Estimation of variance components with MiX99

Napo Vargas Jurado

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

COURSE DAY 1, April 10th, 2025

Brief REML background

Introduction

- Estimation of variance components is a key component
  - Genetic and genomic evaluation
  - Expected rates of genetic gains
  - Development of selection indexes
- Not a trivial task for random regression models
  - Many parameters to estimate
- Even more computationally intensive with genomic models
  - GBLUP
- MiX99 uses a very efficient strategy

Luke

(Very brief) REML background

- Idea: maximize the likelihood after accounting for the fixed effects.
- Model assumptions (single trait):

$$\begin{aligned}\mathbf{y} &= \mathbf{X\beta} + \mathbf{Zu} + \mathbf{e}\\ \mathrm{Var}[\mathbf{y}] &= \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}\\ \mathrm{Var}[\mathbf{u}] &= \mathbf{G} = \sigma_u^2\mathbf{A}\\ \mathrm{Var}[\mathbf{e}] &= \mathbf{R} = \sigma_e^2\mathbf{I}\end{aligned}$$

- Likelihood function:

$$\ell(\boldsymbol{\theta}) = constant + \mathbf{y}'\mathbf{P}\mathbf{y} - \log|\mathbf{V}| - \log(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})$$
$$\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$$

- $\boldsymbol{\theta}$ is the vector of VC parameters (σ_u^2, σ_e^2)
  - Estimated by differentiating and solving the score (likelihood) function

REML background (cont.)

- Recall mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \frac{1}{\sigma_u^2}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

- In a univariate model the REML updates are given by:

$$\hat{\sigma}_u^2 = \frac{1}{q}\left[\hat{\mathbf{u}}'\mathbf{A}^{-1}\hat{\mathbf{u}} + \mathrm{tr}(\mathbf{A}^{-1}\mathbf{C}^{uu})\right]$$

$$\hat{\sigma}_e^2 = \frac{1}{n}\left[\hat{\mathbf{e}}'\hat{\mathbf{e}} + \mathrm{tr}(\mathbf{W}\mathbf{C}\mathbf{W}')\right]$$

- $\mathbf{C}$ is the inverse of the coefficient matrix (very large in RRM), $\mathbf{W} = [\mathbf{X}\ \mathbf{Z}]$
  - Permanent environmental and genetic effects

Monte Carlo EM-REML

- Motivation:
  - Inverse of coefficient matrix ($\mathbf{C}$) is computationally demanding (cubic cost w.r.t. number of eq.)
  - $\mathbf{C}$ needs to be calculated at each iteration (not efficient)
- Generate "pseudo" records according to:

$$\tilde{\mathbf{y}}^h = \mathbf{X}\tilde{\boldsymbol{\beta}}^h + \mathbf{Z}\tilde{\mathbf{u}}^h + \tilde{\mathbf{e}}^h$$
$$\tilde{\mathbf{u}}^h \sim N(\mathbf{0}, \hat{\sigma}^2_u \mathbf{A})$$
$$\tilde{\mathbf{e}}^h \sim N(\mathbf{0}, \hat{\sigma}^2_e \mathbf{I})$$
$$h = 1, \ldots, s$$

- Estimate $\hat{\mathbf{u}}$ by solving MME with pseudo-records using PCG
  - Memory efficient
  - Calculate residuals

Luke

MC EM-REML (cont.)

- Approximate traces by

$$\mathrm{tr}(\mathbf{A}^{-1}\mathbf{C}^{uu}) \approx \frac{1}{s}\sum_{h=1}^{s} q\hat{\sigma}_u^2 - \hat{\mathbf{u}}^{h\prime}\mathbf{A}^{-1}\hat{\mathbf{u}}^h$$

$$\mathrm{tr}(\mathbf{WCW}') \approx \frac{1}{s}\sum_{h=1}^{s} n\hat{\sigma}_e^2 - \hat{\mathbf{e}}^{h\prime}\hat{\mathbf{e}}^h$$

- Then update VC

$$\hat{\sigma}_u^2 = \frac{1}{q}\left[\hat{\mathbf{u}}'\mathbf{A}^{-1}\hat{\mathbf{u}} + \mathrm{tr}(\mathbf{A}^{-1}\mathbf{C}^{uu})\right]$$

$$\hat{\sigma}_e^2 = \frac{1}{n}\left[\hat{\mathbf{e}}'\hat{\mathbf{e}} + \mathrm{tr}(\mathbf{WCW}')\right]$$

Some properties of the MC EM REML method

- Using MC method for approximating traces does not require explicit inversion of LHS
  - Can handle larger models that plain EM-REML cannot
- The number of MC samples (*s*) need not be too large (5 – 25) for a good approximation
  - More MC samples improve accuracy of approximation - slower
- There is an associated MC error (especially with smaller *s*)
- More difficult to assess convergence
  - Useful to confirm visually that this is the case

In MiX99

VCE In MiX99

- Need an additional solver file (.slv)
  - Maximum number of iterations
  - Convergence criteria
  - Number of Monte Carlo samples
  - Random number generator
- Parameters to keep fixed during estimation
  - Kept at value provided in parameter file
- Can use genomic information (GBLUP)
  - Need to modify CLIM file accordingly

