzugehen, die im folgenden Anwendung finden.

3 Zufallsvariablenerzeugung

Während die multivariate Normalverteilung durch den Erwartungswertvektor und die Kovarianzmatrix bereits vollständig beschrieben ist, kann man bei nichtnormalen multivariaten Verteilungen beliebig viele mit gleichem Erwartungswertvektor und gleicher Kovarianzmatrix finden. Wir werden uns hier auf die Mischung von Normalverteilungen und Johnsons Translationssystem beschränken. Mit $\mu_i^*, i = 1, \ldots, g$ wollen wir den Erwartungswertvektor der i-ten Klasse bezeichnen und Σ^* sei die allen Klassen gemeinsame Kovarianzmatrix.

Diese Größen sollen wie in Rudolph u.a. [9] im Simulationsexperiment vorgegeben werden. Es sind also entsprechende nichtnormale Verteilungen mit Erwartungswertvektor μ_i^* und Kovarianzmatrix Σ^* zu erzeugen.

3.1 Mischung von Normalverteilungen

Gemischte Normalverteilungen, in der Robustheitsliteratur als verschmutzte Normalverteilungen mit einer sehr breiten Anwendung bekannt, entstehen durch Mischen zweier n-dimensionaler Normalverteilungen mit Wahrscheinlichkeit $p \quad (0 < p < 1) : \underline{X} \sim p \cdot N_n(\mu_1, \Sigma_1) + (1-p) \cdot N_n(\mu_2, \Sigma_2)$.

Der Erwartungswert von $\underline{X}$ ist $\mu^* = E(\underline{X}) = p\mu_1 + (1-p)\mu_2$ und die Kovarianzmatrix $\Sigma^* = \mathrm{Cov}(\underline{X}) = p\Sigma_1 + (1-p)\Sigma_2 + p(1-p)(\mu_1 - \mu_2)(\mu_1 - \mu_2)'$.

Um derartige Verteilungen mit Erwartungswertvektoren μ_i^* und gemeinsamer Kovarianzmatrix Σ^* mit dem Makro von Tuchscherer u.a. [8] erzeugen zu können, sind für jede Klasse die Wahrscheinlichkeit p und eine Ausgangsnormalverteilung derart vorzugeben, dass sich die zweite aus den obigen Beziehungen ergibt.

3.2 Johnsons Translationssystem

Sei $\underline{Y} = (\underline{y}_1, \ldots, \underline{y}_n)' \sim N_n(\mu, \Sigma)$. Wendet man auf die Komponenten von $\underline{Y}$ eine der folgenden Transformationen (Tabelle 1) an, so hat der resultierende Vektor $\underline{X} = (\underline{x}_1, \ldots, \underline{x}_n)'$ eine Verteilung im multivariaten Johnsonsystem. Die Parameter $a1_i, a2_i \in \mathbb{R}, i = 1, \ldots, n$, wurden zur Kontrolle als Skalen- und Lokationsparameter der $\underline{x}_i$ eingeführt.

Tabelle 1: Transformationen in Johnsons Translationssystem

Transformation $T_j(\underline{y}); j \in \{N, L, U, B\}$	Bezeichnung
$\underline{x}_i = T_N(\underline{y}_i) = \underline{y}_i$	Normalverteilung (N)
$\underline{x}_i = T_L(\underline{y}_i) = a1_i \exp(\underline{y}_i) + a2_i$	Lognormalverteilung (L)
$\underline{x}_i = T_U(\underline{y}_i) = a1_i \sinh(\underline{y}_i) + a2_i$	Sinh^{-1}-Normalverteilung (U)
$\underline{x}_i = T_B(\underline{y}_i) = a1_i(1 + \exp(\underline{y}_i))^{-1} + a2_i$	Logit-Normalverteilung (B)

Die einfache Form der Transformationsfunktionen führt sofort und sehr einfach zum Erzeugungsalgorithmus für Verteilungen des multivariaten Johnsonsystems. Hierbei ist noch zu erwähnen, dass auf jede Komponente von $\underline{Y}$ eine beliebige Transformation T_j aus der obigen Tabelle angewendet werden kann.

Die einzige Schwierigkeit bei der praktischen Anwendung des Johnsonsystems besteht in der Spezifikation der Verteilungsparameter der Ausgangsverteilung. Für Simulationsuntersuchungen ist es notwendig, aus den Parametern einer Johnsonverteilung die Parameter der Ausgangsnormalverteilung zu bestimmen. Wie man dabei vorgeht, soll hier nur am Beispiel der Lognormalverteilung erläutert werden. Ansonsten sei auf Johnson (1987), [2] verwiesen.

