Probleme mit MIXED bei vorgegebener Kovarianzmatrix der Zufallseffekte

Volker Guiard

Forschungsinstitut für die Biologie landwirtschaftlicher Nutztiere
FB Genetik und Biometrie
Wilhelm-Stahl-Allee 2, 18196 Dummerstorf
Telefon: 038208 / 68906
eMail: guiard@fbn-dummerstorf.de

Abstract

Bei der Vorgabe einer Kovarianzmatrix **G** mit Hilfe der **GDATA**-Option im **RANDOM**-Statement der Prozedur **MIXED** ist sorgfältig darauf zu achten, daß die Reihen von **G** mit der Reihenfolge der Stufen der Zufallsfaktoren, also mit den Spalten der Matrix **Z**, korrespondieren. Sollte durch Ausfallwerte eine Stufe eines Zufallsfaktors entfallen, so ist zu klären, wie sich dieses auf die Korrespondenz zwischen den Spalten von **Z** und den Reihen von **G**, bzw. auf das Analyseergebnis auswirkt.

Einleitung

In der Tierzucht kann die Kovarianzmatrix $\mathbf{G}$ der genetischen Effekte aller Tiere durch $\mathbf{G} = \mathbf{A} \cdot \sigma_g^2$ beschrieben werden, wobei $\mathbf{A}$ die bekannte Verwandtschaftsmatrix darstellt und σ_g^2 die unbekannte, zu schätzende genetische Varianz. Ungünstigerweise kann dieses Modell nicht mit der Prozedur **MIXED** behandelt werden, da mit der **GDATA**-Option des **RANDOM**-Befehls eine als völlig bekannt angenommene Matrix $\mathbf{G}$ vorgebbar ist, ein unbekannter Proportionalitätsfaktor σ_g^2 ist nicht zugelassen. Es mag aber sicherlich andere Situationen geben, bei denen von einer bekannten Matrix $\mathbf{G}$ ausgegangen werden kann.

Bei der Vorgabe von $\mathbf{G}$ ist sorgfältig darauf zu achten, dass die Reihen von $\mathbf{G}$ mit der Reihenfolge der Zufallseffekte, also mit den Spalten der Matrix $\mathbf{Z}$ korrespondieren. Die Kontrolle dieser Korrespondenz ist dem Anwender überlassen, sie wird nicht durch **MIXED** unterstützt. Bei der Erzeugung der $\mathbf{G}$-Matrix ist also zu berücksichtigen, dass in der $\mathbf{Z}$-Matrix die Zufallseffekte in der gleichen Reihenfolge erscheinen wie in der **RANDOM**-Anweisung. Weiterhin ergibt sich die Reihenfolge der Stufen eines Faktors aus der nicht immer einfach anzuwendenden **ORDER**-Option. Bei Wechselwirkungseffekten wird durch **MIXED** die Reihenfolge der einzelnen Faktoren einer Wechselwirkung umgeordnet, und zwar in die Reihenfolge mit der diese Faktoren in der **CLASS**-Anweisung angegeben wurden. Die Reihenfolge der einzelnen Faktorstufenkombinationen erhält man, indem man diese derart sortiert, dass der Index der des letzten Faktors am schnellsten und derjenige des ersten Faktors am langsamsten läuft.
Zur Kontrolle der Reihenfolge sollte man in dem **RANDOM**-Befehl die Option **SOLUTION** einfügen, damit gibt **MIXED** die Lösungen für alle Zufallseffekte aus, und zwar in der Reihenfolge der Spalten aus $\mathbf{Z}$.

Auswirkung von nur mit Fehlwerten belegten Faktorstufen

1. Beispielergebnisse ohne Fehlwerte

Im folgenden soll untersucht werden, ob die Korrespondenz zwischen den Reihen von **G** und den Zufallseffekten gestört wird, falls einige Faktorstufen entfallen, da sie nur mit Ausfallwerten besetzt sind. Zur Demonstration verwenden wir ein Beispiel, welches gut überschaubare Ergebnisse liefert. Der Zufallsfaktor B sei dem Zufallsfaktor A hierarchisch untergeordnet und es gelte das Modell (Zufallsgrößen wurden unterstrichen):

$$\underline{y}_{ijk} = \mu + \underline{a}_i + \underline{b}_{ij} + \underline{e}_{ijk}$$

mit $V(\underline{a}_i) = \sigma_a^2$, $V(\underline{b}_{ij}) = \sigma_b^2$ und $V(\underline{e}_{ijk}) = \sigma_R^2$.

