MiX99

Solving Large Mixed Model Equations

MTT

Variance component estimation

Kaarina Matilainen, MTT

Content

- Introduction
- Technical instructions and Example
- Features: update of preconditioner matrix, stopping criterion, standard errors, keeping some parameters unchanged
- Check
- Future development

Introduction

Introduction

- Analytical REML may need a lot of memory for the inversion of the coefficient matrix.
- MCMC methods may be time consuming.
- García-Cortés et. Al (1995) presented a Monte Carlo (MC) method for variance component estimation
  - Idea is to use MC method to estimate the prediction error variances. This is done by repeatedly simulating the data and estimating the location parameters of the simulated data.

$$\hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}^T \mathbf{A}^{-1} \hat{\mathbf{u}} + tr(\mathbf{A}^{-1} \mathbf{C}^{uu})}{q} \quad \Rightarrow \quad \hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}^T \mathbf{A}^{-1} \hat{\mathbf{u}} + \frac{1}{s} \sum_{h=1}^{s} (\tilde{\mathbf{u}}^h - \hat{\mathbf{u}}^h)^T \mathbf{A}^{-1} (\tilde{\mathbf{u}}^h - \hat{\mathbf{u}}^h)}{q}$$

Introduction

- MC EM REML algorithm as presented by Matilainen et. Al (2012) implemented in MiX99.
- Makes REML feasible for large data sets and complex models for which the inversion of the coefficient matrix would be too memory and time consuming.
- Does not support models
  - Which include external correlation structure matrix
  - Which include effect with an autoregressive correlation structure
  - Threshold models
  - Gompertz models

Technical instructions with example

Example

Simple run with simulated data (Matilainen et. Al, 2012)

- Records resembled 305-day milk and fat yields for 3000 animals assigned to 100 herds.
  - 30 animals per herd on average, with a minimum of 13 animals and a maximum of 46 animals in a herd.
- The base generation comprised 150 unrelated sires without observation records. Each base generation sire had 20 daughters, whose dams were assumed to be unknown and unrelated.
- Genetic and residual variance components were chosen to give heritabilities of 0.40 and 0.36 for milk and fat, respectively.

MiX99 instruction file

- Use of CLIM is recommended.
  - File for pre-processor must be named MiX99_DIR.DIR and CLIM creates this automatically.
- Instruction file is written as for breeding value estimation.
- Variance component file consists of starting values (if named as parfile, overwritten during the analysis).

Example

Simple run with simulated data (Matilainen et. Al, 2012)

- simMF.clm

```
TITLE     simulated milk and fat 305-day records

DATAFILE  data.dat
INTEGER   HERD SIRE ANI
REAL      MILK FAT
MISSING   -99999.0

PARFILE   parin
PEDFILE   data.ped
PEDIGREE  ANI am

MODEL
   MILK = HERD ANI
   FAT  = HERD ANI
```

- data.dat

```
1          7        281   -822.8917      -25.74623
1         19        516   -2717.256      -101.0321
1         25        639   -125.0038      -26.74563
1         28        692    752.3760       64.92542
1         37        883    1675.095       82.17327
1         72       1577   -118.8838       10.40209
1         73       1596   -1832.491      -57.13416
1         73       1610   -1269.790      -28.81219
1         79       1728    668.6235      -6.698683
1         86       1859    536.5456       70.56316
.          .          .          .              .
.          .          .          .              .
```

- data.ped

```
   .   . . .
   .   . . .
3141 150 0 0
3142 150 0 0
3143 150 0 0
3144 150 0 0
3145 150 0 0
3146 150 0 0
3147 150 0 0
3148 150 0 0
3149 150 0 0
3150 150 0 0
```

- parin

```
1 1 1 1
1 1 2 0
1 2 2 1
2 1 1 1
2 1 2 0
2 2 2 1
```

MiX99 solver option file

- Option e on the VAROPT line following by three additional lines:
  1. Three entries: Maximum number of REML rounds (integer), Number of MC samples (integer), Stopping cirterion (real). Default values 1000, 5 and 1.0e-9 are suitable in many cases.
  2. Type of seed used by the rundom number generator (D=default initialization of seeds, R=seeds initialized based on the system clock, G=user specified seeds).
  3. Directory path for pre-processor executable

