Ein SAS-Makro zur Erzeugung multivariat normalverteilter Zufallsgrößen

Armin Tuchscherer[1)*], Paul Eberhard Rudolph[1)], Bernd Jäger[2)], Margret Tuchscherer[1)]

1) Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere Dummerstorf-Rostock, Wilhelm-Stahl-Allee 2, D-18196 Dummerstorf
* *E-mail:* atuchsch@fbn-dummerstorf.de
2) Ernst-Moritz-Arndt - Universität Greifswald, Institut für Biometrie u. Medizinische Informatik, Sauerbruchstraße, D-17489 Greifswald

1. Einleitung

Eigenschaften insbesondere multivariater statistischer Verfahren sind häufig nicht mehr analytisch, sondern nur mittels Simulationen bestimmbar. Zur Durchführung von Simulationen im Zusammenhang mit multivariaten Verfahren benötigt man geeignete Programme zur Erzeugung multivariat verteilter Zufallsgrößen. Um derartige Simulationen in SAS durchzuführen, ist es naheliegend, die entsprechenden Zufallszahlengeneratoren als SAS-Makros bereitzustellen.
Unter Verwendung von SAS/IML wird ein SAS-Makro vorgestellt, mit dem sich für eine vorgegebene Dimension, einen zugehörigen Mittelwertvektor und eine ebenfalls vorgegebene Kovarianzstruktur multivariat normalverteilte Zufallsgrößen erzeugen lassen. Schätzungen der vorgegeben Parameter aus mit diesem Makro erzeugten Stichproben für verschiedene Stichprobenumfänge und Dimensionen zeigen, wie gut multivariate Normalverteilungen mit diesem Makro nachgebildet werden können, falls die Stichprobenumfänge hinreichend groß sind.

2. Die multivariate Normalverteilung

Zunächst soll die multivatiate Normalverteilung kurz beschrieben werden, gefolgt von Bemerkungen zu Eigenschaften, die für die Erzeugung entsprechender Zufallsvariablen benötigt werden. Ein p-dimensionaler Zufallsvektor $\underline{\mathbf{y}} = (\underline{y}_1,\ldots,\underline{y}_p)'$ heißt p-variat normalverteilt mit Erwartungswertvektor

$$E(\underline{\mathbf{y}}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \tag{1}$$

und Kovarianzmatrix

$$\operatorname{cov}(\underline{\mathbf{y}},\underline{\mathbf{y}}') = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_2^2 & & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_p^2 \end{bmatrix}, \tag{2}$$

wenn die Dichtefunktion $f(\mathbf{y})$ die Gestalt

$$f(\mathbf{y}) = ((2\pi)^p |\boldsymbol{\Sigma}|)^{-1/2} \cdot e^{-\frac{(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}} \tag{3}$$

hat. Die Kurzbezeichnung

$$\underline{\mathbf{y}} = (\underline{y}_1,\ldots,\underline{y}_p)' \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{4}$$

für die p-variate Normalverteilung einer Zufallsvariablen $\underline{\mathbf{y}}$ ist allgemein gebräuchlich.
Für $p = 2$ kann man die Dichtefunktion (3) z.B. sehr schön mit der SAS-Prozedur `g3d` grafisch darstellen (Abb. 1).

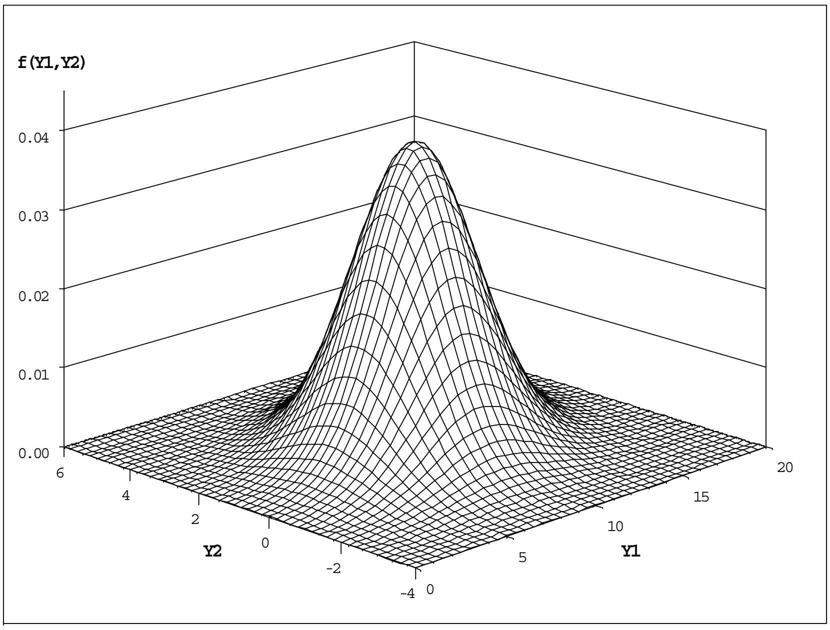


Abbildung 1: Dichtefunktion $f(y_1, y_2)$ der bivariaten Normalverteilung $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ mit

$$\boldsymbol{\mu} = \begin{bmatrix} 10 \\ 1 \end{bmatrix} \text{ und } \boldsymbol{\Sigma} = \begin{bmatrix} 9 & 1 \\ 1 & 2 \end{bmatrix}$$

