MiX99

Solving Large Mixed Model Equations

MTT

MiX99 model derivatives

Martin Lidauer & Ismo Strandén

Preface

- Over the years many useful options have been implemented to provide information for model development and validation
- To demonstrate some of the options let's consider a single trait RR model with 4 fixed effects and 3 RR factors for both permanent environment and additive genetic effect:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zp} + \boldsymbol{Qa} + \boldsymbol{e}$$

where $\boldsymbol{b}$, $\boldsymbol{p}$, $\boldsymbol{a}$, $\boldsymbol{e}$ are vectors of fixed effects, random permanent environmental, additive genetic and residual effects.

The corresponding MME is:

$$\begin{bmatrix} \boldsymbol{X'R^{-1}X} & \boldsymbol{X'R^{-1}Z} & \boldsymbol{X'R^{-1}Q} \\ \boldsymbol{Z'R^{-1}X} & \boldsymbol{Z'R^{-1}Z + P^{-1}} & \boldsymbol{Z'R^{-1}Q} \\ \boldsymbol{Q'R^{-1}X} & \boldsymbol{Q'R^{-1}Z} & \boldsymbol{Q'R^{-1}Q + A^{-1} \otimes G^{-1}} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{p}} \\ \hat{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'R^{-1}y} \\ \boldsymbol{Z'R^{-1}y} \\ \boldsymbol{Q'R^{-1}y} \end{bmatrix}$$

Residuals

- Derivate:

$$\hat{e} = y - X\hat{b} - Z\hat{p} - Q\hat{a}$$

- Specify Y in RESID option
- Residuals will be written to the files ***eHat.data***(*i*),
  - where *i* =0, ..., number of used cores-1
- Order of output lines in the file is the same as in the MiX99 input data file
- Order of output columns follow order of columns with observations in the MiX99 input data file
- Output can be binary of text format
- Missing values are coded by -8192.0

Solver option file

```
# RAM:  RAM demand: high, medium, low
        H
# STOP: Max.Iter., Stopping criteria
        2500        1.0e-5      d     f
# RESID: Calculate residuals?
        Y
# VALID:
        N
# VAROPT
        N
# SOLTYP:
        Y
```

Predicted observations

- Derivate:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{b}} + \boldsymbol{Z}\hat{\boldsymbol{p}} + \boldsymbol{Q}\hat{\boldsymbol{a}}$$

- Specify P in VALID option
- Predictions will be written to the files ***yHat.data****(i)*, which have same setup as ***eHat.data****(i)*

Solver option file

```
# RAM:  RAM demand: high, medium, low
        H
# STOP: Max.Iter., Stopping criteria
        2500        1.0e-5      d     f
# RESID: Calculate residuals?
        N
# VALID:Predictions
        P
# VAROPT
        N
# SOLTYP:
        Y
```

Sum of selected model factors

- Given the order of effects in the model is:
  - 4 fixed effects
  - 3 perm. env. rand. regression effects
  - 3 additive rand. regression effects
- then, e.g. summing of fixed effects ($\hat{\boldsymbol{s}} = \boldsymbol{X}\hat{\boldsymbol{b}}$) requires:
- Specifying S for VALID option
- Specifying 1 for included model factors; order of model factors is same as in CLIM
- Sums will be written to the files ***sHat.data***(*i*), which have same setup as ***eHat.data***(*i*)

Solver option file

```
# RAM:  RAM demand: high, medium, low
        H
# STOP: Max.Iter., Stopping criteria
        2500        1.0e-5      d     f
# RESID: Calculate residuals?
        N
# VALID: Summing fixed effects
        S
# FACTOR: H L1 L2 L3 L4 p1 p2 p3 a1 a2 a3
          1  1  1  1  0  0  0  0  0  0  0
# VAROPT
        N
# SOLTYP:
        Y
```

Yield deviations

- Derivate:
  $$\mathbf{YD} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{b}} - \boldsymbol{Z}\hat{\boldsymbol{p}}$$
- Given, order of model effects are specified in CLIM input as:
  - 4 fixed effects
  - 3 perm. env. rand. regression effects
  - 3 additive rand. regression effects
- Specify Y for VALID
- Specify 1 for model factors included in YD calculation, i.e. genetic effects
- YDs will be written to the files ***YD.data***(*i*), which have same setup as ***eHat.data***(*i*)