Die Daten seien durch die folgende **DATA**-Anweisung gegeben.

```
data sas1;
  input a b c y ;
  cards;
 1 1 101 1
 1 1 101 3
 1 2 102 3
 1 2 102 5
 2 1 201 5
 2 1 201 7
 2 2 202 7
 2 2 202 9
;
run;
```

Die gemäß

```
proc mixed data=sas1 ;
  class a b ;
  model y= ;
  random  a b(a) /  s ;
run;
```

angewendete Prozedur **MIXED** liefert folgende Ergebnisse:

```
          Covariance Parameter Estimates (REML)
              Cov Parm       Estimate
              A            7.00000000
              B(A)         1.00000000
              Residual     2.00000000

              Solution for Random Effects
Effect  A  B      Estimate       SE Pred    DF       t  Pr > |t|
A       1      -1.75000000    1.98431348     4   -0.88    0.4276
A       2       1.75000000    1.98431348     4    0.88    0.4276
B(A)    1  1   -0.62500000    0.85695683     4   -0.73    0.5062
B(A)    1  2    0.37500000    0.85695683     4    0.44    0.6843
B(A)    2  1   -0.37500000    0.85695683     4   -0.44    0.6843
B(A)    2  2    0.62500000    0.85695683     4    0.73    0.5062
```

Wir berechnen nun die Stufenkombinationen (i, j) als Stufen eines zufälligen Faktors C gemäß der folgenden Stufenzuordnungstabelle

a	b	c
1	1	101
1	2	102
2	1	201
2	2	202

Die Effekte der Stufen von C ergeben sich aus der Summe

$$\underline{c}_{i0j} = \underline{a}_i + \underline{b}_{ij}.$$

Der Vektor der vier Werte c_{i0j} $(i = 1,2; j = 1,2)$ hat hier bekanntlich die Kovarianzmatrix

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \sigma_a^2 + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \sigma_b^2.$$

Unterstellen wir, dass die Schätzwerte $\hat{\sigma}_a^2 = 7$ und $\hat{\sigma}_b^2 = 1$ den tatsächlichen Werten der Varianzkomponenten entsprechen, so kann die Kovarianzmatrix $\mathbf{G}$ der c_{i0j} durch den **DATA**-Schritt

```
data g4;
  input row col1-col4
  cards;
1  8  7  0  0
2  7  8  0  0
3  0  0  8  7
4  0  0  7  8
;
run;
```

erzeugt werden. Die mit dem Programm

```
proc mixed data=sas1 ;
  class c ;
  model y= ;
  random  c / gdata=g4 g s ;
run;
```

zu erwartenden Ergebnisse müssten dann den oben genannten entsprechen. Man erhält jedoch anstelle der vorgegebenen **G**-Matrix die Matrix

```
                          G Matrix
Effect  C     Row     COL1       COL2      COL3      COL4
C       101     1    8.000
C       102     2              8.0000
C       201     3                        8.000
C       202     4                                  8.000
```

MIXED übernimmt also seltsamerweise aus der vorgegebenen **G**-Matrix nur die Hauptdiagonale. Dieses entspricht dem Matrixtyp TYPE=SIM, der verwendet wird, falls die TYPE-Option in der RANDOM-Anweisung nicht angegeben wurde. Obwohl der Typ der Matrix **G** sich durch die Vorgabe von **G** ergibt, muss trotzdem der Typ UN (unstrukturiert) angegeben werden, damit die komplette Matrix gelesen wird. Wir ersetzen daher die RANDOM-Zeile durch

```
random c / type=un gdata=g4 g s;
```

und erhalten somit

```
                        G Matrix
Effect  C     Row      COL1      COL2      COL3      COL4
C       101     1     8.000     7.000
C       102     2     7.000     8.000
C       201     3                         8.000     7.000
C       202     4                         7.000     8.000

              Covariance Parameter Estimates (REML)
              Cov Parm       Estimate
              Residual     2.00000000
                   Solution for Random Effects
Effect  C        Estimate       SE Pred    DF        t  Pr > |t|
C       101   -2.37500000    2.05775970     4    -1.15    0.3127
C       102   -1.37500000    2.05775970     4    -0.67    0.5406
C       201    1.37500000    2.05775970     4     0.67    0.5406
C       202    2.37500000    2.05775970     4     1.15    0.3127
```

Die **G**-Matrix wurde nun also korrekt verarbeitet, die Restvarianzschätzung ergibt hier ebenfalls $\hat{\sigma}_R^2 = 2$ und die Lösungen für die Werte c_{i0j} entsprechen den Summen $a_i + b_{ij}$ der mit dem vorherigen Programm erhaltenen Lösungen.