2.1 Lineare Transformation multivariat normalverteilter Zufallsvariabler

Betrachten wir eine m-variate normalverteilte Zufallsvariable $\underline{\mathbf{z}} \sim N_m(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)$ und eine lineare Transformation der Variablen der Form

$$\underline{\mathbf{y}} = \mathbf{L} \cdot \underline{\mathbf{z}} + \boldsymbol{\mu}, \tag{5}$$

wobei $\mathbf{L}$ eine $(p \times m)$-Matrix mit $p \leq m$ ist und $\boldsymbol{\mu}$ ein p-dimensionaler reeller Vektor, dann ist $\underline{\mathbf{y}}$ p-variat normalverteilt mit Erwartungswertvektor $E(\underline{\mathbf{y}}) = \mathbf{L} \cdot E(\underline{\mathbf{z}}) + \boldsymbol{\mu}$ und der Kovarianzmatrix $\mathrm{cov}(\underline{\mathbf{y}}, \underline{\mathbf{y}}') = \mathbf{L} \cdot \mathrm{cov}(\underline{\mathbf{z}}, \underline{\mathbf{z}}') \cdot \mathbf{L}'$, d.h. $\boldsymbol{\Sigma} = \mathbf{L}\boldsymbol{\Sigma}_z\mathbf{L}'$.
Wählt man $\underline{\mathbf{z}} \sim N_m(\mathbf{0}, \mathbf{I}_m)$, d.h. die Komponenten von $\underline{\mathbf{z}}$ sind unabhängig (0,1)-normalverteilt, so gilt für die lineare Transformation (5):

$$\underline{\mathbf{y}} \sim N_p(\boldsymbol{\mu}, \mathbf{LL}'),$$

d.h. $\underline{\mathbf{y}}$ ist p-variat normalverteilt mit Erwartungswert $E(\underline{\mathbf{y}}) = \mathbf{L} \cdot E(\underline{\mathbf{z}}) + \boldsymbol{\mu} = \mathbf{L} \cdot \mathbf{0} + \boldsymbol{\mu} = \boldsymbol{\mu}$ und der Kovarianzmatrix $\mathrm{cov}(\underline{\mathbf{y}}, \underline{\mathbf{y}}') = \mathbf{L} \cdot \mathrm{cov}(\underline{\mathbf{z}}, \underline{\mathbf{z}}') \cdot \mathbf{L}' = \mathbf{L} \cdot \mathbf{I}_m \cdot \mathbf{L}' = \mathbf{LL}'$. Mit $\mathbf{LL}' = \boldsymbol{\Sigma}$ läßt sich somit jede p-variat normalverteilte Zufallsvariable $\underline{\mathbf{y}} = (\underline{y}_1,\ldots,\underline{y}_p)' \sim N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ als lineare Transformation von p unabhängig (0,1)-normalverteilten Zufallsvariablen darstellen.
Es ist somit nur noch erforderlich, eine Matrix $\mathbf{L}$ zu bestimmen, die der Bedingung $\mathbf{LL}' = \boldsymbol{\Sigma}$ genügt.

2.2 Zerlegung der Kovarianzmatrix $\mathbf{\Sigma}$

Die symmetrische und positiv definite Matrix $\mathbf{\Sigma}$ ist in der Form

$$\mathbf{\Sigma} = \mathbf{LL}' \tag{6}$$

darzustellen. Eine mögliche Wahl der Matrix $\mathbf{L}$ ist die $(p \times p)$-Dreiecksmatrix $\mathbf{L} = \mathbf{A}$

$$\mathbf{A} = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}, \tag{7}$$

die man mit Hilfe der folgenden rekursiven Gleichungen (Choleski-Zerlegung; JOHNSON, 1987) bestimmen kann:

$$\left.\begin{array}{ll} a_{i1} = \sigma_{i1} / \sigma_1^{1/2} & ;\quad 1 \le i \le p \\ a_{ij} = \left(\sigma_{ij} - \sum_{k=1}^{j-1} a_{ik} a_{jk}\right) \Big/ a_{jj} & ;\quad 1 < j < i \le p \\ a_{ii} = \left(\sigma_i - \sum_{k=1}^{i-1} a_{ik}^2\right)^{1/2} & ;\quad 2 \le j = i \le p \end{array}\right\}. \tag{8}$$