Example

Simple run with simulated data (Matilainen et. Al, 2012)

- simFM.slv

```
# RAM: RAM demand: H=high, M=medium, L=low
   H
# STOP: Maximum_number_of_iterations, Stopping_criteria (CR/CA)
   2000           1.0e-4                d
# RESID: Calculate residuals? (Y/N)
   N
# VALID: N=no, P=prediction, S=sum of effects, Y=YD, D=DYD, I=IDD
   N
# HETVAR: adjust for heterogeneous variance (N, S, C)
   E
      # REML rounds, number of MC samples, stopping criteria
         1000  100  1.0e-9
      # Seed for random number generator (D, R, G)
         D
      # Directory path for pre-processor executable
         /share/apps/common/use/MiX99/release/13.07/
# TYPSOL: type of solution files? (N,Y,A)
   Y
```

Solution files

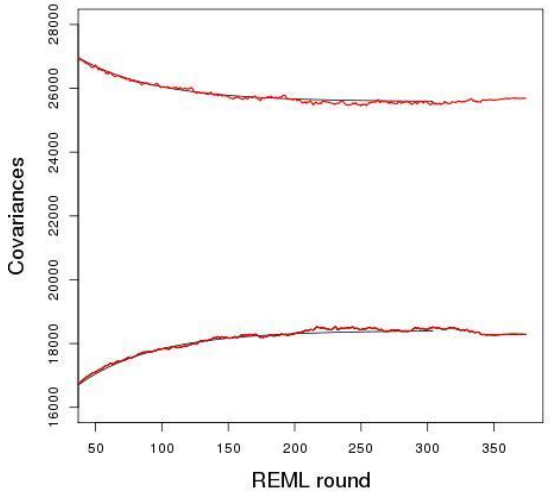
- parfile
  - Contains the lates solutions of variance component estimates.
  - Structure of the file is the same as in the parameter file.
- resfile (when multiple residual variance matrices)
  - Contains the lates solutions of residual variance component estimates.
  - Structure of the file is the same as in the multiple residual variance file.
- REMLlog
  - Contains the estimates of the variance components at every REML round.
  - First column specifies the REML round, after which as many columns as parameters to be estimated.
  - First three lines specify the variance components (as in parameter file).
  - Forth column contains the starting values.

Example

Simple run with simulated data (Matilainen et. Al, 2012)

- parfile

```
1    1    1    611875.47
1    2    1    18523.568
1    2    2    917.38461
2    1    1    677600.11
2    2    1    25493.094
2    2    2    1291.3373
```



- REMLlog

```
  0 1.0000000      1.0000000      1.0000000      2.0000000      2.0000000      2.0000000
  0 1.0000000      2.0000000      2.0000000      1.0000000      2.0000000      2.0000000
  0 1.0000000      1.0000000      2.0000000      1.0000000      1.0000000      2.0000000
  0 1.0000000      0.0000000      1.0000000      1.0000000      0.0000000      1.0000000
  1 267804.45      9206.8835      463.31028      344870.61      11942.134      600.97528
  2 371361.97      12708.859      640.18175      568671.98      19727.522      989.81225
  3 414802.72      14144.115      713.08890      691453.17      24047.361      1207.0739
  4 434874.98      14772.941      744.99125      752167.76      26195.685      1314.5610
  5 445243.33      15083.187      760.65028      779120.92      27171.103      1365.1000
  .         .              .              .              .              .              .
  .         .              .              .              .              .              .
374 602992.59      18285.905      911.48585      683982.71      25686.102      1297.6036
```

Some features

Update of preconditioner matrix

- Updating of the preconditioner matrices was found crucial to enhance convergence.
- In the current version the MiX99 solver will automatically make a system call to start a mix99i pre-processing run, which will update the preconditioner matrices with the most current variance component estimates (because of the system call, instruction file MiX99_DIR.DIR is needed).
  - The updating is done every 10th REML round during the first 100 REML rounds and on every 100th REML round thereafter.
  - If the updating did not succeed, some error messages are printed and the program stops.

