

# Estimating variance components with MC EM REML using MiX99

Anna-Kaisa Ylitalo

MiX99 course: test-day models and single step genomic prediction

COURSE DAY 1, April 10<sup>th</sup>, 2025



# Example, Feed efficiency data

## Statistics:

- Number of records: 22071
- Number of individuals with records: 791
- Number of individuals: 5586
- Number of herds: 4

## For milk yield:

- 22071 observations
- Mean 27.2kg, SD 4.7, min 5.1, max 44.5

## Variables in data file:

- ANI = animal ID
- HERD = herd ID
- HTM = herd x test month
- DIM = days in milk
- cage1 = calving age linear (in years)
- cage2 = calving age quadratic (in years<sup>2</sup>)
- milk = milk yield
- L0 = Legendre: intercept
- L1 = Legendre: linear
- L2 = Legendre: quad.
- L3 = Legendre: cubic
- W005 =  $\exp(-0.05 \cdot \text{DIM})$  (Wilmink function)

# Model

$$milk = \overbrace{cage_1 + cage_2 + f_1(lc, DIM) + HTM}^{\text{fixed}} + \underbrace{f_2(pe, DIM) + f_3(a, DIM) + e}_{\text{random}}$$

where

- milk = milk yield observation
- cage1 = calving age linear (in years)
- cage2 = calving age quadratic (in years<sup>2</sup>)
- $f_1(lc, DIM)$  = fixed lactation curve regression function (for HERD)
- HTM = herd x test month
- $f_2(pe, DIM)$  = random permanent environment regression function (for ANIMAL)
- $f_3(a, DIM)$  = random additive genetic regression function (for ANIMAL)
- e = random residual vector

Functions  $f_1, f_2, f_3$  consists of Legendre polynomials  $\phi_k$  and Wilmink function:

- $f_1(lc, DIM) = \beta_{0k}\phi_0(t) + \beta_{1k}\phi_1(t) + \beta_{2k}\phi_2(t) + \beta_{3k}\phi_3(t) + \beta_{4k}e^{-0.5*t}, k = 1, \dots, 4,$
- $f_2(pe, DIM) = \gamma_{0i}\phi_0(t) + \gamma_{1i}\phi_1(t) + \gamma_{2i}\phi_2(t) + \gamma_{3i}e^{-0.5*t}$
- $f_3(a, DIM) = \alpha_{0i}\phi_0(t) + \alpha_{1i}\phi_1(t) + \alpha_{2i}\phi_2(t) + \alpha_{3i}e^{-0.5*t}$

$$t = 2 \left( \frac{DIM - DIM_{min}}{DIM_{max} - DIM_{min}} \right) - 1$$

# CLIM file

## RR-milk.clm

```
TITLE  RR model for milk using Finnish FE data

DATAFILE  ../data/random_regression_FEdata/data/weekly_FE_FIN062022.dat
PEDFILE   ../data/random_regression_FEdata/data/pedigree.FIN062022.G24

PARFILE   IDENTITY

INTEGER   ANI HERD HTM DIM
REAL      cage1 cage2 milk L0 L1 L2 L3 W005

MISSING   -9999.0000

MODEL
milk = cage1 cage2 LACCRV(L0 L1 L2 L3 W005|HERD) HTM PE(L0 L1 L2 W005| ANI) G(L0 L1 L2 W005| ANI)

RANDOM     PE G

PEDIGREE  G am

DATASORT  BLOCK=HERD PEDIGREECODE=ANI

WITHINBLOCKORDER G PE HTM
              |   |   |   # fixed effects
PRECON       b  b  b      b

TMPDIR     tmpMiX
```



# Solver option file (.slv)

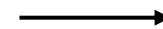
Specific solver options can be specified in a standard input file for MiX99s:

reml.slv

Options for PCG  
( $CD \leq 10^{-5}$ , max  
5000 iterations)