Dabei wird folgendes Schema abgearbeitet:

$$
\begin{array}{cc|c}
 & & \begin{array}{ccccccc} & & & \mathbf{j} & & & \\ 1 & 2 & 3 & 4 & 5 & \dots & p \end{array} \\ \hline
\begin{array}{c} \\ \\ \mathbf{i} \\ \\ \\ \\ \end{array} & \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ p \end{array} & \left[\begin{array}{ccccccc}
a_{11} & 0 & 0 & 0 & 0 & \dots & 0 \\
a_{21} \longrightarrow & a_{22} & 0 & 0 & 0 & \dots & 0 \\
a_{31} \longrightarrow & a_{32} \longrightarrow & a_{33} & 0 & 0 & \dots & 0 \\
a_{41} \longrightarrow & a_{42} \longrightarrow & a_{43} \longrightarrow & a_{44} & 0 & \dots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots \\
a_{p1} \longrightarrow & a_{p2} \longrightarrow & a_{p3} \longrightarrow & a_{p4} \longrightarrow & a_{p5} \longrightarrow & \dots \longrightarrow & a_{pp}
\end{array}\right]
\end{array}
$$

2.3 Algorithmus zur Erzeugung p-variat normalverteilter Zufallsvariablen

(i) Einlesen des Erwartungswertvektors $\boldsymbol{\mu}$ und der Kovarianzmatrix $\mathbf{\Sigma}$

(ii) Berechnung von $\mathbf{A}$ aus $\mathbf{\Sigma}$ nach (8), so daß $\mathbf{AA}' = \mathbf{\Sigma}$

(iii) Erzeugung von $\mathbf{z} = (z_1, \ldots, z_p)'$ nach $\underline{z} \sim N_p(\mathbf{0}, \mathbf{I})$ unter Verwendung einer Prozedur zur Erzeugung unabhängiger standardnormalverteilter Zufallsvariabler

(iv) Berechnung von $\mathbf{y} = \mathbf{A} \cdot \mathbf{z} + \boldsymbol{\mu}$

3. Umsetzung des Algorithmus zur Erzeugung p-variat normalverteilter Zufallsvariabler als SAS-Makro

3.1 SAS-Makro zur Erzeugung p-variat normalverteilter Zufallsvariabler

Der oben beschriebene Algorithmus soll durch ein SAS-Makro unter Verwendung der SAS-Prozedur `iml` umgesetzt werden, wobei vorausgesetzt wird, daß der Mittelwertvektor und die Kovarianzmatrix jeweils als (permanente bzw. temporäre) SAS-Datei vorliegen.
Ein Beispiel für den Fall temporärer SAS-Dateien und $p = 4$ enthält die folgende Tabelle.

Temporäre SAS-Datei für Mittelwertvektor	temporäre SAS-Datei für Kovarianzmatrix
Data **WORK.MITTEL**; input mue; cards; 10 1 5 8 run;	data **WORK.KOVAR**; input co1-co4; cards; 9 1 0.5 -1 1 2 0.5 2 0.5 0.5 3 1 -1 2 1 7 run;

Die im folgenden Schema jeweils grau hinterlegten Programmzeilen des Makros repräsentieren die Schritte des oben beschriebenen Algorithmus (Nummer in der linken Spalte von Tab. 1).

Tabelle 1: SAS-Makro zur Erzeugung p-variat normalverteilter Zufallsvariabler mit Kommentaren (rechte Spalte)

SAS-Makro: %MACRO PNORMAL(M, MEANS, COVAR, DATA, NAMES)

	%MACRO PNORMAL(M, MEANS, COVAR, DATA, NAMES); proc iml;	❶
(i)	use &MEANS; read all into MUE; use &COVAR; read all into COV;	

	p=ncol(COV);
	de=det(COV);
	ERROR='Die Kovarianzmatrix ist nicht positiv semidefinit! Determinante:';
	if de<0 then print ERROR de;	

(ii)	A=j(p,p,0);
	do i=1 to p;
	do j=1 to i;
	if j=1 then do; A[i,1]=cov[i,1]/sqrt(cov[1,1]); end;
	if j<i then do;
	sum=0;
	do k=1 to j-1;
	sum=sum+A[i,k]*A[j,k];
	end;
	A[i,j]=(COV[i,j]-sum)/A[j,j] ;
	end;
	if i=j then do;
	sm=0;
	do kk=1 to i-1;	