Standard errors

- Available since MiX99 version 13.07.
- Calculated by NR-method as in Matilainen et. Al (2013).
- Printed after the last REML round to output only.
- Printed automatically if number of MC samples is at least 10 (50 or more recommended).
- If fewer MC samples used for REML estimates, one additional REML round could be done with user defined number of MC samples.

Example

Simple run with beef cattle data (Matilainen et. Al, 2014)

- Data had 25,220 birth weight observations and 7,715 yearling weight observations.
- Analysis with 10 MC samples gives:

```
---------------------------------------------------------------

 Matrix type   : Information    Matrix number=           1
 Trait         :          11  diagonal= -2.078807938480970E-014

 Approximated information matrix is non positive definite.
 More MC samples may be needed to obtain standard errors
 of variance component estimates.

---------------------------------------------------------------
```

- Run one additional REML round with, for example, 50 MC samples:
  - Change PARFILE in CLIM command file to be parfile of previous run.
  - Change maximum number of REML rounds to be one and number of MC samples to be 50 in the solver option file.

Example

Simple run with beef cattle data (Matilainen et. Al, 2014)

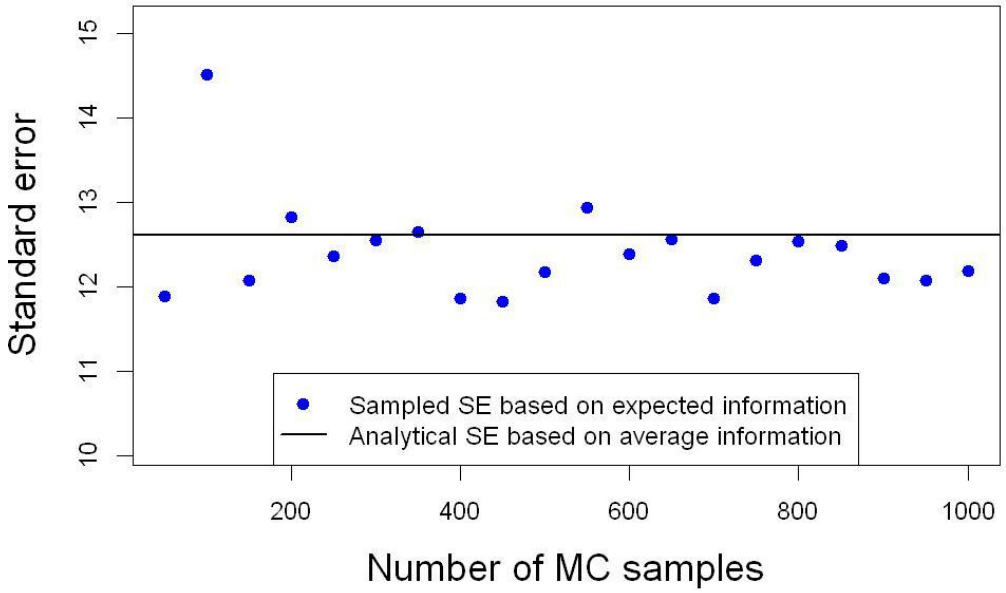
- Analysis with 50 MC samples gives:

```
----------------------------------------------------------------

 Approximated standard errors for MC EM REML estimates:
   1    1    1  0.579623
   1    2    1   4.20352
   1    2    2   76.6673
   2    1    1  0.984137
   2    2    1   6.91980
   2    2    2   88.6640
   2    3    1  0.938816
   2    3    2   8.79753
   2    3    3   1.74581
   2    4    1   11.2420
   2    4    2   104.318
   2    4    3   16.8186
   2    4    4   208.897
   3    1    1   1.02988
   3    2    2   147.731
   3    2    1   10.5421

----------------------------------------------------------------
```

Example

Simple run with beef cattle data (Matilainen et. Al, 2014)

- Approximated SE for direct genetic covariance between birth and yearling weight with different number of MC samples.