Angenommen $\underline{Y} = (\underline{y}_1, \ldots, \underline{y}_n)' \sim N_n(\mu, \Sigma)$ sei die Ausgangsnormalverteilung. Im Falle der Lognormalverteilung können wir $\mu = 0$ setzen und erhalten für die Parameter der Johnsonverteilung von $\underline{z}_i = \exp(\underline{y}_i), i = 1, \ldots, n$:

$$\mathrm{E}(\underline{z}_i) = \mu_i^z = \exp(\sigma_i^2/2), \mathrm{Var}(\underline{z}_i) = \sigma_i^{z2} = (\exp(2\sigma_i^2) - \exp(\sigma_i^2))$$

und

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

Mit $\underline{x}_i = a1_i \cdot \dfrac{\underline{z}_i - \mu_i^z}{\sigma_i^z} + a2_i$ kann man nun lognormalverteilte Zufallsgrößen mit Mittelwert $a2_i$ und Varianz $a1_i^2$ erzeugen.

Aus der Beziehung $\mathrm{Cor}(\underline{z}_i, \underline{z}_j) = \mathrm{Cor}(\underline{x}_i, \underline{x}_j)$ und

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

kann man nun die Kovarianzmatrix Σ der Ausgangsverteilung bestimmen. Dabei ist zu testen, ob alle $|\rho_{ij}| \leq 1$ sind und im Falle von $n > 2$ ist auch die positive Definitheit von Σ zu prüfen.

4 Simulationsexperiment

Mittels Simulation multivariat nichtnormalverteilter Lernstichproben und Testdaten sollen die Fehlklassifikationswahrscheinlichkeiten bei 'crossvalidation' der Lernstichproben und bei der Klassifizierung aufgrund von Testdaten für verschiedene Diskriminanzanalyseverfahren bei unterschiedlichen Umfängen in der Lernstichprobe ermittelt werden.

Die Dimension der Beobachtungsvektoren sei $n = 3$ in $n_k = 2$ Klassen mit in den Klassen jeweils gleichem Umfang $K[i] = k$, $k \in \{10, 20, 50, 100\}$ für die entsprechende Lernstichprobe. Die Anzahl der Testdatensätze wurde $n_t = 1000$ je Klasse und simulierte Wiederholung einer Lernstichprobe gewählt. Dieser Prozess wird $Nsim = 10$ mal wiederholt.

Wegen der Vergleichbarkeit mit den Ergebnissen von Rudolph u.a. [9] wählen wir die Erwartungswertvektoren $\mu_1^* = (3\ \ 7\ \ 5)'$, $\mu_2^* = (2\ \ 8\ \ 4)'$ für die Klassen 1 und 2 und die für beide Klassen identische Kovarianzmatrix

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

Mit den Makros von Tuchscherer u.a. [8] sind dann die Daten zu erzeugen, die in diesem Beitrag am Beispiel einer gemischten Normalverteilung sowie einer Lognormalverteilung aus Johnsons Translationssystem zur Schätzung der entsprechenden Fehlklassifikationswahrscheinlichkeiten für crossvalidation der Lernstichproben und die Testdaten folgender Diskriminanzanalyseverfahren (in Klammern: Abkürzung der Verfahren) verwendet werden:

Parametrisches Verfahren:

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

Parameterfreie Verfahren:

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

- **Kerndichteschätzung** mit gleicher Kovarianzmatrix in beiden Klassen und gleichen a-priori-Wahrscheinlichkeiten sowie Bandbreite R nach Tabelle 2 mit

Gleichverteilungskern:	KERNEL=UNI (**KU**),
Normalverteilungskern:	KERNEL=NOR (**KN**),
Epanechnikow-Kern:	KERNEL=EPA (**KE**).

Tabelle 2: Bandbreite R für verschiedene Kerne in Abhängigkeit von n und $K[i]$

		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

Die SAS-Statements für die Durchführung der $Nsim$ Diskriminanzanalysen für die Kerndichteschätzung mit Normalverteilungskern und Klassenumfang der Lernstichprobe von je 10 hätten dann z.B. folgende Gestalt:

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

Die temporäre SAS-Datei 'lsp3' enthält dabei jeweils die Daten von $Nsim$ Lernstichproben mit Klassenumfang $K[i]$ und die temporäre SAS-Datei 'sim' enthält die $Nsim$ Testdaten des Umfangs 1000 je Lernstichprobe.

In die temporären SAS-Dateien outcross='cros....' bzw. testout='s....' werden die Fehlklassifikationen der Kreuzvalidierung der Lernstichproben bzw. der Testdaten geschrieben, die anschließend zusammengefasst und ausgewertet werden. Die Dateien 'lsp3' und 'sim' werden jeweils für die Lernstichpro-

benumfänge von 10, 20, 50 bzw. 100 je Klasse für die oben beschriebenen multivariaten Verteilungen erzeugt.