SAS-Makro: %MACRO PNORMAL(M, MEANS, COVAR, DATA, NAMES)

```
              sm=sm+A[i,kk]*A[i,kk];
              end;

            A[i,i]=sqrt(COV[i,i]-sm) ;
         end;
      end;
    end;

    Test=A*A`;
    print 'Testdruck für die Zerlegung der Kovarianzmatrix: TEST=A*A`' ;   ❷
    print A;
    print COV '    ' TEST;

    DAT=j(&m,p,0);
    MUE_EST=j(p,1,0);
    COV_EST=J(p,p,0);

    do i=1 to &M;

(iii)  Z=j(p,1,0);                                                       ❸
       Z=NORMAL(Z);

(iv)   Y=A*Z+mue;

       MUE_EST=MUE_EST+Y;
       do j=1 to p;
          DAT[i,j]=Y[j];
       end;
    end;

    MUE_EST=MUE_EST/&m;
    print ' ';
    print 'Vergleich Parameter - Schätzung bei Stichprobenumfang:' &M;    ❹
    Mattrib MUE_EST format=10.2;
    print 'Parameter'  '        Schätzung'  ;
    print '------------------------------';
    print  MUE '       '   MUE_EST ;
    do i=1 to &M;
       do k=1 to p;
          do j=1 to p;
          COV_EST[k,j]=COV_EST[k,j]+(DAT[i,k]-MUE_EST[k])*(DAT[i,j]-
                       MUE_EST[j]);
          end;
       end;
    end;

    COV_EST=COV_EST/(&m-1);

    Mattrib COV_EST format=10.2;                                          ❹
    print '  ' COV '      ' COV_EST ;

    c=&NAMES ;                                                            ❺
    create &Data from DAT [colname=c];
    append from DAT;

quit;
%mend MNORMAL;
```

❶ Parameter des Makros PNORMAL(M, MEANS, COVAR, DATA, NAMES)
INPUT:

`M`	Anzahl der zu erzeugenden Datensätze, z.B.:	`10`
`MEANS`	SAS-Datei die den Mittelwertvektor enthält, z.B.:	`WORK.Mittel` oder `SASUSER.Mittel`
`COVAR`	SAS-Datei die die Kovarianzmatrix enthält, z.B.:	`WORK.KOVAR` oder `SASUSER.KOVAR`

OUTPUT:

`DATA` SAS-Datei in die die erzeugten Datensätze geschrieben werden, z.B.:
`WORK.Daten` oder
`SASUSER.Daten`

`NAMES` Variablennamen in `WORK.Daten` bzw. `SASUSER.Daten`, z.B.:
`{'Y1' 'Y2' 'Y3' 'Y4'}` oder
`'Y1' : 'Y4'`

Das Makro könnte zum Beispiel wie folgt aufgerufen werden:

```
%PNORMAL(10, Mittel, KOVAR, Daten, {'Y1' 'Y2' 'Y3' 'Y4'});
```
(9)

Mit diesem Aufruf werden `M = 10` Datensätze mit 4-dimensional normalverteilten Variablen Y1,...,Y4 auf der Basis der temporären SAS-Dateien 'Mittel' (Mittelwertvektor) und 'KOVAR' (Kovarianzmatrix) erzeugt, die in der temporären SAS-Datei 'Daten' abgelegt werden (siehe ❺).

❷ Kontrollausdruck für die Zerlegung der Kovarianzmatrix $\boldsymbol{\Sigma}$ (im Beispiel: `COV`) in **AA`** mit der Dreiecksmatrix (7). `COV` und `TEST` müssen identisch sein. Mit dem Makroaufruf (9) erhält man folgenden Testdruck:

```
            Testdruck für die Zerlegung der Kovarianzmatrix: TEST=A*A`


                             A
                             3         0         0         0
                     0.3333333 1.3743685         0         0
                     0.1666667 0.3233808 1.6934128         0
                     -0.333333  1.536059 0.3299984 2.1025016


      COV                                        TEST
        9         1       0.5        -1            9         1       0.5        -1
        1         2       0.5         2            1         2       0.5         2
      0.5       0.5         3         1          0.5       0.5         3         1
       -1         2         1         7           -1         2         1         7
```

❸ Durch Nullsetzung des p-dimensionalen Vektors **z** wird mit dem Aufruf `NORMAL(z)` die `IML`-Funktion `NORMAL(.)` mit zufälligen `SEED`-Werten (Systemzeit) gestartet und ein p-dimensionaler Vektor mit unabhängig (0,1)-normalverteilten Komponenten erzeugt.
❹ Druck der Parameter (Mittelwertvektor `'MUE'`, und Kovarianzmatrix 'COV') und ihrer Schätzungen `'MUE_EST'` bzw. 'COV_EST' aus den M erzeugten Datensätzen (hier z.B. `M = 10`).