Solver option file

```
# RAM: RAM demand: high, medium, low
       H
# STOP: Max.Iter., Stopping criteria
       2500       1.0e-5     d    f
# RESID: Calculate residuals?
       N
# VALID:Yield deviations
       Y
# FACTOR: H L1 L2 L3 L4 p1 p2 p3 a1 a2 a3
          0  0  0  0  0  0  0  0  1  1  1
# VAROPT
       N
# SOLTYP:
       Y
```

Individual daughter deviations

- Derivate:

$$\mathrm{IDD}_{(i)} = \boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{b}} - \boldsymbol{Z}_i\hat{\boldsymbol{p}} - 0.5\boldsymbol{Q}_i\hat{\boldsymbol{a}}_{dam}$$

- Given, order of model effects are specified in CLIM input as:
  - 4 fixed effects
  - 3 perm. env. rand. regression effects
  - 3 additive rand. regression effects
- Specify I for VALID
- Specify 1 for model factors included in IDD calculation, i.e. genetic effects
- IDDs will be written to the files ***IDD.data***(*i*), which have same setup as ***eHat.data***(*i*)

Solver option file

```
# RAM: RAM demand: high, medium, low
       H
# STOP: Max.Iter., Stopping criteria
       2500        1.0e-5      d     f
# RESID: Calculate residuals?
       N
# VALID:Individual daughter deviations
       I
# FACTOR: H L1 L2 L3 L4 p1 p2 p3 a1 a2 a3
       0  0  0  0  0  0  0  0  1  1  1
# VAROPT
       N
# SOLTYP:
       Y
```

Daughter yield deviations

- Implemented as described by **Mrode & Swanson 2004**, (Livest.Prod.Sci.86:), which is an generalization of VanRanden & Wiggan's method (1991, J. Dairy Sci.74:)
- A vector of DYDs for a sire (*i*) is calculated:

$$\mathbf{DYD}_{(\mathrm{i})} = \left(\sum_j \kappa_j \, \boldsymbol{G}^{-1} \mathbf{W}_{\mathrm{ij}} \right)^{-1} \sum_j \kappa_j \, \boldsymbol{G}^{-1} \mathbf{d}_{\mathrm{ij}}$$

- where

$$\mathbf{d}_{\mathrm{ij}} = \left[\left(\mathrm{a}^{\mathrm{jj}} \boldsymbol{G}^{-1} + \boldsymbol{Q}'_j \boldsymbol{R}^{-1} \boldsymbol{Q}_j \right)^{-1} \boldsymbol{Q}'_j \boldsymbol{R}^{-1} \left(\boldsymbol{y}_j - \boldsymbol{X}_j \widehat{\boldsymbol{b}} - \boldsymbol{Z}_j \widehat{\boldsymbol{p}} \right) \right] - 0.5 \widehat{\boldsymbol{a}}_{\mathrm{dam}}$$

$$\mathbf{W}_{\mathrm{ij}} = \left(\mathrm{a}^{\mathrm{jj}} \boldsymbol{G}^{-1} + \boldsymbol{Q}'_j \boldsymbol{R}^{-1} \boldsymbol{Q}_j \right)^{-1} \boldsymbol{Q}'_j \boldsymbol{R}^{-1} \boldsymbol{Q}_j$$

- Note, vector $\mathbf{DYD}_{(\mathrm{i})}$ contains regression coefficients in case of random regression models
- DYDs can be calculated for different daughter groups of a sire, e.g. BY groups

Daughter yield deviations

Instruction for preprocessor mix99i:

- Column indicator for the pedigree file column where the daughter classification variable is given
- IF no classification of DYD within sire, then a 0 is specified
- Block diagonal preconditioner for the additive genetic animal effect

Instruction for solvers mix99s/p:

- Specify D for VALID
- Specify 1 for model factors included in DYD calculation
- DYDs will be written to file **Soldyd** in same format as for the additive genetic animal effect solutions

CLIM syntax

```
.
PEDFILE            /home/ejo31/pedigree 0
PEDIGREE           AN am
.
.
RANDOM             P G
WITHINBLOCKORDER G PE H
PRECON             b d  d  b
MODEL
M1k= H L(1 t2 t3 t4) P(t1 t2 t3|A) G(t1 t2 t3|A)
```

Solver option file

```
.
.
# RESID: Calculate residuals?
         N
# VALID: Daughter yield deviations
         D
# FACTOR: H L1 L2 L3 L4 p1 p2 p3 a1 a2 a3
          0  0  0  0  0  0  0  0  1  1  1
# VAROPT
        N
# SOLTYP:
        Y
```