Luke

VCE using solver option file (.slv)

- Option E on the VAROPT line followed by three additional lines:
  - STOPE Maximum number of REML rounds, Number of MC samples, Stopping criterion.
    - Default values: 1000, 5 and 1.0e-9 suitable in many cases
  - SEED Type of the seed used by the random number generator
    - D=default initialization of seeds
    - R=seeds initialized based on the system clock
    - G=user specified seeds
  - Directory path for MiX99 preprocessor

Luke

Example of solver option file (reml.slv)

reml.slv

```
# RAM: RAM demand: H=high, M=medium, L=low
   H
# 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
```

Keeping variance components fixed

- Some components may need to be fixed for the model to be estimable
  - RRM with sire as the main genetic effect
  - PE and residual variances not estimable, need to fix residuals to a small value
- 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 the number of parameters to keep fixed
  - Random effect number, row, column, value
- Variance component will be fixed to the values provided in the PARFILE and given in the CLIM file

Luke

Example .slv file

Fixing 42 components:

```
# RAM: RAM demand: H=high, M=medium, L=low
   H
# STOP: Max.num. iterations, Stopping criterion, Convergence indicator, enforce
   5000                          5.0e-5                  d                  f
# RESID: Calculate residuals? (Y/N)
   N
# VALID: N=none, P=prediction, S=sum of effects, Y=YD, D=DYD, I=IDD
   N
# VAROPT: adjust for heterogeneous variance (N, E, S, C)
   E f 42
   4  1  1  0.0100000000000
   4  2  1  0.0000000000000
   4  2  2  0.0100000000000
   4  3  1  0.0000000000000
   4  3  2  0.0000000000000
   4  4  1  0.0000000000000
   4  4  2  0.0000000000000
   4  4  3  0.0000000000000
   4  4  4  0.0100000000000
   4  5  1  0.0000000000000
   4  5  2  0.0000000000000
   4  5  3  0.0000000000000
   4  5  4  0.0000000000000
   4  5  5  0.0100000000000
```

VCE with genomic information

- VCE can be estimated from a GBLUP model
- CLIM file needs to be modified as:
  - GBLUP: keyword
  - Random effect identifier (as given in the model)
  - Inverse of genomic relationship matrix (lower triangle)

```
MODEL
Yield1 = YxT PE G(ID)
Yield2 = YxT PE G(ID)
Yield3 = YxT PE G(ID)
WntDmg = YxT PE G(ID)
Dval1  = YxT PE G(ID)
Dval2  = YxT PE G(ID)

GBLUP G GInverse.dat
RANDOM PE G
WITHINBLOCKORDER G
```

Luke

Output files

- parfile: contains the latest solutions of variance component estimates
  - Lower (or upper) triangle
  - Effect, row, column, value
- REMLlog: contains the VC estimates at each REML round
  - Convergence criterion
  - Can track changes in convergence of each component
- .log files: it's a good practice to save the log files (from the preprocessor and solver) and check them!

Luke

Standard errors

- Define $\boldsymbol{\theta}$ as the vector of VC, then the variance of the estimates is:

$$Var\left[\widehat{\boldsymbol{\theta}}\right] = \boldsymbol{\mathcal{I}}$$
$$SE\left(\widehat{\boldsymbol{\theta}}_i\right) = \sqrt{\boldsymbol{\mathcal{I}}_{ii}}$$

- $\boldsymbol{\mathcal{I}}$ is the Information matrix (large for RRM)
  - 162 by 162 (13203 elements) in example
- MiX99 can produce standard errors for the variance components in the model
  - Information matrix can be approximated by the variance of the score functions over MC samples
  - Need a larger number of MC samples to get an appropriate estimate

Luke

Standard errors (cont.)

- Run MiX99 again with small changes in .clm and .slv files:
  1) Use the final VC estimates as the starting values (PARFILE in CLIM)
  2) Run only 1 REML round
  3) Use a large MC sample size (e.g. 100). For RRM may need more samples
- As an output, you get two files:
  - vceSE: includes standard errors of the VC estimates
  - vcel: information matrix

Luke

SE for VCE .slv file

```
# RAM: RAM demand: H=high, M=medim, L=low
   H
# STOP: Maximum_number_of_iterations, Stopping_criterion, Criterion(A/R/D)
   5000          1.0e-5              d       f
# RESID: Calculate residuals? (Y/N)
   Y
# 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 VC by EM
   E       #  f  l
   # STOPE maximum number of EM steps, samples/step, convergenc crit,
         1              500                          1.0e-11
   # RANDG random number generator
           R
   # mix99i
/data/projects/animalgenetics/software/MiX99/develop/22.12/

# TYPSOL: Solution files? (N)o, (Y)es, (A)itken, (H)alf-Chebychev
   Y
```

Example

- Simulated data
- Three-trait RRM from first lactation
  - Milk
  - Protein
  - Fat
- 5586 animals in pedigree
- 22071 test-day records