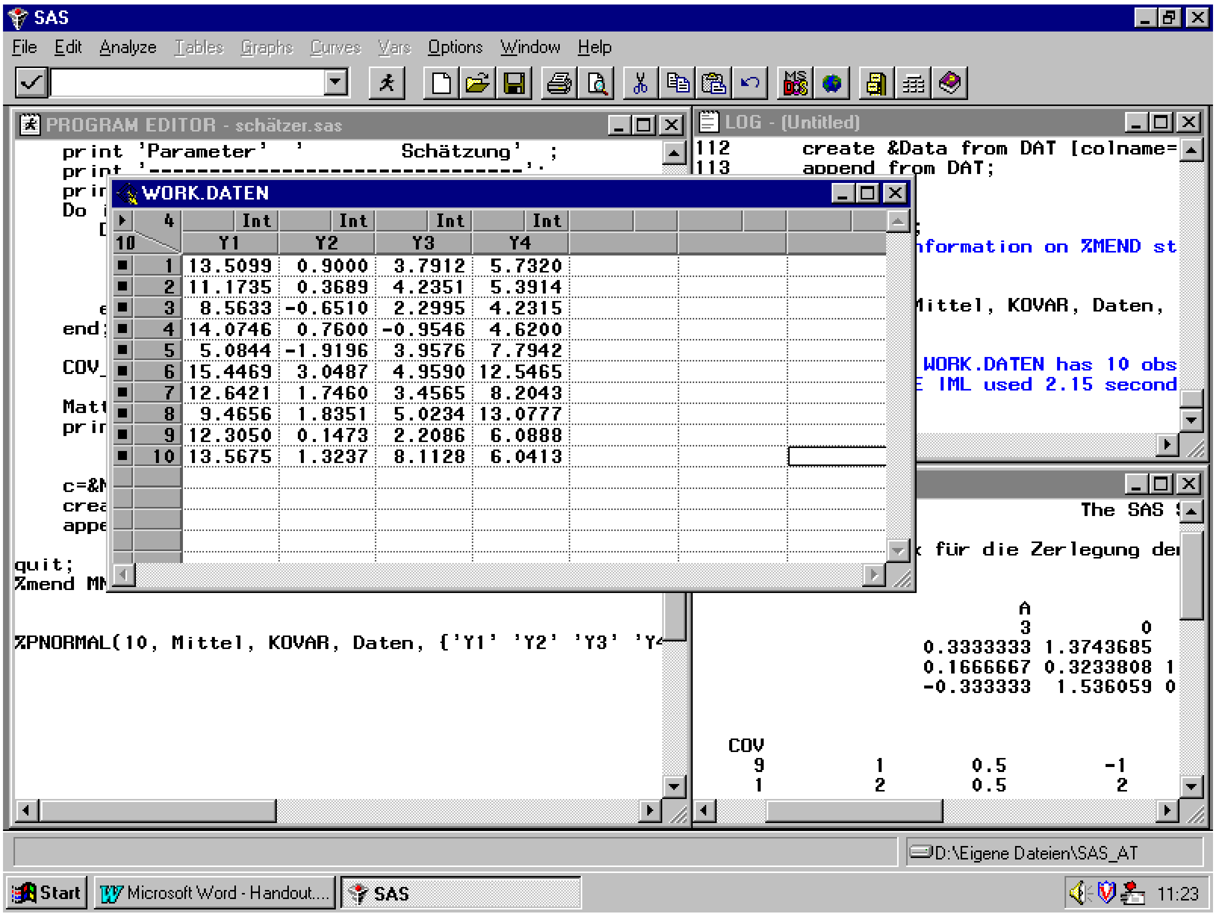
```
          Vergleich Parameter - Schätzung bei Stichprobenumfang:       10

                             Parameter          Schätzung
                           -----------------------------

                                   MUE            MUE_EST
                                    10              11.58
                                     1               0.76
                                     5               3.71
                                     8               7.37


      COV                                         COV_EST
        9          1        0.5        -1            9.60      3.38      0.04      0.31
        1          2        0.5         2            3.38      1.93      0.95      2.52
      0.5        0.5          3         1            0.04      0.95      5.45      2.89
       -1          2          1         7            0.31      2.52      2.89      9.74
```

❺ Die erzeugten Datensätze werden von der Matrix `DAT` in die SAS-Datei `WORK.DATEN` geschrieben, die z.B. bei Makroaufruf (9) folgende Gestalt haben könnte.



Das Laufzeitverhalten des Makros `PNORMAL(M, MEANS, COVAR, DATA, NAMES)` wurde für unterschiedliche Dimensionen und Anzahlen der zu erzeugenden multivariat normalverteilten Zufallsvektoren geprüft. Die Ergebnisse sind in Tabelle 2 zusammengefaßt.