Options for VCE  
estimation ( $CD \leq 10^{-9}$ ,  
2 REML samples, max  
1000 REML iterations)



```
# RAM: RAM demand: H=high, M=medium, L=low
H   nt 10
# STOP: Maximum_number_of_iterations, Stopping_criterion, Criterion (A/R/D)
5000      1.0e-5      d      f
# RESID: Calculate residuals? (Y/N)
N
# VALID: N=no, P=prediction, S=sum of effects, Y=YD, D=DYD, I=IDD, G=generate
N
# VAROPT: (N)o HV, (S)tart HV, (C)ontinue HV, (F)inale, (E)stimation of VC by EM
E
# STOPE maximum number of EM steps, samples/step, convergenc crit.
      1000      2      1.0e-9
# SEED Type of the seed used by the random number generator
      R
# mix99i preprocessor
/home/L1677/bin/
# SOLTYP: Solution files? (N)o, (Y)es, (A)itken, (H)alf-Chebychev
Y
```

Run:

```
mix99i RR-milk.clm > p.log
mix99s < reml.slv > s.log
```

# Checking the output files

- ✓ OK\_mix99s file appeared
- ✓ s.log file

s.log (end)

```
MiX99_SOLVE:      --- D O N E ---           Time: 12:57:38.7  07.04.2025

|| | Copyright(C) 2024 Natural Resources Institute Finland (Luke)
```

REML convergence (next slide)



Random effect:  
1 = PE  
2 = G  
3 = residual

parfile

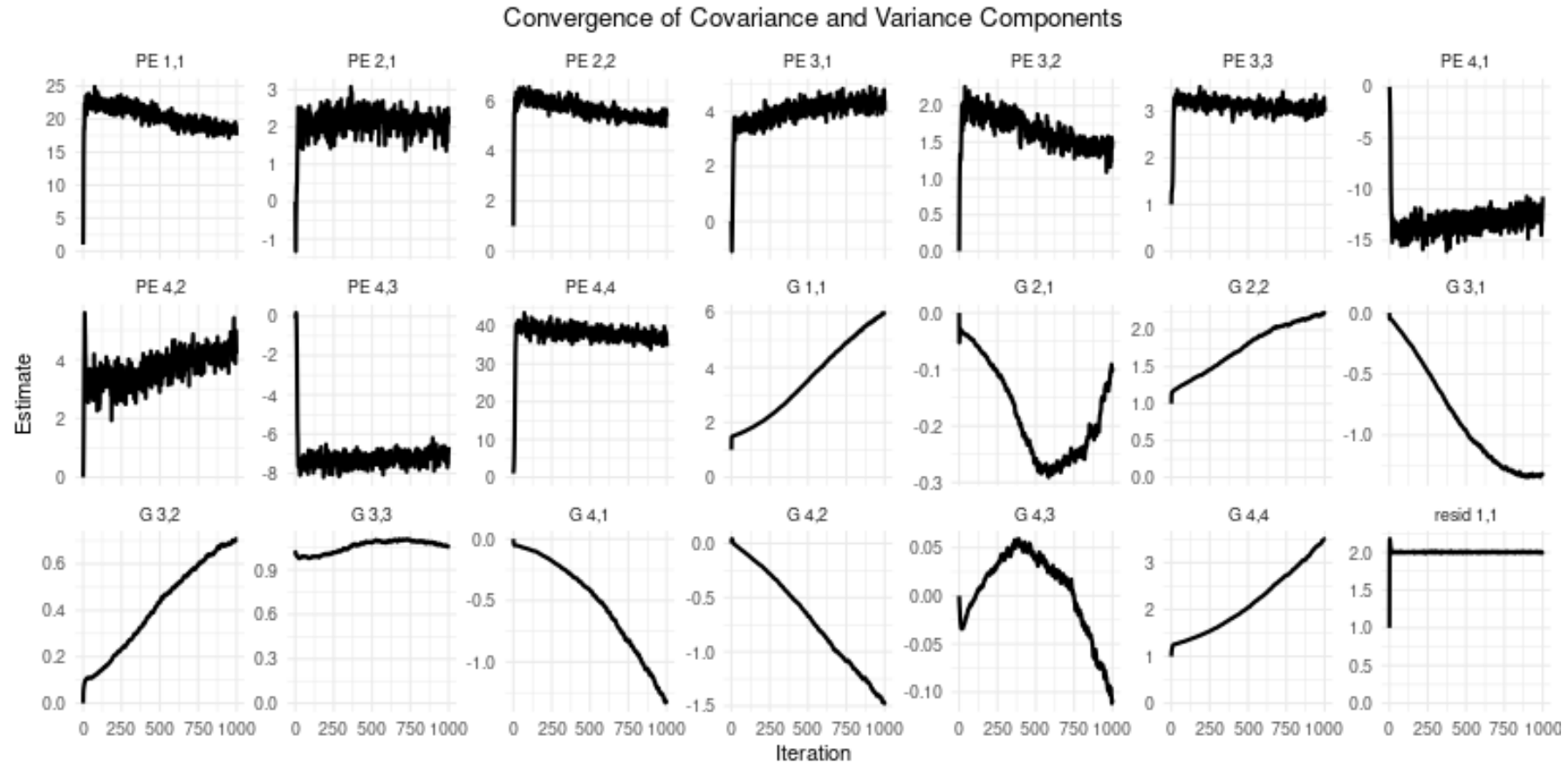
1	1	1	17.533733
1	2	1	2.2442572
1	2	2	5.3533178
1	3	1	4.0155096
1	3	2	1.5013200
1	3	3	3.0039102
1	4	1	-10.740962
1	4	2	4.1961244
1	4	3	-6.6513802
1	4	4	34.635314
2	1	1	5.9906178
2	2	1	-0.96667632E-01
2	2	2	2.2249144
2	3	1	-1.3218683
2	3	2	0.70551988
2	3	3	1.0569965
2	4	1	-1.3359755
2	4	2	-1.4700748
2	4	3	-0.10867647
2	4	4	3.5041189
3	1	1	2.0003345