4.1 Gemischte Normalverteilung

Die gemischten Normalverteilungen für die Klasse 1 und 2 wurden mit $p_1 = 0.4$ und $p_2 = 0.1$ als $VN(\mu_1^*, \Sigma^*) = p_1 N(\mu_{11}, \Sigma_{11}) + (1 - p_1)N(\mu_{12}, \Sigma_{12})$ und $VN(\mu_2^*, \Sigma^*) = p_2 N(\mu_{21}, \Sigma_{21}) + (1 - p_2)N(\mu_{22}, \Sigma_{22})$ mit den Parametern aus Tabelle 3 erzeugt.

Tabelle 3: Parameter der Ausgangsnormalverteilungen

μ_{11}	μ_{12}	μ_{21}	μ_{22}	Σ_{11}			Σ_{12}			Σ_{21}			Σ_{22}		
2.70	3.2	-0.7	2.3	3.15	2.58	0.43	0.8	0.4	0.5	0.9	2.2	2.3	1.0	0.9	-0.3
6.85	7.1	7.1	8.1	2.58	3.86	2.81	0.4	0.9	0.1	2.2	11.1	9.3	0.9	1.0	0.1
4.85	5.1	2.2	4.2	0.43	2.81	5.21	0.5	0.1	1.0	2.3	9.3	12.6	-0.3	0.1	1.2

Dabei wurden für die Ausgangsverteilungen je ein Erwartungswertvektor und eine Kovarianzmatrix vorgegeben und die anderen so berechnet, dass sich die gemischte Normalverteilung mit den geforderten Parametern ergab.

4.2 Lognormalverteilung aus Johnsons Translationssystem

Die Lognormalverteilungen für die Klasse 1 und 2 wurden mit den Parametern aus Tabelle 4 erzeugt.

Tabelle 4: Parameter für die Erzeugung der Lognormalverteilung

$a1_1^2 = a1_2^2$	$a2_1$	$a2_2$	$\Sigma_1 = \Sigma_2 = \Sigma$
1.8	3	2	1.0 1.3 0.0
2.1	7	8	1.3 1.0 1.2
2.7	5	4	0.0 1.2 1.0

5 Simulationsergebnisse und Diskussion

Die Ergebnisse von Rudolph u.a. [9] zeigten deutlich, dass kleine Stichprobenumfänge in den Lernstichproben häufig zu sehr ungenauen Schätzungen der Parameter der zugrunde liegenden multivariaten Normalverteilung und

damit auch zu einer großen Variation der ermittelten Fehlklassifikationswahrscheinlichkeiten führen. Es wurde außerdem festgestellt, dass es bei multivariater Normalverteilung für kleine Stichproben auch anhand der Ergebnisse der crossvalidation kaum möglich ist, ein 'bestes' Verfahren aus den untersuchten herauszufinden.

Für den Stichprobenumfang 100 waren jedoch in fast allen Fällen die Fehlklassifikationswahrscheinlichkeiten unter crossvalidation bei der parametrischen Methode (ML) am kleinsten und würden hier zur Wahl eines 'richtigen' Verfahrens führen.

Die Ergebnisse der Simulation zur Beurteilung der Fehlklassifikationsraten für Lernstichproben und Testdaten, die einer multivariaten gemischten Normalverteilung genügen, sind in Tabelle 5 zusammengefasst.

Tabelle 5: Minima, Mittelwerte und Maxima der Fehlklassifikationswahrscheinlichkeiten über $Nsim = 10$ Wiederholungen ohne Berücksichtigung der Klassen bei gemischt normalverteilten Daten

Minima der Wahrscheinlichkeiten der

	Fehlklassifikation bei Kreuzvalidierung der Lernstichprobe					Fehlklassifikation bei Testdaten				
K[i]	ML	NN3	KU	KN	KE	ML	NN3	KU	KN	KE
10	0.00	0.00	0.10	0.00	0.00	0.01	0.01	0.11	0.01	0.05
20	0.00	0.00	0.05	0.00	0.05	0.00	0.01	0.12	0.01	0.08
50	0.00	0.00	0.04	0.00	0.02	0.00	0.00	0.08	0.00	0.06
100	0.00	0.01	0.07	0.00	0.05	0.00	0.01	0.08	0.00	0.04