Tabelle 2: Mittlere Rechenzeit des Makros `PNORMAL(M, MEANS, COVAR, DATA, NAMES)`

Dimension der Zufallsvektoren P	Anzahl der Datensätze M	Rechenzeit [sec] PC 128 MHz 46 MB RAM	Rechenzeit [sec] PC 300 MHz 64 MB RAM
4	1	0.21	0.14
	10	0.23	0.15
	100	0.26	0.19
	1000	0.42	0.25
	10000	2.21	1.09
2	1	0.20	0.14
	10	0.23	0.14
	100	0.25	0.15
	1000	0.38	0.21
	10000	1.50	0.78

Die Tabelle 2 verdeutlicht, daß selbst Simulationsumfänge von 10.000 Datensätzen im Sekundenbereich mit Standard-PC zu bewältigen sind. Bei kleinen Umfängen kann man erkennen, daß ein wesenlicher Teil der Rechenzeit für das Einlesen und Ausgeben der Daten benötigt wird. Für intensive Simulationsuntersuchungen ist daher zu empfehlen, möglichst viele Operation in `IML` auszuführen, um die Anzahl der zeitaufwendigen Einlese-, Ausgabeoperationen und `Datasteps` so gering wie möglich zu halten.

Mit Hilfe des folgenden Beispiels soll nun der Einfluß des Simulationsumfanges auf die Genauigkeit der Schätzung des Mittelwertvektors und der Kovarianzmatrix demonstriert werden, ohne auf Genauigkeitsabschätzung oder eventuelle Versuchsplanung einzugehen.
Zu diesem Zweck wählen wir Umfänge von 10, 100, 1000, 10000 und rufen das Makro mit den entsprechenden Parametern auf:

```
%PNORMAL(   10, Mittel, KOVAR, Daten1, {'Y1' 'Y2' 'Y3' 'Y4'});
%PNORMAL(  100, Mittel, KOVAR, Daten2, {'Y1' 'Y2' 'Y3' 'Y4'});
%PNORMAL( 1000, Mittel, KOVAR, Daten3, {'Y1' 'Y2' 'Y3' 'Y4'});
%PNORMAL(10000, Mittel, KOVAR, Daten3, {'Y1' 'Y2' 'Y3' 'Y4'});
```

Bei kleinen Umfängen muß man mit recht ungenauen Schätzungen für den Mittelwertvektor `MUE` und besonders für die Kovarianzmatrix `COV` rechnen, wie im folgenden `SAS-OUTPUT` zu sehen ist. Wie erwartet, erhält man für hinreichend große Stichprobenumfänge auch sehr gut geschätzte Mitttelwerte und Kovarianzmatrizen.

`SAS-OUTPUT` (Parameter und Schätzungen bei unterschiedlichen Umfängen):

```
          Vergleich Parameter - Schätzung bei Stichprobenumfang:       10

                            Parameter        Schätzung
                          ------------------------------
                                  MUE          MUE_EST
                                   10            10.44
                                    1             1.21
                                    5             4.88
                                    8             8.48

COV                                            COV_EST
  9         1       0.5        -1             4.97      -1.00      0.22      -4.12
  1         2       0.5         2            -1.00       1.33      0.02       2.37
0.5       0.5         3         1             0.22       0.02      0.73      -0.48
 -1         2         1         7            -4.12       2.37     -0.48       8.16

          Vergleich Parameter - Schätzung bei Stichprobenumfang:      100

                            Parameter        Schätzung
                          ------------------------------
                                  MUE          MUE_EST
                                   10             9.76
                                    1             1.23
                                    5             5.01
                                    8             8.21

COV                                            COV_EST
  9         1       0.5        -1             7.65       0.55     -0.82      -1.46
  1         2       0.5         2             0.55       1.84      0.41       2.03
0.5       0.5         3         1            -0.82       0.41      2.91       1.35
 -1         2         1         7            -1.46       2.03      1.35       6.31

          Vergleich Parameter - Schätzung bei Stichprobenumfang:     1000

                            Parameter        Schätzung
                          ------------------------------
                                  MUE          MUE_EST
                                   10             9.82
                                    1             1.10
                                    5             5.06
                                    8             8.15

COV                                            COV_EST
  9         1       0.5        -1             9.52       1.11      0.41      -1.04
  1         2       0.5         2             1.11       1.86      0.50       1.84
0.5       0.5         3         1             0.41       0.50      2.88       0.97
 -1         2         1         7            -1.04       1.84      0.97       6.66

          Vergleich Parameter - Schätzung bei Stichprobenumfang:    10000

                            Parameter        Schätzung
                          ------------------------------
                                  MUE          MUE_EST
                                   10             9.94
                                    1             0.99
                                    5             5.02
                                    8             8.00

COV                                            COV_EST
  9         1       0.5        -1             8.95       0.98      0.44      -0.94
  1         2       0.5         2             0.98       1.96      0.51       2.01
0.5       0.5         3         1             0.44       0.51      3.00       1.02
 -1         2         1         7            -0.94       2.01      1.02       6.98
```