VC estimates

Row and column indices of  
the variance-covariance matrix

1 = L0      3 = L2  
2 = L1      4 = W005

# Convergence based on REMLlog



✘ REML convergence

# Output files

s.log

```
66  0.3244E-06  0.1137E-06  0.2984E-06  0.7888E-05  -0.1643E-03
67  0.2828E-06  0.6335E-07  0.2571E-06  0.8751E-05  -0.1809E-03
Stopping criterion CD < 0.1E-4 achieved in 67 iterations.
NOTE: CD criterion must be met by two consecutive iterations!

Solutions have converged according to CD criterion of the last iteration.

-----

MiX99_SOLVE: End of PCG Iteration           Time: 12:57:34.1  07.04.2025

-----

Last PCG iteration of data sample:    1
243  0.1307E-05  0.4370E-05  0.1686E-05  0.8242E-05  0.5256E-04
Last PCG iteration of data sample:    2
274  0.1129E-05  0.1474E-05  0.1482E-05  0.9257E-05  -0.7316E-04

REML ROUND 1000  CONV Cd  0.3245E-07

-----

MiX99_SOLVE: End of MC EM REML iteration   Time: 12:57:37.5  07.04.2025

-----

REML convergence criterion 0.1E-8 was _not_ achieved in 1000 rounds!
```