Example - model

- RRM with intercept (0), linear (1), quadratic (2) and Wilmink (3) terms
- Model

$$y = AGE + AGE^2 + HTM + \sum_{j=0}^{4} \beta_j \phi_j + \sum_{j=0}^{3} \alpha_j \phi_j + \sum_{j=0}^{3} \gamma_j \phi_j + \epsilon$$

$$\text{Var}\begin{bmatrix} \alpha_{0_M} \\ \alpha_{0_P} \\ \alpha_{0_F} \\ \vdots \\ \alpha_{3_M} \\ \alpha_{3_P} \\ \alpha_{3_F} \end{bmatrix} = \mathbf{K}_\alpha, \text{Var}\begin{bmatrix} \gamma_{0_M} \\ \gamma_{0_P} \\ \gamma_{0_F} \\ \vdots \\ \gamma_{3_M} \\ \gamma_{3_P} \\ \gamma_{3_F} \end{bmatrix} = \mathbf{K}_\gamma, Var[\epsilon] = \mathbf{R}$$

- $\mathbf{K}_\alpha$ and $\mathbf{K}_\gamma$ are 12 by 12 covariance matrices (78 elements each)
- **R** is a 3 by 2 matrix

Data

- Integers:
  - ID, herd, herd-year-season, DIM, mean (intercept)
- Reals:
  - Calving age (linear, quadratic), milk yield, protein, fat, LP 0, LP 1, LP 2, LP 3, Wilmink

```
1638 1 111 119981 103 11998091 1 1 45 7 2 1 -0.2493 0.0622 5.527467 5.682810 3.262739 0.7071 -0.9526 0.6442 -0.018 0.1054
1638 1 111 119981 104 11998101 1 1 52 8 3 1 -0.2493 0.0622 3.674568 3.834023 2.667054 0.7071 -0.905 0.5043 0.1868 0.0743
1638 1 111 119981 104 11998101 1 1 59 9 3 1 -0.2493 0.0622 2.751818 3.479482 1.945793 0.7071 -0.8573 0.3716 0.3601 0.0523
1638 1 111 119981 104 11998101 1 1 66 10 3 1 -0.2493 0.0622 2.973819 4.095433 1.867726 0.7071 -0.8097 0.246 0.5038 0.0369
1638 1 111 119981 104 11998101 1 1 73 11 3 1 -0.2493 0.0622 5.274148 5.858224 4.220061 0.7071 -0.7621 0.1277 0.6194 0.026
1638 1 111 119981 104 11998101 1 1 80 12 4 1 -0.2493 0.0622 2.026425 1.863144 3.524207 0.7071 -0.7144 0.0165 0.7086 0.0183
1638 1 111 119981 104 11998111 1 1 87 13 4 1 -0.2493 0.0622 3.500561 4.065313 3.677488 0.7071 -0.6668 -0.0875 0.773 0.0129
1638 1 111 119981 104 11998111 1 1 94 14 4 1 -0.2493 0.0622 2.779573 4.226952 2.999407 0.7071 -0.6192 -0.1844 0.8144 0.0091
1638 1 111 119981 104 11998111 1 1 101 15 4 1 -0.2493 0.0622 5.680336 6.364255 3.332927 0.7071 -0.5715 -0.2741 0.8343 0.0064
1638 1 111 119981 104 11998111 1 1 108 16 5 1 -0.2493 0.0622 3.309149 3.234915 3.102585 0.7071 -0.5239 -0.3566 0.8343 0.0045
```

Luke

Pedigree file

- ID
- Sire
- Dam
- Code

```
1638 1314 1500 1
1639 1640 648 1
1641 1642 1435 1
1643 1644 1373 1
1672 1673 1569 1
1708 1709 1707 1
1738 1314 1352 1
1739 1740 1616 1
1749 1314 1263 1
1774 1377 1581 1
```

Luke

Covariable table file

- DIM
- LP: intercept, linear, quadratic, cubic, Wilmink (t1 to t5 in CLIM)

```
5 0.707106781186547 -1.224744871391589 1.581138830084 19 -1.870828693386971 0.7788007830714049
6 0.707106781186547 -1.217940733217191 1.554859717121216 -1.80889999717125 0.7408182206817179
7 0.707106781186547 -1.211136595042793 1.528727005901769 -1.747832613548 42 0.7046880897187134
8 0.7071067811865476 -1.204332456868396 1.502740696425848 -1.687621730716286 0.6703200460356393
9 0.7071067811865476 -1.197528318693998 1.476900788693454 -1.628262536872655 0.6376281516217733
10 0.7071067811865476 -1.190724180519 6 1.451207282704586 -1.56975022021533 0.606530659712633 4
```

CLIM file

```
TITLE   RR test-day model for 1st lactation milk, protein and fat simulated observations

DATAFILE /homeappl/home/ejo31/MiX99_course_Italy2025/martin/MTRRpolished/lactation1sim.dat

PEDFILE /homeappl/home/ejo31/MiX99_course_Italy2025/martin/MTRRpolished/pedigree.ped

PARFILE /homeappl/home/ejo31/MiX99_course_Italy2025/Timo/Settingupandsolving/params/lact1noHTD.par

          #   1      2    3    4    5    6    7   8    9     10      11    12
INTEGER     ANI   HERD  H5Y  HPY  HYS  HTM  Y5P  YR  DIM  LCTWK  LCTMTH   MEAN

             #  1      2      3      4     5   6   7   8   9    10
REAL        cage1  cage2  milk protein  fat  L0  L1  L2  L3  W005

MISSING   -99
```

