MiX99

Solving Large Mixed Model Equations

MiX99 features for large models

Martin Lidauer

MiX99 workshop 2014, Tuusula, Finland

3.12.2014

Which models are large?

Models for which, given a certain computing environment, solving is reaching limits

- Solving time is long or even unacceptable long
- Memory requirements are reaching limitations
- Hard disk space is limited

Large? Maybe if:

  - Number of traits >>10
  - Number of records >> 1 million
  - Number of animals > 1 million
  - Number of equations > 50 million
  - Non-linear models

MiX99's large-model features aim for

- Reducing computing time
  - Avoiding reading of redundant information
  - Optimizing iteration-on-data calculations by providing information about the structure of the data and model
  - Improving convergence
  - Parallel computing
- Reducing memory requirements
  - Options for fitting models with reduced dimension
  - Reduced memory use option for iteration-on-data

Features for large models

MiX99 is offering several opportunities to speed up solving or reduce memory requirements

- Data sorting
- Input data design
  - Grouping of traits
  - Table values for covariables
- Parallel processing
- Pre-conditioning
- Detection of convergence
- Options for fitting models with reduced dimensions

The suitability of a feature depends on the model

Data sorting

Beneficial for data with repeated observations and essential for parallel computing

- 3 sorting levels:
  - by data blocks

    herd, cohort, country, …
    - by relationship code within block

      animal ID, sire ID, progeny ID, …

      CAUTION: For models with a maternal or paternal effect and a sire-maternal-grand-sire relationship matrix the sire ID must not be used as sorting variable!

      MiX99 stores relationship information only once for all records with same relationship code!
      - by trait group within relationship code

Data sorting

- CLIM syntax

```
.
DATASORT BLOCK=Herd PEDIGREECODE=Animal
.
```

- Sorting by BLOCK or PEDIGREECODE is optional
- Sorting by trait group code is mandatory
- If BLOCK is specified, it has to be given in the pedigree and data file

Both, pedigree and data file have to be sorted by the BLOCK variable

Only data file has to be sorted by the PEDIGREECODE variable

pedigree file

1	2	3	4
Animal	Sire	Dam	Herd
1	-10	-20	100
2	-10	-20	100
3	1	2	100
4	-15	-20	100
5	1	2	102
6	4	-25	102
7	3	-25	103
8	3	7	103
.	.	.	.

data file

1	2	3	4	1	2
Animal	Herd	Year-Season	Age	Milk	Protein
5	102	3	17	5123.5	180.4
6	102	3	13	7597.0	243.8
7	103	4	25	6410.3	-9999.0
8	103	3	20	-9999.0	210.7
.	.	.	.	.	.

Input data design

For certain multiple-trait models it is possible to structure data in a way that significantly speeds up iteration-on-data

For such models MiX99 allows grouping of traits, given

- traits are measured at different time or environment
- and there exists no residual correlations between traits of different trait groups
- or the residual correlation between traits of different trait groups is modelled by a random (e.g. permanent environment) effect

Input data design

Example A: RRM with 4 traits: milk and protein in 1st and 2nd lactation

Data file
without grouping of traits

1	2	3	4	5	6	1	2	3	4	5	6	7	8
Classification variables						Covariables				Traits			
Hrd	Ani	Ag1	Ag2	Se1	Se2	C11	C12	C21	C22	M1	P1	M2	P2
.	.	.	.	.	.	.	.	.	.	.	.	.	.
34	10	7	0	5	0	.967	.042	-16.	-16.	12.1	3.40	-16.	-16.
34	10	7	0	6	0	.562	.084	-16.	-16.	8.7	3.52	-16.	-16.
34	10	0	17	0	10	-16.	-16.	.661	.035	-16.	-16.	28.2	3.37
34	10	0	17	0	10	-16.	-16.	.430	.087	-16.	-16.	32.7	-16.
.	.	.	.	.	.	.	.	.	.	.	.	.	.

data file
with grouping of traits

1	2	3	4	5	1	2	3	4
Hrd	Ani	Trg	Age	Sea	Cv1	Cv2	Mlk	Prt
.	.	.	.	.	.	.	.	.
34	10	1	7	5	.967	.042	12.1	3.40
34	10	1	7	6	.562	.084	8.7	3.52
34	10	2	17	10	.661	.035	28.2	3.37
34	10	2	17	10	.430	.087	32.7	-16.
.	.	.	.	.	.	.	.	.