# How to continue VCE (REML iterations)

reml.slv

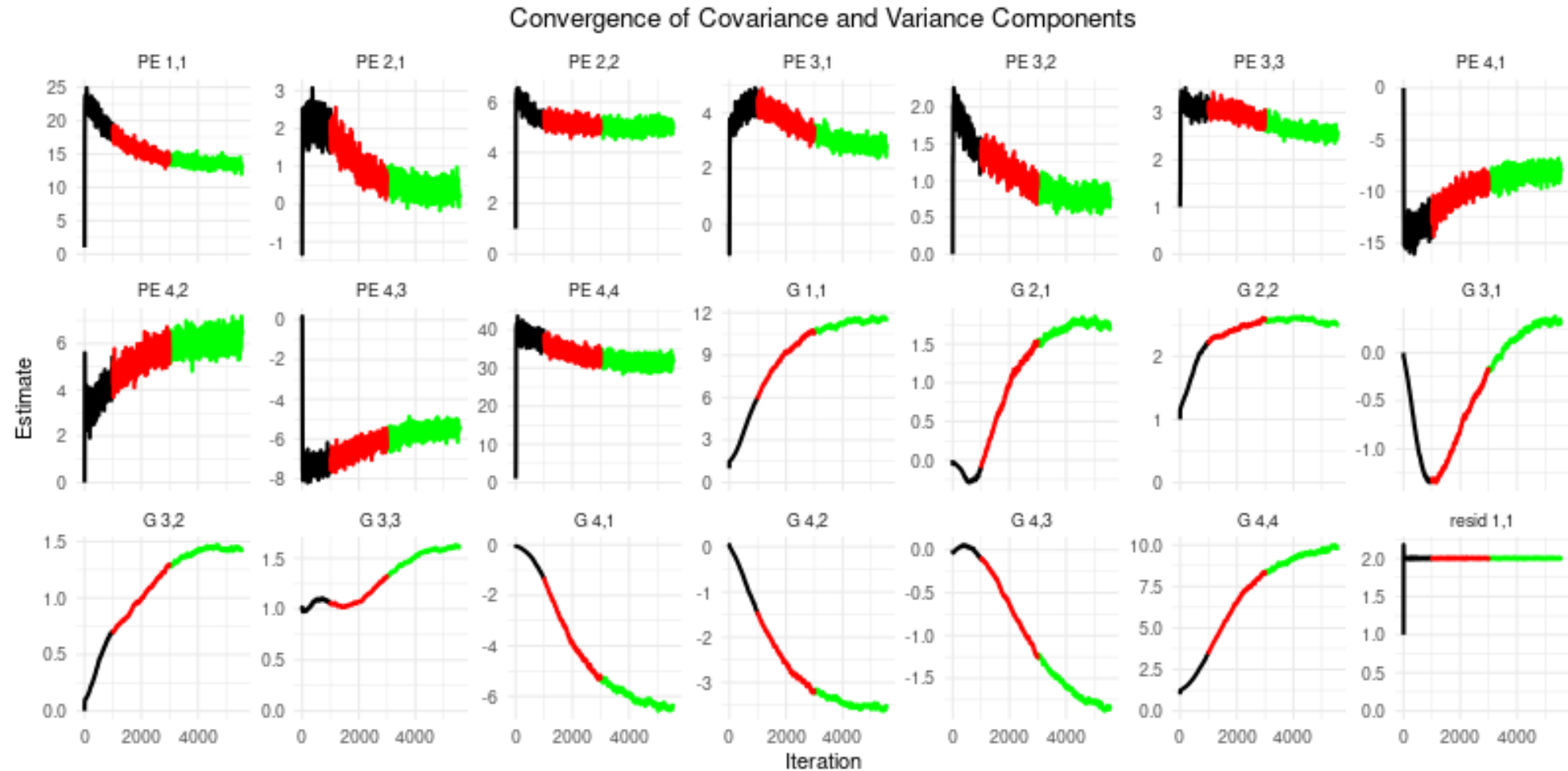
```
# RAM: RAM demand: H=high, M=medium, L=low
H   nt 10
# STOP: Maximum_number_of_iterations, Stopping_criterion, Criterion (A/R/D)
5000      1.0e-5      d      f
# RESID: Calculate residuals? (Y/N)
N
# VALID: N=no, P=prediction, S=sum of effects, Y=YD, D=DYD, I=IDD, G=generate
N
# VAROPT: (N)o HV, (S)tart HV, (C)ontinue HV, (F)inale, (E)stimation of VC by EM
E
# STOPE maximum number of EM steps, samples/step, conv.c crit. REML, PCG conv. crit. sampled data
-3000      2      1.0e-9      1.0e-4
# SEED Type of the seed used by the random number generator
R
# mix99i preprocessor
/home/L1677/bin/
# SOLTYP: Solution files? (N)o, (Y)es, (A)itken, (H)alf-Chebyshev
Y
```