Initial values

CLIM file (cont.)

```
MODEL
milk     =  cage1 cage2 LACCRV(t1 t2 t3 t4 t5|HERD)  HTM  PE(t1 t2 t3 t5|ANI)  G(t1 t2 t3 t5 |ANI)
protein  =  cage1 cage2 LACCRV(t1 t2 t3 t4 t5|HERD)  HTM  PE(t1 t2 t3 t5|ANI)  G(t1 t2 t3 t5 |ANI)
fat      =  cage1 cage2 LACCRV(t1 t2 t3 t4 t5|HERD)  HTM  PE(t1 t2 t3 t5|ANI)  G(t1 t2 t3 t5 |ANI)

RANDOM  PE G
PEDIGREE  G  am

DATASORT  BLOCK=HERD PEDIGREECODE=ANI

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

TABLEFILE /homeappl/home/ejo31/MiX99_course_Italy2025/random_regression_FEdata/data/3rdOrd.LegPolWil005.cov
TABLEINDEX  DIM
TMPDIR  tmpMiX
```

Permanent environmental and genetic variance are 12 x 12 matrices

Parameter file

- Initial values (identity in this case)
- Lower (or upper triangle)
  - 12 by 12
- Effect 1: PE (78 rows)
- Effect 2: Genetic (78 rows)
- Effect 3: Residual (6 rows)

```
1 1 1 1.0
1 2 1 0.0
1 2 2 1.0
1 3 1 0.0
1 3 2 0.0
1 3 3 1.0
1 4 1 0.0
1 4 2 0.0
1 4 3 0.0
1 4 4 1.0
```

```
2 1 1 1.0
2 2 1 0.0
2 2 2 1.0
2 3 1 0.0
2 3 2 0.0
2 3 3 1.0
2 4 1 0.0
2 4 2 0.0
2 4 3 0.0
2 4 4 1.0
```

```
3 1 1 1.0
3 2 1 0.0
3 2 2 1.0
3 3 1 0.0
3 3 2 0.0
3 3 3 1.0
```

Solver file for VCE for example

```
# RAM: RAM demand: H=high, M=medim, L=low
   H
# STOP: Maximum_number_of_iterations, Stopping_criterion, Criterion(A/R/D)
   5000          1.0e-5                d       f
# RESID: Calculate residuals? (Y/N)
   Y
# 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 VC by EM
   E       #  f  l
   # STOPE maximum number of EM steps, samples/step, convergenc crit,
          5000                5                              1.0e-11
   # RANDG random number generator
          R
   # mix99i
/data/projects/animalgenetics/software/MiX99/develop/22.12/

# TYPSOL: Solution files? (N)o, (Y)es, (A)itken, (H)alf-Chebychev
   Y
```

REMLlog

Iteration number

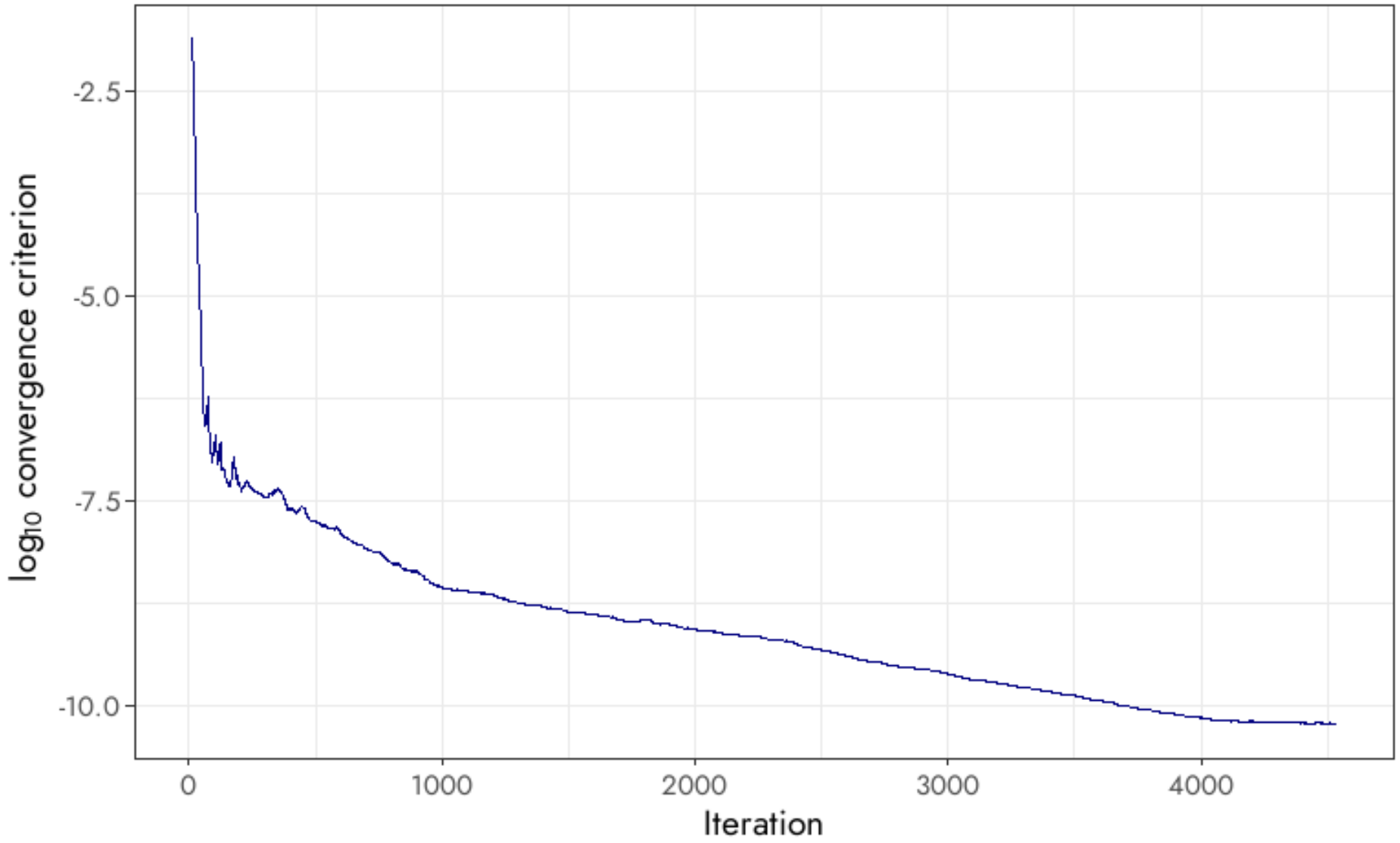
Parameter file (transpose)