2. Beispielergebnisse mit Fehlwerten

2.1 Berechnung mit reduzierter G-Matrix

Nehmen wir nun an, dass für eine zusätzliche Faktorstufenkombination $(i, j) = (1,3)$ noch zwei weitere Datensätze

```
a b  c  y
1 3 103 .
1 3 103 .
```

vorliegen, bei denen für das Merkmal y jedoch nur Ausfallwerte vorliegen, so könnte man vermuten, dass, da hier im Prinzip die Stufe $(i, j) = (1,3)$ nicht existiert, die **G**-Matrix um die Reihen dieser Stufe zu reduzieren ist, d.h. in unserem Fall, dass die bisher erzeugte **G**-Matrix wieder zum gleichen Ergebnis führt. **MIXED** liefert jedoch in diesem Fall die Fehlerausschrift, dass in der **G**-Matrix nicht 4 sondern 5 Spalten erwartet werden. Verwenden wir die zweite **G**-Darstellung mit den Variablen **ROW**, **COL**, **VALUE**:

```
data g4a;
  input row col value
  cards
1  1  8
2  2  8
3  3  8
4  4  8
1  2  7
2  1  7
3  4  7
4  3  7
;
run;
```

dann liefert **MIXED** das folgende Ergebnis:

G Matrix

Effect	C	Row	COL1	COL2	COL3	COL4	COL5
C	101	1	8.000	7.000			
C	102	2	7.000	8.000			
C	103	3			8.000	7.000	
C	201	4			7.000	8.000	
C	202	5					0.000

Inv(G) Matrix

Effect	C	Row	COL1	COL2	COL3	COL4	COL5
C	101	1	0.533	-0.467			
C	102	2	-0.467	0.533			
C	103	3			0.533	-0.467	
C	201	4			-0.467	0.533	
C	202	5					0.000

Solution for Random Effects

Effect	C	Estimate	SE Pred	DF	t	Pr > |t|
C	101	-4.33288493	1.28904006	4	-3.36	0.0283
C	102	-3.43591014	1.28904006	4	-2.67	0.0561
C	103	-0.91226134	1.80044800	4	-0.51	0.6390
C	201	-1.04258439	1.33602543	4	-0.78	0.4788
C	202	0.00000000	.	.	.	.

Offensichtlich hat **MIXED** nicht vier, sondern fünf Zufallseffekte angenommen und für den letzten Effekt in der $\mathbf{G}$-Matrix Null-Reihen hinzugefügt. Die korrekte Zuordnung zwischen den Reihen von $\mathbf{G}$ und den Spalten von $\mathbf{Z}$ ist somit zerstört. Auch die durch die hier zusätzlich verwendete Option **gi** ermittelte Inverse der Matrix $\mathbf{G}$ zeigt, dass auch bei der weiteren Verarbeitung der $\mathbf{G}$-Matrix dieser Zuordnungsfehler nicht korrigiert wird. Damit gilt:

> *In* **MIXED** *wird für jede in den Daten auftretende Stufe(nkombination) eine Spalte in der* $\mathbf{Z}$*-Matrix angelegt, auch dann, wenn eine Stufe nur mit Ausfallwerten vertreten ist.*

In der Dokumentation zu **MIXED** ist dieses im Abschnitt *"Interaction Effects"* des Kapitels *"Parametrization of Mixed Models"* erwähnt. Es trifft aber genauso auch für *"Main Effects"* zu.

2.2 Berechnung mit nicht reduzierter G-Matrix

Gemäß dem vorherigen Abschnitt ist es also nicht erforderlich, die $\mathbf{G}$-Matrix zu reduzieren, wenn für einige Stufenkombinationen nur Ausfallwerte vorliegen sollten. Es fragt sich jedoch, ob man mit der kompletten $\mathbf{G}$-Matrix die richtigen Ergebnisse erhält, obwohl es zu einigen ihrer Reihen keine Daten (d.h., nur Datensätze mit Fehlwerten) gibt. Die in der Mixed-Model-Gleichung auftretende Inverse der reduzierten $\mathbf{G}$-Matrix ergibt sich nämlich nicht einfach durch Streichung überflüssiger Reihen aus der Inversen der kompletten $\mathbf{G}$-Matrix.
In Modellen der Tierzucht z.B. ist es in diesem Falle üblich, Datensätze, die einen Ausfallwert enthalten, mit einem Gewichtsfaktor Null zu versehen, die anderen Datensätze erhalten den Gewichtsfaktor 1. Ist $\mathbf{W}$ die Diagonalmatrix der Gewichte, so kann man analog zur gewichteten Regression die gewichtete Mixed-Model-Gleichung zu dem Modell