In the upcoming version of MiX99, one can set a different convergence criteria for the sampled data (in REML)

We run already 1000 REML rounds,  
let's continue 2000 more:

Put -3000 ( = 1000 + 2000)

# Convergence based on REMLlog



REML convergence criterion ( $0.1E-8$ ) was achieved in 5533 rounds and additional 30 rounds were performed to reduce variation in variance component estimates.

s.log

# Variance component estimates

Random effect:  
1 = PE  
2 = G  
3 = residual

parfile

1	1	1	13.195983
1	2	1	0.28868815
1	2	2	4.9370093
1	3	1	2.7890333
1	3	2	0.76124452
1	3	3	2.5466688
1	4	1	-7.8752794
1	4	2	6.2571967
1	4	3	-5.4354612
1	4	4	31.262229
2	1	1	11.552537
2	2	1	1.7052058
2	2	2	2.4986931
2	3	1	0.31509603
2	3	2	1.4265949
2	3	3	1.6141484
2	4	1	-6.3919248
2	4	2	-3.5331418
2	4	3	-1.8408422
2	4	4	9.8103408
3	1	1	2.0013035

VC estimates

Row and column  
indices of the variance-  
covariance matrix

# Example on calculating standard errors

## New CLIM file (obs! not fully visible here)

RR-milk\_SE.clm

```
TITLE  RR model for milk using Finnish FE data

DATAFILE  ../data/random_regression_FEdata/data/weekly_FE_FIN062022.dat
PEDFILE   ../data/random_regression_FEdata/data/pedigree.FIN062022.G24
PARFILE   par-file

INTEGER   ANI HERD HTM DIM
REAL      cage1 cage2 milk L0 L1 L2 L3 W005

MISSING   -9999.0000

MODEL
milk = cage1 cage2 LACCRV(L0 L1 L2 L3 W005|HERD) HTM PE(L0 L1 L2 W005| ANI) G(L0 L1 L2 W005| ANI)

RANDOM    PE G
```

## New solver option file

reml\_SE.slv

```
# RAM: RAM demand: H=high, M=medium, L=low
| H nt 10
# STOP: Maximum_number_of_iterations, Stopping_criterion, Criterion (A/R/D)
| 5000 1.0e-5 d f
# RESID: Calculate residuals? (Y/N)
| N
# VALID: N=no, P=prediction, S=sum of effects, Y=YD, D=DYD, I=IDD, G=generate
| N
# VAROPT: (N)o HV, (S)tart HV, (C)ontinue HV, (F)inale, (E)stimation of VC by EM
| E
# STOPE maximum number of EM steps, samples/step, conv.c crit. REML, PCG conv. crit. sampled data
| 1 50 1.0e-9 1.0e-4
# SEED Type of the seed used by the random number generator
| R
# mix99i preprocessor
| /home/L1677/bin/
# SOLTYP: Solution files? (N)o, (Y)es, (A)itken, (H)alf-Chebyshev
| Y
```

Run:

```
mix99i RR-milk_SE.clm > p_SE.log
mix99s < reml_SE.slv > s_SE.log
```



# Output files

vceSE (standard errors)

```

1 1 1 2.31003
1 2 1 1.63365
1 2 2 1.30376
1 3 1 1.03933
1 3 2 0.822907
1 3 3 0.924752
1 4 1 3.14761
1 4 2 1.92922
1 4 3 2.34946
1 4 4 6.58858
2 1 1 3.18296
2 2 1 1.58406
2 2 2 1.51478
2 3 1 1.27455
2 3 2 0.977628
2 3 3 0.982481
2 4 1 3.27822
2 4 2 2.20557
2 4 3 2.24556
2 4 4 5.90200
3 1 1 0.286625E-01