3.2 SAS-Makro zur Erzeugung multivariat normalverteilter Zufallsgrößen in Klassen

Wie bereits im vorherigen Abschnitt beschrieben, ist es ratsam, bei umfangreichen Simulationen so viele Operationen wie möglich in `IML` auszuführen, um die Zahl der zeitaufwendigen Einlese-, Ausgabeoperationen und Datasteps auf ein Minimum zu reduzieren.
Zur Beurteilung multivariater Verfahren wie z.B. der Diskriminanzanalyse werden multivariat normalverteilte Zufallsvariable in Klassen benötigt, die möglicherweise durch unterschiedliche Mittelwertvektoren und eventuell auch durch unterschiedliche Kovarinzmatrizen charakterisiert sind.
Aus diesem Grund wurde das Makro

`PNORMAL(M, MEANS, COVAR, DATA, NAMES)` (10)

derart modifiziert, daß mit dem resultierenden Makro

`PNORMALK(M, MEANS, COVAR, DATA, NAMES)` (11)

p-variat normalverteilte Zufallsvariablen in Klassen erzeugt werden können.
Da (11) für die Zufallsvariablenerzeugung zur Beurteilung von Diskriminanzanalyseverfahren gedacht ist, wurde hier auf die Schätzung der Mittelwertvektoren und der Kovarianzmatrizen in den Klassen innerhalb des Makros verzichtet, da diese mit Hilfe der SAS-Prozedur `DISCRIM` ohnehin optional bestimmt werden können.
Ein weiterer Unterschied von (11) im Vergleich zu (10) besteht in der Struktur der Eingabegrößen `MEANS` und `COVAR`.
In der SAS-Datei `MEANS` werden bei (11) die Mittelwertvektoren der Klassen als Zeilenvektoren erwartet, wobei das erste Element einer Zeile immer der Umfang der entsprechenden Klasse ist. Durch die Anzahl der Zeilen von `MEANS` wird automatisch die Dimension p bestimmt. In der SAS-Datei `COVAR` stehen die Kovarianzmatrizen der einzelnen Klassen untereinander.
Die SAS-Datei DATA, als Ergebnis des Makros (11), enthält als erste Variable die Nummer der Simulationswiederholung (1,...,M), gefolgt von der Klassennummer als zweite Variable und den p Komponenten der erzeugten Zufallsvektors als dritte bis (p+2)-te Variable.
Das vollständige Mako mit Kommentaren, Aufruf-, Testbeispiel und Ergebnisdatei ist als Anlage angefügt.

Literatur

Geißler, H.; Ortseifen, C.: SAS-Makroprogrammierung: Eine Einführung. Universitätsrechenzentrum Heidelberg 1995, 54 S.

Johnson, M. E.: Multivariate Statistical Simulation. J. Wiley, New York 1987.

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

Kleijnen; J.; van Groenendaal, W.: Simulation: A Statistical Perspective. J. Wiley, Chichester 1992, 241 pp.

SAS Institute Inc., SAS/IML Software: Usage and Reference, Version 6, First Edition Cary, NC: SAS Institute Inc., 1989, 501 pp.

SAS Institute Inc., SAS Language: Reference, Version 6, First Edition Cary, NC: SAS Institute Inc., 1990, 1042 pp.