Generated Observations

- The solver **mix99s** allows to simulate solutions and observations for almost all models possible with MiX99
- Following García-Cortés et al. (1992) true solutions are generated by

$$\tilde{\boldsymbol{b}} = \mathbf{0}$$

$$\tilde{\boldsymbol{p}} = \left(\boldsymbol{I}_{n_p} \otimes \boldsymbol{T}_p\right)\boldsymbol{x}_{n_p}$$

$$\tilde{\boldsymbol{a}} = (\boldsymbol{L} \otimes \boldsymbol{T}_a)\boldsymbol{x}_{n_a}$$

$$\tilde{\boldsymbol{e}} = \left(\boldsymbol{I}_{n_r}\sigma_e^{1/2}\right)\boldsymbol{x}_{n_r}$$

  - where $\boldsymbol{L}$, $\boldsymbol{T}_p$ and $\boldsymbol{T}_a$ are Cholesky decompositions of $\mathbf{A}$ and of the corresponding VCV matrices
  - $\boldsymbol{x}_n \sim \mathrm{N}(\mathbf{0}, \boldsymbol{I}_n)$ are random samples from stand. N.D.
- and observations are formed as:

$$\tilde{\boldsymbol{y}} = \boldsymbol{X}\tilde{\boldsymbol{b}} + \boldsymbol{Z}\tilde{\boldsymbol{p}} + \boldsymbol{Q}\tilde{\boldsymbol{a}} + \tilde{\boldsymbol{e}}$$

Generated Observations

Model and data

- For setting up the model and providing information about data and variance components
  → usual CLIM instructions

mix99s solver options

- Specify G for VALID
- and the type of SEED
  - d … Default (random seed compiler spec.)
  - r … Initialization based on system clock
  - g … Given seeds by user 2 integers
- True solutions will be written to standard solution files
- Generated observations will be written to ***ySIM.data0***

Solver option file

```
# RAM:  RAM demand: high, medium, low
         H
# STOP:  Max.Iter., Stopping criteria
         2500        1.0e-5      d    f
# RESID: Calculate residuals?
        N
# VALID: Generate observations
        G
# VAROPT
        N
# SEED:  Default
        d
# SOLTYP:
        Y
```

Generated Observations

- Generated observations file ***ySIM.data0*** can be read by pre-processor **mix99i** to replace the real observations by generated observations
  - Same data set and model has to be applied as for the MiX99 run to get simulated observations
  - Option not in CLIM yet: Thus, run CLIM to get MiX99_DIR.DIR ...
  - **... and modify MiX99_DIR.DIR**
    specify **g** at VAR line,
    and the location
    of ***ySim.data0***

```
# DATAFILE:
            /home/ejo31/original.data
# VAR:                              replace Y generated Y
              8      1        f                 g
             # directory where ySIM.data0 is located
               /home/ejo31/generatedY
.
```

- Applications so far:
  - for simulation studies based on real data structure
  - re-estimating genetic variances for subsets on animals (using a full model sampling approach based on Mendelian sampling terms)

MiX99 in

DEREGRESSION

Deregression & MiX99

- Need to solve equations that look like mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{-1} \otimes \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

- Breeding values of animals are known: **a**
- General mean is unknown: μ
- Right hand side is unknown: observations **y** will have deregressed proofs

- Need to solve a non-linear system of equations

- Require: random phantom parent group(s)

Base deregression equation system

Unknowns

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1} & \mathbf{0} & \mathbf{0} \\ \mathbf{R}^{-1}\mathbf{X} & \mathbf{R}^{-1}+\mathbf{A}^{bb}\otimes\mathbf{G}_0^{-1} & \mathbf{A}^{ba}\otimes\mathbf{G}_0^{-1} & \mathbf{A}^{bg}\otimes\mathbf{G}_0^{-1} \\ \mathbf{0} & \mathbf{A}^{ab}\otimes\mathbf{G}_0^{-1} & \mathbf{A}^{aa}\otimes\mathbf{G}_0^{-1} & \mathbf{A}^{ag}\otimes\mathbf{G}_0^{-1} \\ \mathbf{0} & \mathbf{A}^{gb}\otimes\mathbf{G}_0^{-1} & \mathbf{A}^{ga}\otimes\mathbf{G}_0^{-1} & \left(\mathbf{A}^{gg}+\mathbf{I}\right)\otimes\mathbf{G}_0^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}} \\ \hat{\mathbf{t}}_b \\ \hat{\mathbf{t}}_a \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_\mu \\ \mathbf{r}_b \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