$$\underline{\mathbf{y}} = \mathbf{X}\cdot\mathbf{b} + \mathbf{Z}\cdot\underline{\mathbf{u}} + \underline{\mathbf{e}},\ E(\underline{\mathbf{u}}) = \mathbf{0},\ E(\underline{\mathbf{e}}) = \mathbf{0},\ V(\underline{\mathbf{u}}) = \mathbf{G},\ V(\underline{\mathbf{e}}) = \sigma^2\cdot\mathbf{W}$$

in folgender Weise schreiben:

$$\begin{pmatrix} \mathbf{X'WX} & \mathbf{X'WZ} \\ \mathbf{Z'WX} & \mathbf{Z'WZ} + \mathbf{G}^{-1} \end{pmatrix} \cdot \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'Wy} \\ \mathbf{ZWy} \end{pmatrix}.$$

Durch die Einführung der Gewichtsmatrix $\mathbf{W}$, die für Ausfallwerte nur Nullen enthält, werden in den Matrizen $\mathbf{Y}$, $\mathbf{X}$ und $\mathbf{Z}$ die zu Ausfallwerten gehörenden Zeilen eliminiert. Bezeichnen wir nun die entsprechenden reduzierten Matrizen der Einfachheit halber wieder mit $\mathbf{Y}$, $\mathbf{X}$ und $\mathbf{Z}$, so erhält man wieder die übliche Mixed-Model-Gleichung

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{pmatrix}.$$

Der Vektor $\mathbf{u}$ der Zufallseffekte kann nun in zwei Teilvektoren $\mathbf{u}' = (\mathbf{u}_1', \mathbf{u}_2')$ zerlegt werden, wobei $\mathbf{u}_2$ aus allen Zufallseffekten bestehen soll, zu denen nur Ausfallwerte vorhanden waren. Zerlegt man die $\mathbf{Z}$-Matrix analog in $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$, so ist also $\mathbf{Z}_2$ eine Nullmatrix. In gleicher Weise sind auch die Zeilen und Spalten von $\mathbf{G}$ und $\mathbf{G}^{-1}$ zu gruppieren, so dass man

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} \\ \mathbf{G}_{21} & \mathbf{G}_{22} \end{pmatrix} \text{ bzw. } \mathbf{G}^{-1} = \begin{pmatrix} \mathbf{G}^{(11)} & \mathbf{G}^{(12)} \\ \mathbf{G}^{(21)} & \mathbf{G}^{(22)} \end{pmatrix}$$

erhält, womit die Mixed-Model-Gleichung in drei Matrizengleichungen zerlegt werden kann:

$$\mathbf{X'X}\hat{\mathbf{b}} + \mathbf{X'Z}_1 \cdot \hat{\mathbf{u}}_1 = \mathbf{X'y}$$

$$\mathbf{Z}_1\mathbf{'X}\hat{\mathbf{b}} + (\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{G}^{(11)})\hat{\mathbf{u}}_1 + \mathbf{G}^{(12)} \cdot \hat{\mathbf{u}}_2 = \mathbf{Z}_1\mathbf{'y}$$

$$\mathbf{G}^{(21)} \cdot \hat{\mathbf{u}}_1 + \mathbf{G}^{(22)} \cdot \hat{\mathbf{u}}_2 = \mathbf{0}.$$

Verwendet man nun für Blöcke von $\mathbf{G}^{-1}$ die Formeln der blockweisen Inversion, so gehen die beiden letzten Gleichungen über in

$$\mathbf{Z}_1'\mathbf{X}\hat{\mathbf{b}} + (\mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{G}_{11}^{-1} + \mathbf{G}_{11}^{-1}\mathbf{G}_{12}\mathbf{K}\mathbf{G}_{21}\mathbf{G}_{11}^{-1})\hat{\mathbf{u}}_1 - \mathbf{G}_{11}^{-1}\mathbf{G}_{12}\mathbf{K}\hat{\mathbf{u}}_2 = \mathbf{Z}_1'\mathbf{y}$$

$$-\mathbf{K}\mathbf{G}_{21}\mathbf{G}_{11}^{-1}\hat{\mathbf{u}}_1 + \mathbf{K}\hat{\mathbf{u}}_2 = \mathbf{0}$$

mit $\mathbf{K} = (\mathbf{G}_{22} - \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\mathbf{G}_{12})^{-1}$. Multipliziert man die letzte Gleichung von links mit $\mathbf{G}_{11}^{-1}\mathbf{G}_{12}$ und addiert das Ergebnis zur vorletzten Gleichung, so erhält man die gleiche Mixed-Model-Gleichung

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z}_1 \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{G}_{11}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z}_1'\mathbf{y} \end{pmatrix},$$

wie sie auch entstanden wäre, wenn man vorher die Datensätze mit Ausfallwerten eliminiert und auch nur die entsprechend reduzierte $\mathbf{G}$-Matrix angegeben hätte.