Anlage

A1 SAS-Makro zur Erzeugung multivariat normalverteilter Zufallsgrößen in Klassen

```
/*---------------------------------------------------------------------------*
 |  %PNORMALK(M, MEANS, COVAR, Data, names)                                  |
 |  ======================================                                   |
 |                                                                           |
 |  erzeugt p-dimensional normalverteilte Merkmalsvektoren  Y=(y1,...,yp)    |
 |         in NK Klassen vom Umfang k(i)  mit                                |
 |                                                                           |
 |         Erwartungswertvektor mue_i=(mue_1(1), ..., mue_1(p))'   und       |
 |                                                                           |
 |                                   /                           \           |
 |                                  | COV_i(1,1)  ...  COV_i(1,p) |          |
 |         Kovarianzmatrix      COV_i=|  ...                       |          |
 |                                  | COV_i(p,1)  ...  COV_i(p,p) |          |
 |                                   \                           /           |
 |                                                                           |
 |         in der Klasse i , (i=1,...,NK).                                   |
 |  ---------------------------------------------------------------------    |
 |  INPUT:                                                                   |
 |  ======                                                                   |
 |                                                                           |
 |    M         Anzahl der Simulationswiederholungen in jeder Klasse i       |
 |              (i=1,...,NK)                                                 |
 |                                                                           |
 |    MEANS     SAS-Datei der Form                                           |
 |                                                                           |
 |               k(1)   mue_1(1)  ... mue_1(p)                               |
 |               .       .             .                                     |
 |               .       .             .                                     |
 |               k(NK)  mue_NK(1) ... mue_NK(p)                              |
 |                                                                           |
 |    COVAR     SAS-Datei in der Form                                        |
 |                                                                           |
 |               COV_1(1,1)  ...  COV_1(1,p)                                 |
 |                .               .                                          |
 |                .               .                                          |
 |               COV_1(p,1)  ...  COV_1(p,p)                                 |
 |                            .                                              |
 |                            .                                              |
 |               COV_NK(1,1) ...  COV_NK(1,p)                                |
 |                .               .                                          |
 |                .               .                                          |
 |               COV_NK(p,1) ...  COV_NK(p,p)                                |
 |                                                                           |
 |  OUTPUT:                                                                  |
 |  =======                                                                  |
 |    DATA       SAS-Datei, mit der Datensatzstruktur:                       |
 |               Simulations-Nr. Klassen-Nr. y1 y2 ... yp                    |
 |                                                                           |
 |    NAMES      Namen der Variablen der Datei 'DATA'                        |
 |               z.B. bei p=4 in der Form:                                   |
 |                    {'Nsim' 'CLASS_NR' 'Y1' 'Y2' 'Y3' 'Y4'}                |
 *---------------------------------------------------------------------------*/
```

```
%macro PNORMALK(M, MEANS, COVAR, DATA, NAMES);

proc iml;
    use &MEANS; read all into MEANS;
    use &COVAR;  read all into COVAR;
    p=ncol(means)-1;
    nk=nrow(means);

    mue=j(p,1,0);
    cov=j(p,p,0);

    NKi=0;
    do i=1 to NK;
    NKi=NKi+means[i,1];
    end;

    DAT=j(&m*NKi, p+2,0);

    L=0;

    do KNR=1 to nk;

       do i=1 to p;
       mue[i]=means[KNR,i+1];
       kk=(KNR-1)*p+i;
          do j=1 to p;
           cov[i,j]=COVAR[kk,j];
          end;
       end;

       de=det(COV);
       ERROR='Die Kovarianzmatrix ist nicht positiv semidefinit in Klasse:';
       if de<0 then print ERROR KNR;

       A=j(p,p,0);
       do i=1 to p;
         do j=1 to i;
           if j=1 then do; A[i,1]=cov[i,1]/sqrt(cov[1,1]); end;
           if j<i then do;
           sum=0;
              do k=1 to j-1;
              sum=sum+A[i,k]*A[j,k];
              end;
            A[i,j]=(COV[i,j]-sum)/A[j,j] ;
            end;
            if i=j then do;
            sm=0;
              do kk=1 to i-1;
              sm=sm+A[i,kk]*A[i,kk];
              end;
            A[i,i]=sqrt(COV[i,i]-sm) ;
            end;
         end;
       end;
```

```
       do k=1 to &M;
          do i=1 to means[KNR,1] ;
            Z=j(p,1,0);
            Z=NORMAL(Z);
            Y=A*Z+mue;
            L=L+1;
            DAT[L,2]=KNR;
            DAT[L,1]=k;
            do j=1 to p;
               DAT[L,j+2]=Y[j];
            end;
          end;
       end;

    end;

    c=&names;
    create &Data from DAT [colname=c];
    append from DAT;

quit;

%mend PNORMALK;
```

A2 Testbeispiel

```
/*==================================================================================*
 |  Testbeispiel:                                                                   |
 |  =============                                                                   |
 |  1. Erzeugen der temporären SAS-Dateien 'Mittel' und 'KOVAR' (Voraussetzung)     |
 |  2. Aufruf des Makros 'PNORMALK':                                                |
 |     %PNORMALK(10, Mittel, KOVAR, Daten, {'Nsim' 'Klasse' 'X1' 'X2' 'X3' 'X4'});|
 |  3. Ergebnis: temporäre SAS-Datei 'Daten' mit den Variablen:                     |
 |                                                     Nsim  Klasse  X1  X2  X3  X4 |
*==================================================================================*/

/*--------------------------------------------------------------------*
 |  Erzeugen der Datei 'Mittel' (Klassenumfänge, Mittelwertvektoren)  |
 *--------------------------------------------------------------------*/

data Mittel;
input  Ki MUE1-MUE4;
cards;
     5  1  3  6  0.5
     4  3  3  5   1
     3  2  4  2   6
run;

/*----------------------------------------------------*
 |  Erzeugen der Datei 'KOVAR' (Kovarianzmatrizen)     |
 *----------------------------------------------------*/

data KOVAR;
input   co1-co4;
```

```
cards;
       9     1     0.5  -1
       1     2     0.5   2
       0.5   0.5   3     1
      -1     2     1     7
       8     1     0.5  -1
       1     4     0.5   2
       0.5   0.5   6     1
      -1     2     1     5
       4     1     0.5  -1
       1     9     0.5   2
       0.5   0.5   8     1
      -1     2     1     9
run;

/*---------------------------------------------------*
 |  Aufruf des Makros PNORMALK                        |
 *---------------------------------------------------*/
%PNORMALK(10, Mittel, KOVAR, Daten, {'Nsim' 'Klasse' 'X1' 'X2' 'X3''X4'});A3
           Ergebnis
```

A3 Ergebnis