0	0	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000	1.000000
0	0	1.0000000	2.0000000	2.0000000	3.0000000	3.0000000	3.000000
0	0	1.0000000	1.0000000	2.0000000	1.0000000	2.0000000	3.000000
0	0	1.0000000	0.0000000	1.0000000	0.0000000	0.0000000	1.000000
1	0.000	1.5833991	0.71297972	1.1540085	0.54328270	0.40956823	
2	0.000	2.4544583	1.4929014	1.5356451	1.1233143	0.84646708	
3	0.000	3.2635432	2.1435245	1.8898264	1.6260489	1.2031398	
4	0.000	3.6596838	2.4679343	2.0514827	1.8926559	1.3898476	
5	0.000	3.8949261	2.6666328	2.1494984	2.0493684	1.5013750	
6	0.000	4.1324323	2.8525513	2.2473872	2.1981071	1.6057587	
7	0.000	4.1669751	2.8743452	2.2228860	2.2312441	1.6197686	
8	0.000	4.2660616	2.9589724	2.2658288	2.2942561	1.6665906	
9	0.000	4.4396280	3.0921752	2.3470645	2.3946631	1.7390174	
10	0.1362E-01	4.4022409	3.0739072	2.3226811	2.3839209	1.7313119	
11	0.1223E-01	4.4976423	3.1523249	2.3705051	2.4442528	1.7762206	
12	0.1142E-01	4.5321819	3.1878526	2.3918014	2.4673401	1.7962297	
13	0.9655E-02	4.4429563	3.1332383	2.3479803	2.4207971	1.7654656	
14	0.8393E-02	4.4624780	3.1418131	2.3425195	2.4328253	1.7699962	
15	0.6910E-02	4.5433264	3.2066222	2.3869305	2.4831302	1.8097709	