Damit erhält man also für $\hat{\mathbf{b}}$ und $\hat{\mathbf{u}}_1$ korrekte Ergebnisse. Zusätzlich erhält man aus dem letzten Teil der oben genannten Mixed-Model-Gleichung auch für die Zufallseffekte, zu denen keine Beobachtungswerte vorliegen, die Lösung

$$\hat{\mathbf{u}}_2 = \mathbf{G}_{21}\mathbf{G}_{11}^{-1}\hat{\mathbf{u}}_1$$

welche sich lediglich aus der Beziehung zwischen $\mathbf{u}_1$ und $\mathbf{u}_2$ ergibt. Führen wir nun in unserem Beispiel mit Hilfe der zusätzlich in den **DATA**-Schritt eizufügenden Zeile

```
f=1; if not y then f=0;
```

für Ausfallwerte den Gewichtsfaktor 0 und für reelle Werte den Gewichtsfaktor 1 ein und geben die $\mathbf{G}$-Matrix unter Einbeziehung aller in den Daten auftretenden Stufen vor, auch

wenn einige Stufen nur Ausfallwerte enthalten sollten, so entspricht das dem folgenden SAS-Programm.

```
data g5;
  input row col1 -col5;
  cards;
 1  8 7 7 0 0
 2  7 8 7 0 0
 3  7 7 8 0 0
 4  0 0 0 8 7
 5  0 0 0 7 8
;
run;

proc mixed data=sas1 ;
  class c ;
  model y= ;
  random  c / type=un gdata=g5 g s ;
  weight f;
run;
```

und dem Ergebnis

```
                                 G Matrix
Effect    C    Row      COL1      COL2      COL3      COL4      COL5
C       101      1     8.000     7.000     7.000
C       102      2     7.000     8.000     7.000
C       103      3     7.000     7.000     8.000
C       201      4                                 8.000     7.000
C       202      5                                 7.000     8.000

              Covariance Parameter Estimates (REML)
              Cov Parm       Estimate
              Residual     2.00000000

                   Solution for Random Effects
Effect    C       Estimate       SE Pred    DF       t  Pr > |t|
C       101    -2.37500000    2.05775970     4   -1.15    0.3127
C       102    -1.37500000    2.05775970     4   -0.67    0.5406
C       103    -1.75000000    2.22204860     4   -0.79    0.4750
C       201     1.37500000    2.05775970     4    0.67    0.5406
C       202     2.37500000    2.05775970     4    1.15    0.3127
```

Die Lösungswerte und der Standardfehler zu den C-Stufen 101, 102, 201 und 202 entsprechen dem vorher bereits erhaltenen Ergebnis. Hinzu kommt der Vorhersagewert für die C-Stufe 103, die nur mit Ausfallwerten vertreten ist. Der Standardfehler ist hier größer als bei den sonstigen C-Stufen.
Wiederholt man nun diese Rechnung, jedoch ohne die **WEIGHT**-Anweisung, so erhält man das gleiche Ergebnis. D.h., dass die Prozedur **MIXED** bereits intern diese Wichtungstechnik zur Verarbeitung von Ausfallwerten verwendet, sie muss also nicht explizit programmiert werden. Damit ergibt sich folgende

Zusammenfassung

- Treten in den zur Auswertung mit **MIXED** verwendeten Daten Datensätze auf, in denen für die abhängige Variable y nur ein Ausfallwert vorliegt, so werden diese Datensätze durch **MIXED** intern jeweils mit dem Gewichtsfaktor Null versehen.
- Für alle in den Daten vorkommenden Stufen eines Zufallsfaktors werden in der $\mathbf{Z}$-Matrix entsprechende Spalten angelegt, auch wenn die zu einer Stufe vorliegenden Datensätze nur Ausfallwerte enthalten sollten. Diese Struktur der $\mathbf{Z}$-Matrix ist bei der Vorgabe der $\mathbf{G}$-Matrix zu berücksichtigen.
- **MIXED** liefert zunächst das gleiche Ergebnis als wenn alle ausfallwertbehafteten Datensätze fehlen würden. Zusätzlich werden aber auch zu den Zufallseffekten, zu denen nur Ausfallwerte vorliegen, Vorhersagewerte ausgegeben.
- Nach Vorgabe der Matrix $\mathbf{G}$ muss in der **RANDOM**-Anweisung die Option **TYPE=UN** verwendet werden.