Mittelwerte der Wahrscheinlichkeiten der

	Fehlklassifikation bei Kreuzvalidierung der Lernstichprobe					Fehlklassifikation bei Testdaten				
K[i]	ML	NN3	KU	KN	KE	ML	NN3	KU	KN	KE
10	0.13	0.10	0.39	0.09	0.23	0.10	0.12	0.35	0.11	0.23
20	0.08	0.10	0.29	0.11	0.20	0.07	0.08	0.26	0.09	0.17
50	0.08	0.08	0.19	0.08	0.15	0.08	0.08	0.20	0.07	0.14
100	0.06	0.05	0.15	0.05	0.11	0.07	0.06	0.17	0.07	0.12

Maxima der Wahrscheinlichkeiten der

	Fehlklassifikation bei Kreuzvalidierung der Lernstichprobe					Fehlklassifikation bei Testdaten				
K[i]	ML	NN3	KU	KN	KE	ML	NN3	KU	KN	KE
10	0.40	0.40	1.00	0.40	0.60	0.21	0.22	0.63	0.28	0.49
20	0.25	0.35	0.65	0.35	0.50	0.21	0.18	0.43	0.20	0.28
50	0.22	0.20	0.40	0.24	0.32	0.19	0.16	0.33	0.19	0.25
100	0.15	0.11	0.27	0.13	0.19	0.16	0.13	0.30	0.15	0.21

ML, NN3 und KN schneiden am besten ab. Die Abhängigkeit der Genauigkeit vom Umfang der Lernstichprobe ist deutlich. Die Simulationsergebnisse mit Lernstichproben und Testdaten, die der vorgegebenen multivariaten Lognormalverteilung genügen, sind in Tabelle 6 zusammengestellt.

Tabelle 6: Minima, Mittelwerte und Maxima der Fehlklassifikationswahrscheinlichkeiten über $Nsim = 10$ Wiederholungen ohne Berücksichtigung der Klassen bei lognormalverteilten Daten

Minima der Wahrscheinlichkeiten der

	Fehlklassifikation bei Kreuzvalidierung der Lernstichprobe					Fehlklassifikation bei Testdaten				
K[i]	ML	NN3	KU	KN	KE	ML	NN3	KU	KN	KE
10	0.00	0.00	0.10	0.00	0.00	0.01	0.01	0.13	0.01	0.10
20	0.00	0.00	0.10	0.00	0.00	0.02	0.02	0.09	0.01	0.06
50	0.00	0.02	0.04	0.02	0.04	0.01	0.02	0.10	0.01	0.07
100	0.01	0.01	0.06	0.02	0.04	0.04	0.03	0.08	0.03	0.06

Mittelwerte der Wahrscheinlichkeiten der

	Fehlklassifikation bei Kreuzvalidierung der Lernstichprobe					Fehlklassifikation bei Testdaten				
K[i]	ML	NN3	KU	KN	KE	ML	NN3	KU	KN	KE
10	0.10	0.07	0.26	0.08	0.19	0.07	0.07	0.25	0.07	0.18
20	0.06	0.05	0.21	0.03	0.11	0.06	0.06	0.18	0.06	0.13
50	0.08	0.06	0.13	0.06	0.10	0.06	0.05	0.15	0.05	0.11
100	0.05	0.05	0.12	0.05	0.09	0.06	0.05	0.13	0.05	0.10

Maxima der Wahrscheinlichkeiten der

	Fehlklassifikation bei Kreuzvalidierung der Lernstichprobe					Fehlklassifikation bei Testdaten				
K[i]	ML	NN3	KU	KN	KE	ML	NN3	KU	KN	KE
10	0.20	0.30	0.60	0.20	0.60	0.15	0.17	0.39	0.17	0.30
20	0.25	0.15	0.40	0.10	0.25	0.13	0.13	0.26	0.12	0.19
50	0.18	0.12	0.24	0.12	0.16	0.13	0.12	0.19	0.13	0.16
100	0.08	0.08	0.16	0.08	0.14	0.09	0.07	0.18	0.07	0.14

NN3 und KN zeigen die wenigsten Fehlklassifikationen. Dabei trat zwischen beiden Verfahren kaum ein Unterschied auf. Das ML-Verfahren bringt etwas schlechtere Ergebnisse, was man bei der verwendeten Verteilung durchaus auch erwarten konnte. KU und KE fallen dagegen deutlich ab.

Zusammenfassend kann man feststellen, dass bei stärkerer Abweichung von der multivariaten Normalverteilung die parameterfreien Verfahren besser abschneiden. In unserem Fall waren NN3 und KN zu empfehlen. Wenn man sich bei den Kerndichteverfahren unsicher ist, empfehlen wir NN3 zu nehmen. Dabei kann man im Mittel sicherlich die wenigsten Fehler machen. Dies gilt auch bei Unsicherheit bezüglich der multivariaten Verteilung, die der Lernstichprobe und den Testdaten zugrunde liegt.

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.