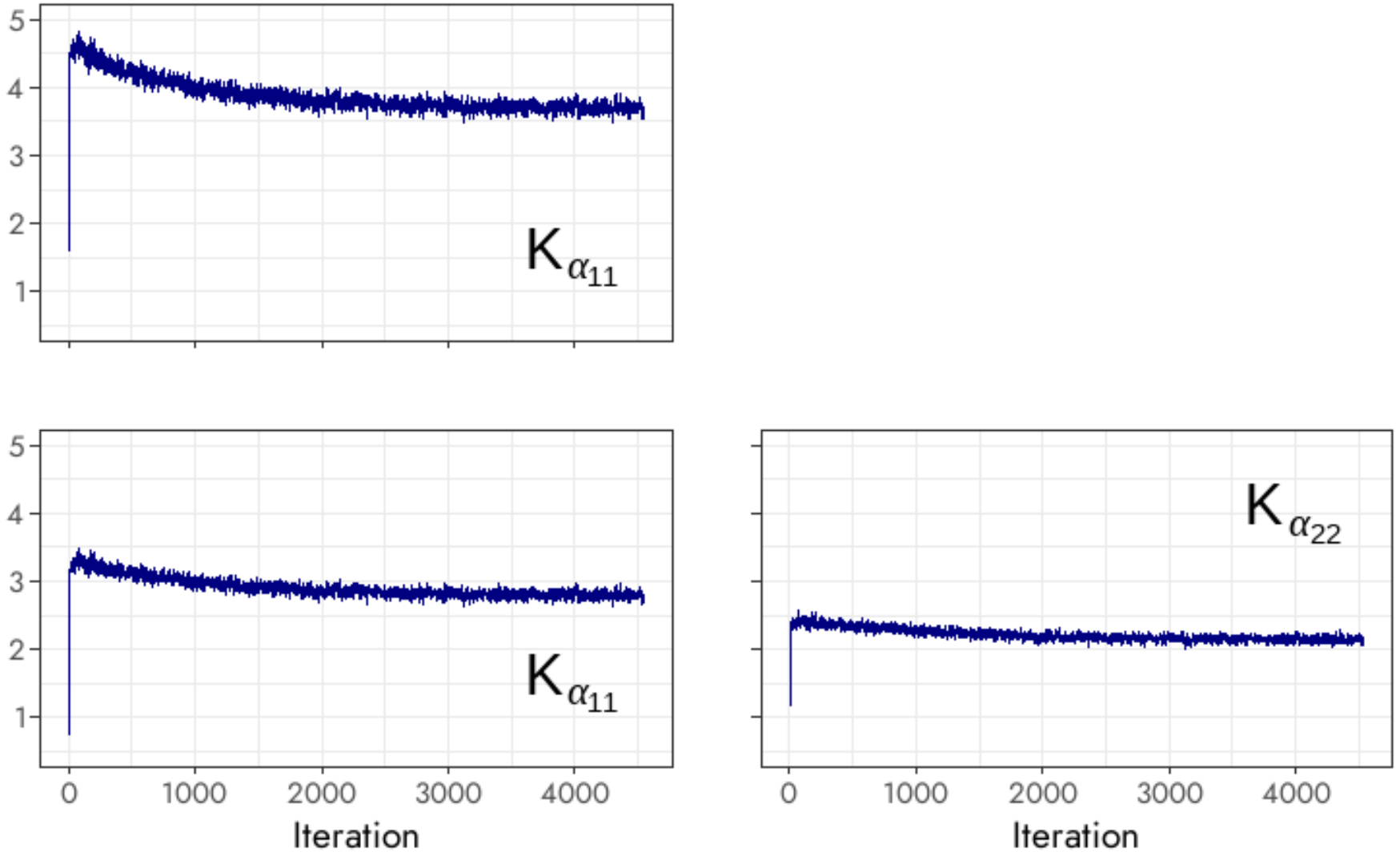
Convergence criterion

Parameter estimates (162 columns)