b: Animals with known EBVs ($\mathbf{a}_b$)

$$\hat{\mathbf{t}}_b = \mathbf{a}_b - \mathbf{X}\hat{\boldsymbol{\mu}}$$

a: Ancestors to animals with known EBVs

g: Random genetic groups

Accelerate solving of general mean

- Methods in MiX99s:
  - None == Gauss-Seidel
  - Bisection
  - Secant
  - Broyden

Data

- Two data sets given by Schaeffer (2001): Country A and B
  - Will consider only country A

- Country A
  - EBVs for 1st, 2nd, 3rd 305-d lactation protein yield
  - 4 analyses: 1, 1+2, 1+2+3 multiple trait,
    1+2+3 as single trait

CLIM instructions

```
TITLE     Multiple trait model

DATAFILE  sch_cntry_A_2.dat # Data file
INTEGER   ones sire         # Integer column names
REAL      w_1 e_1 w_2 e_2 w_3 e_3

DATASORT  PEDIGREECODE=sire
MISSING   -999

PEDFILE   sch_sm.ped
PEDIGREE  sire sm+p 1.0

PARFILE   sch_cntry_A.var

PRECON    b f # Preconditioner: b=block

MODEL
  e_1 = ones sire ! weight=w_1
  e_2 = ones sire ! weight=w_2
  e_3 = ones sire ! weight=w_3
```

		Lactation 1		Lactation 2		Lactation 3	
ones$_1$	sire$_2$	Progeny$_1$	EBV$_2$	Progeny$_3$	EBV$_4$	Progeny$_5$	EBV$_6$
1	12	126	23	0	-999	0	-999
1	12	43	23	43	34	0	-999
1	12	36	23	36	34	36	38
1	13	18	36	0	-999	0	-999
1	13	5	36	5	21	0	-999
1	13	6	36	6	21	6	17
1	14	55	-14	0	-999	0	-999
1	14	21	-14	21	-26	0	-999
1	14	17	-14	17	-26	17	-49
1	15	17	48	0	-999	0	-999
1	15	7	48	7	66	0	-999
1	15	5	48	5	66	5	59
1	16	120	30	0	-999	0	-999
1	16	44	30	44	27	0	-999
1	16	39	30	39	27	39	3

bull$_1$	sire$_2$	maternal grand sire$_3$	maternal grand dam group$_4$
1	-22	-23	-24
2	-22	-23	-24
3	-22	-23	-24
4	-22	-23	-24
5	-22	-23	-24
6	-25	-26	-27
7	-25	-26	-27
8	-25	-26	-27
9	-25	-26	-27
10	-25	-26	-27
11	-25	-26	-27

Random effect$_1$	Row$_2$	Column$_3$	Variance$_1$
1	1	1	96
1	1	2	68
1	1	3	62
1	2	2	160
1	2	3	110
1	3	3	190
2	1	1	1018
2	1	2	128
2	1	3	67
2	2	2	1625
2	2	3	170
2	3	3	1792

Solving by mix99s

- Available solving methods for deregression:
  - Gauss-Seidel
  - Bisection
  - Secant
  - Broyden

F= force to use criteria
1000= max. num. iterations
for non-linear solver

```
H # RAM: RAM demand: H=high, M=medium, L=low
# Max. no. iterations, Stopping criterion, Criterion (A/R/D)
5000                    1.0e-3                    D  F 1000
N # RESID: Calculate residuals? (Y/N)
R b # R=deregression
    # b= Broyden method
    # s= Secant method
    # i= bisection
    # n= Gauss-Seidel
N # adjust for HV?  No
Y # Solution files? Yes
```

Numbers of BLUP solver calls and total number of PCG iterations by analysis

	None	Bisection	Secant	Broyden
BLUP solver calls	169	269	39	7
Num PCG iterations	2197	3497	506	91

Concluding remarks

- Acceleration for non-linear solver has worked very well:
  - No universally best among tested
  - Broyden's method often best when high genetic correlations between traits
  - Secant method best when genetic correlations low
- Convergence affected by definition of genetic groups
  - The more groups the faster convergence
- Random genetic groups essential
- Deregression process almost too easy

THANK YOU