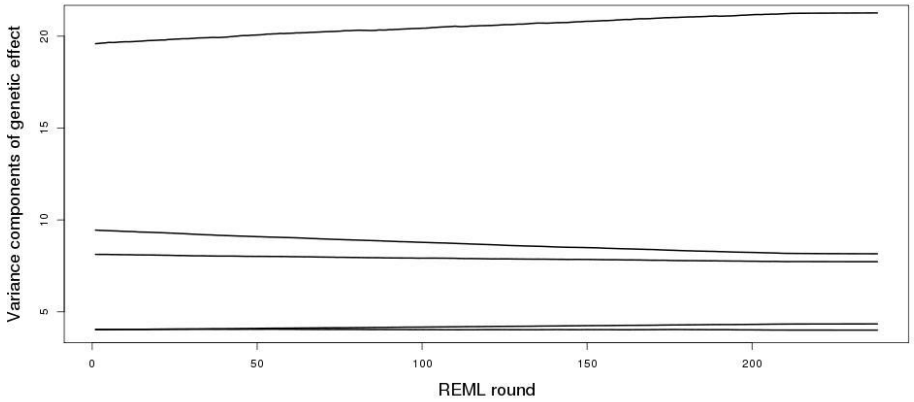


Stopping criteria

- Default stopping criteria for BLUP solutions is suitable in many cases.
- For REML: Convergence indicator which reduce the effect of MC noise in the estimates.
  - Idea:
    1. Linear regression on the latest half of the estimates is fitted.
    2. Variance components at latest REML rounds are predicted by linear regression.
    3. Change in consecutive VC estimates replaced by changes in linear predictions.
  - Suitable values are e.g. 1.0e-8 or 1.0e-9.
  - After specified criterion is reached, additional 30 EM REML rounds are performed (to eliminate MC error by weighted average).

Example
Fertility evaluation

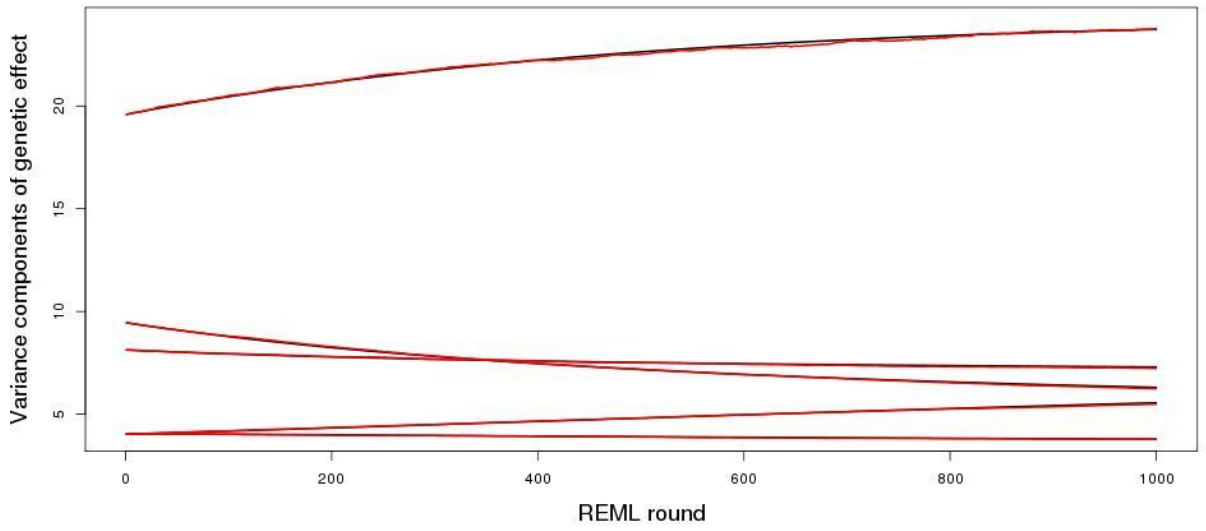
- Five traits with heritabilities 0.01-0.03.
- REML with convergence criteria value 1.0e-9:
  - Stops nicely at REML round 238.
  - However, looks like there is trend for some variance components still.