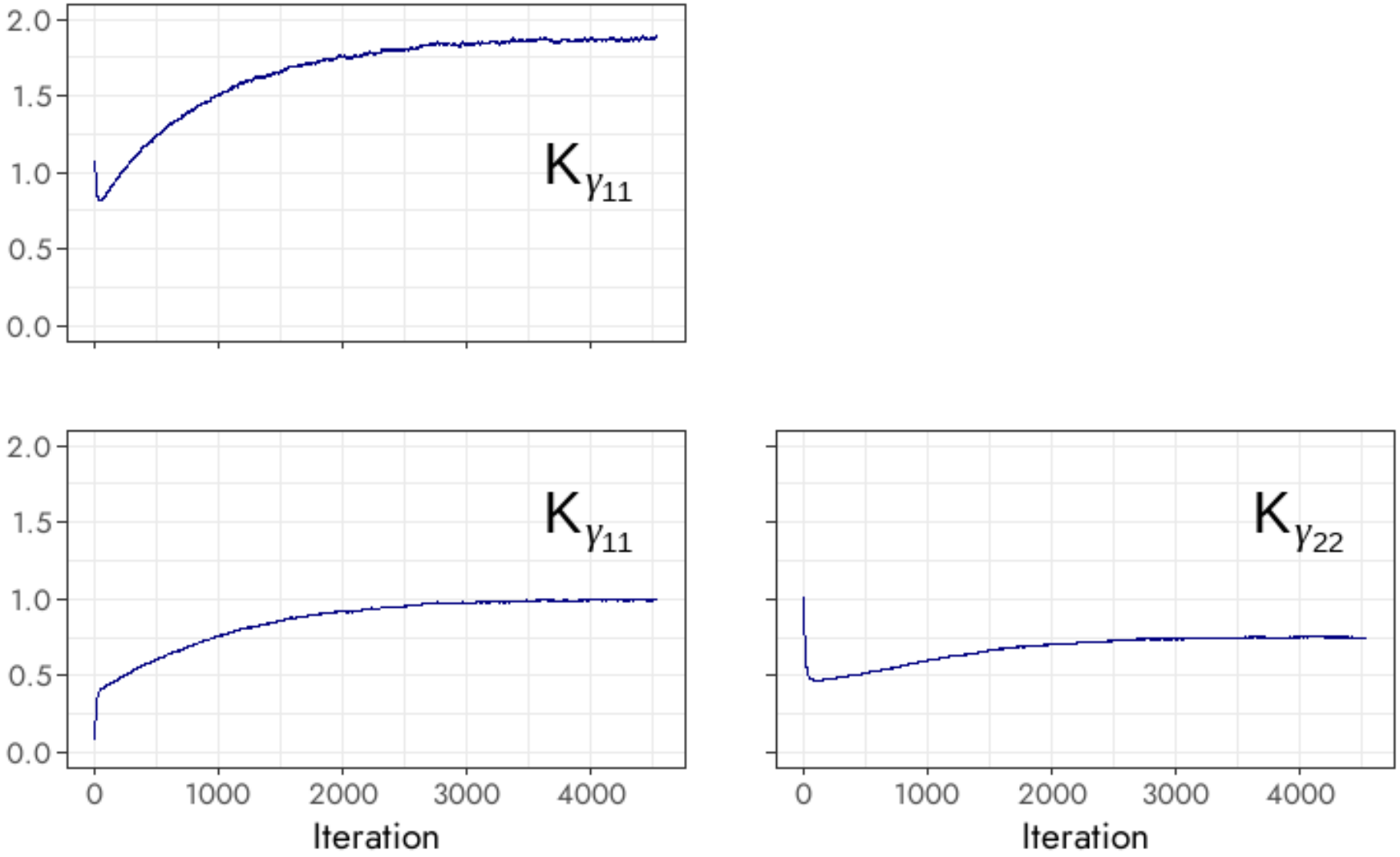
Convergence criterion



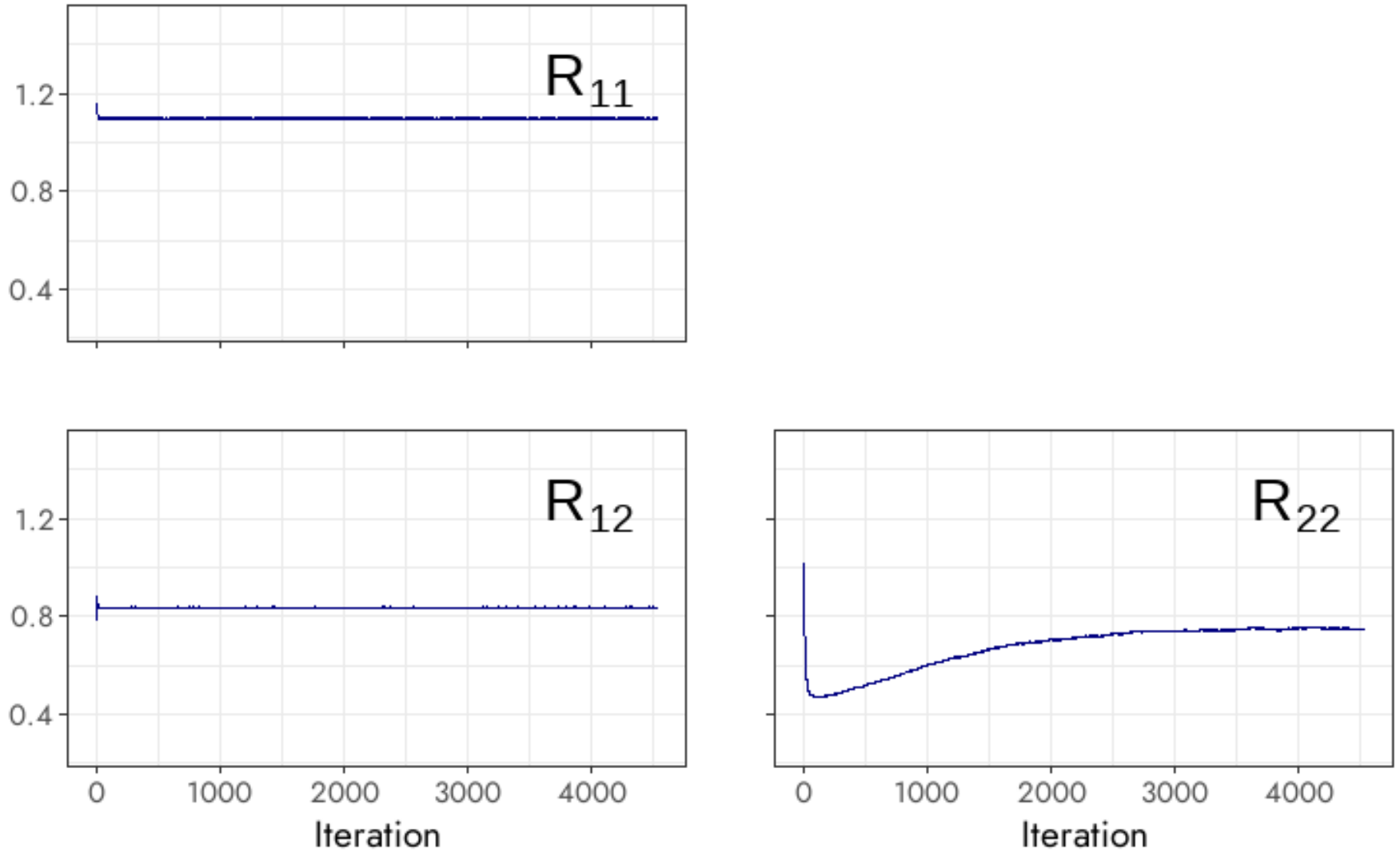
Convergence: permanent environmental variance



Convergence: genetic variance



Convergence: residual variance



Resulting estimates

- parfile

```
1    1    1   3.6934718
1    2    1   2.7956577
1    2    2   2.1477122
1    3    1   2.1081122
1    3    2   1.5975524
1    3    3   1.2794758
1    4    1  0.37596700
1    4    2  0.36171786
1    4    3  0.20203816
1    4    4   1.6504752
1    5    1  0.64850628
1    5    2  0.54702866
1    5    3  0.41669832
1    5    4   1.0773993
1    5    5  0.86017330
```

Standard errors (vceSE file)