```

vcel (information matrix)

```

1 1 1 1 1 1 5.33624
1 2 1 1 1 1 0.498903
1 2 1 1 2 1 2.66880
1 2 2 1 1 1 0.900258
1 2 2 1 2 1 0.241065
1 2 2 1 2 2 1.69978
1 3 1 1 1 1 1.08426
1 3 1 1 2 1 0.947213
1 3 1 1 2 2 0.132760
1 3 1 1 3 1 1.08020
1 3 2 1 1 1 0.729351
1 3 2 1 2 1 0.174896
1 3 2 1 2 2 0.657436
1 3 2 1 3 1 0.364975
1 3 2 1 3 2 0.677176
1 3 3 1 1 1 0.682743
1 3 3 1 2 1 -0.265465
1 3 3 1 2 2 0.268006
1 3 3 1 3 1 0.307752
1 3 3 1 3 2 0.538390
1 3 3 1 3 3 0.855167
1 4 1 1 1 1 -3.85905
1 4 1 1 2 1 1.05917
1 4 1 1 2 2 -0.374665
1 4 1 1 3 1 -1.59251
1 4 1 1 3 2 -0.784957
1 4 1 1 3 3 -1.23127
1 4 1 1 4 1 9.90747
1 4 2 1 1 1 -0.694099

```

# (Toy) example on how to fix variance parameters

Let's fix the variances of the genomic effect

reml\_pfix.slv

```
# RAM: RAM demand: H=high, M=medium, L=low
H nt 10
# STOP: Maximum_number_of_iterations, Stopping_criterion, Criterion (A/R/D)
5000 1.0e-5 d f
# RESID: Calculate residuals? (Y/N)
N
# VALID: N=no, P=prediction, S=sum of effects, Y=YD, D=DYD, I=IDD, G=generate
N
# VAROPT: (N)o HV, (S)tart HV, (C)ontinue HV, (F)inale, (E)stimation of VC by EM
E f 4
2 1 1
2 2 2
2 3 3
2 4 4
# STOPE maximum number of EM steps, samples/step, conv.c crit. REML, PCG conv. crit. sampled data
1000 2 1.0e-9 1.0e-4
# SEED Type of the seed used by the random number generator
R
# mix99i preprocessor
/home/L1677/bin/
# SOLTYP: Solution files? (N)o, (Y)es, (A)itken, (H)alf-Chebyshev
Y
```

Fix 4 parameters

Random effect  
number, row and  
column indices

parfile

```
1 1 1 13.195983
1 2 1 0.28868815
1 2 2 4.9370093
1 3 1 2.7890333
1 3 2 0.76124452
1 3 3 2.5466688
1 4 1 -7.8752794
1 4 2 6.2571967
1 4 3 -5.4354612
1 4 4 31.262229
2 1 1 11.552537
2 2 1 1.7052058
2 2 2 2.4986931
2 3 1 0.31509603
2 3 2 1.4265949
2 3 3 1.6141484
2 4 1 -6.3919248
2 4 2 -3.5331418
2 4 3 -1.8408422
2 4 4 9.8103408
3 1 1 2.0013035
```

Run:

mix99s < reml\_pfix.slv > s\_pfix.log

# Estimated variance components

new VC estimates (parfile)				old VCE
1	1	1	13.691633	13.195983
1	2	1	0.41603627	0.28868815
1	2	2	5.0525610	4.9370093
1	3	1	2.7791162	2.7890333
1	3	2	0.81702440	0.76124452
1	3	3	2.5028028	2.5466688
1	4	1	-7.7716162	-7.8752794
1	4	2	6.2585433	6.2571967
1	4	3	-5.2570694	-5.4354612
1	4	4	30.649230	31.262229
2	1	1	<u>11.552537</u>	<u>11.552537</u>
2	2	1	1.6820589	1.7052058
2	2	2	<u>2.4986931</u>	<u>2.4986931</u>
2	3	1	0.30111517	0.31509603
2	3	2	1.4093865	1.4265949
2	3	3	<u>1.6141484</u>	<u>1.6141484</u>
2	4	1	-6.4649512	-6.3919248
2	4	2	-3.5630788	-3.5331418
2	4	3	-1.8495405	-1.8408422
2	4	4	<u>9.8103408</u>	<u>9.8103408</u>
3	1	1	2.0032369	2.0013035