  - Do another REML run and check whether genetic parameters are changed.

Example
Fertility evaluation

- New REML run with convergence criteria value 1.0e-11 for variance component estimation.
- Stopped for maximum number of REML rounds 1000.



Fixed variance components

- Additional two entries after the option e on the VAROPT line:
  1. Letter f instructs to keep some parameters unchanged.
  2. How many parameters should remain unchanged (integer).
- Insert as many lines as unchanged parameters with three integers:
  1. Random effect number
  2. Row number
  3. Column number
- Variance component value will be the one given in the file of variance components in the MiX99 instruction file.

Example

Simple run with simulated data (Matilainen et. Al, 2012)

- For example, residual variance components are fixed.
  - Change in simMF.slv:

```
E f 3
2 1 1
2 2 1
2 2 2
```

  - parin

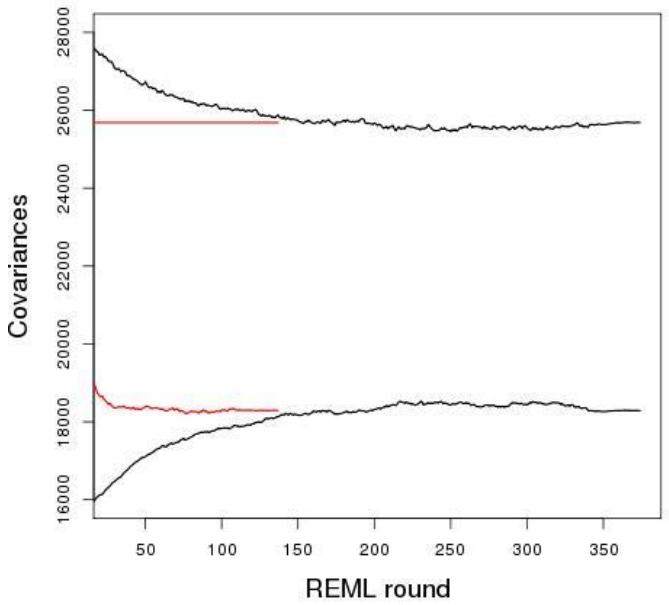
```
1    1    1    900000
1    2    1    0
1    2    2    20000
2    1    1    683982.71
2    2    1    25686.102
2    2    2    1297.6036
```

Example

Simple run with simulated data (Matilainen et. Al, 2012)

- Estimates after 137 REML rounds
  - parfile

```
1    1    1    603158.31
1    2    1    18292.071
1    2    2    912.33121
2    1    1    683982.71
2    2    1    25686.102
2    2    2    1297.6036
```



At the end

Check!

- REML convergence (EM slow especially for small variances and low heritability, difficult data structure)
  - Always check by graph! If there is still some trend in parameter estimates, do another REML run and check whether genetic parameters are changed. In difficult situations, the convergence criteria value for REML estimates may be as strict as 1.0e-11.

Development

- Update of preconditioner more user friendly (so that restriction on MiX99_DIR.DIR file name can be removed).
- Newton-type MC REML for faster convergence.
- To allow models with external correlation structure matrix.

References

- García-Cortés, L.A., Moreno, C., Varona, L., Altarriba, J., 1995. Estimation of prediction-error variances by resampling. J. Anim. Breed. Genet. 112, 176-182.
- Matilainen, K., Mäntysaari, E.A., Lidauer, M.H., Strandén, I., Thompson, R., 2012. Employing a Monte Carlo algorithm in expectation maximization restricted maximum likelihood estimation of the linear mixed model. J. Anim. Breed. Genet., 129, 457-468.
- Matilainen, K., Mäntysaari, E.A., Lidauer, M.H., Strandén, I., Thompson, R., 2013. Employing a Monte Carlo algorithm in Newton-type methods for restricted maximum likelihood estimation of genetic parameters. PLoS ONE 8 12: 1-7.
- Matilainen, K., Strandén, I., Mäntysaari, E.A., 2014. Approximation of standard errors of estimates as a by-product for MC EM REML analysis. In: 10th WCGALP, Canada.