1. Effect
2. Row
3. Column
4. SE

1	1	1	0.710401
1	2	1	0.507069
1	2	2	0.389198
1	3	1	0.433504
1	3	2	0.328822
1	3	3	0.311261
1	4	1	0.418512
1	4	2	0.307389
1	4	3	0.267630
1	4	4	0.437168

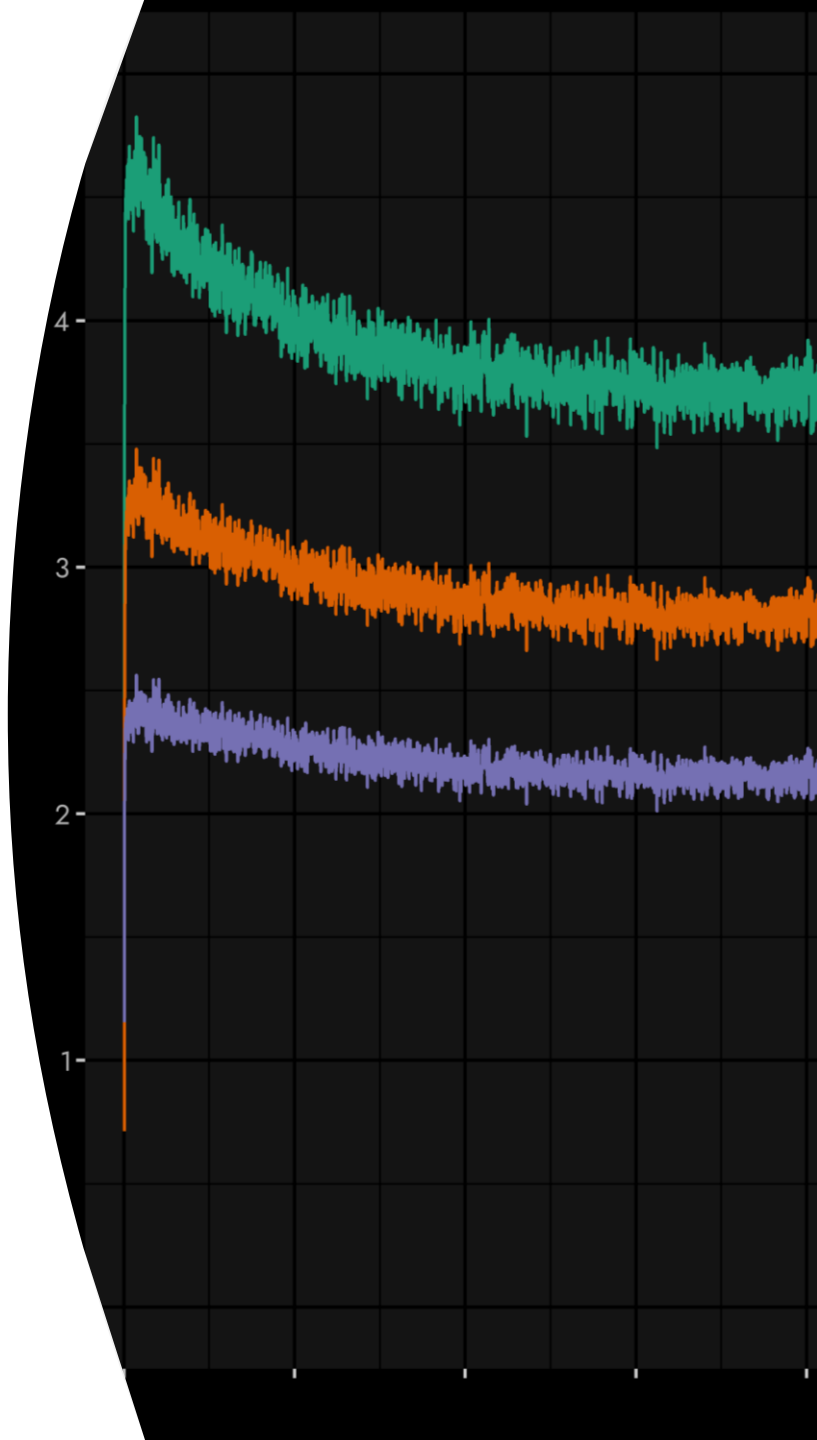
Information matrix (vcel file)

1. Effect parameter 1
2. Row parameter 1
3. Column parameter 1
4. Effect parameter 3
5. Row parameter 2
6. Column parameter 2
7. Covariance between parameter 1 and parameter 2

1	1	1	1	1	1	0.504669
1	2	1	1	1	1	0.349360
1	2	1	1	2	1	0.257119
1	2	2	1	1	1	0.245440
1	2	2	1	2	1	0.191435
1	2	2	1	2	2	0.151475
1	3	1	1	1	1	0.280528
1	3	1	1	2	1	0.202479
1	3	1	1	2	2	0.148512
1	3	1	1	3	1	0.187926

Summary

- Because of the use of MC EM REML MiX99 can estimate variance components for large RRM
  - Many number of parameters to estimate
  - Adjust your time expectations
- Not only for pedigree but also for GBLUP
- Parameters can be kept fixed (constant) at a given value
- Estimates of standard errors can also be obtained



Questions?