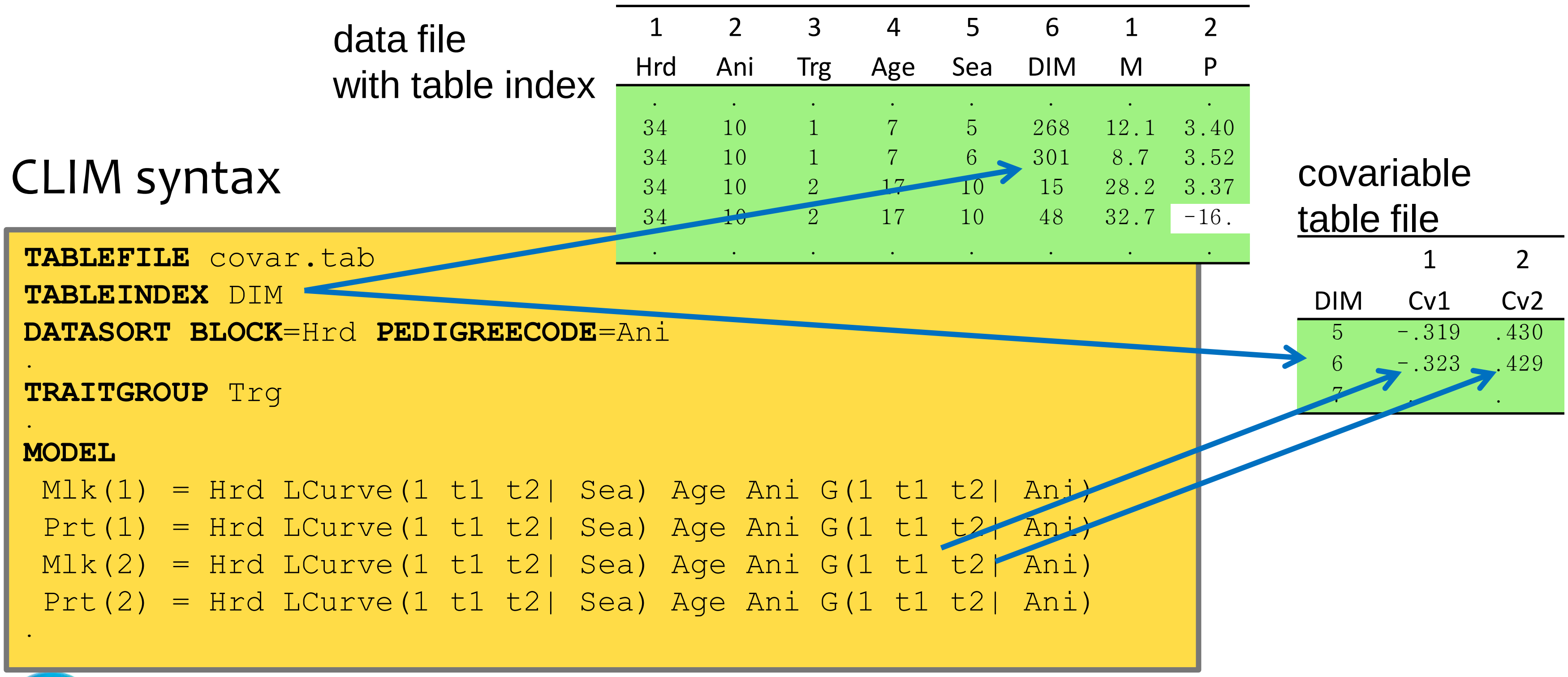
CLIM syntax

```
.
DATASORT BLOCK=Hrd PEDIGREECODE=Ani
.
TRAITGROUP Trg
.
MODEL
 Mlk(1) = Hrd LCurve(1 Cv1 Cv2| Sea) Age Ani G(1 Cv1 Cv2| Ani)
 Prt(1) = Hrd LCurve(1 Cv1 Cv2| Sea) Age Ani G(1 Cv1 Cv2| Ani)
 Mlk(2) = Hrd LCurve(1 Cv1 Cv2| Sea) Age Ani G(1 Cv1 Cv2| Ani)
 Prt(2) = Hrd LCurve(1 Cv1 Cv2| Sea) Age Ani G(1 Cv1 Cv2| Ani)
.
```

Input data design

For some models with regression functions it is possible to store covariables in a table

Example A: There are only 305 different sets of covariables



data file
with table index

1	2	3	4	5	6	1	2
Hrd	Ani	Trg	Age	Sea	DIM	M	P
.	.	.	.	.	.	.	.
34	10	1	7	5	268	12.1	3.40
34	10	1	7	6	301	8.7	3.52
34	10	2	17	10	15	28.2	3.37
34	10	2	17	10	48	32.7	-16.
.	.	.	.	.	.	.	.

CLIM syntax

```
TABLEFILE covar.tab
TABLEINDEX DIM
DATASORT BLOCK=Hrd PEDIGREECODE=Ani
.
TRAITGROUP Trg
.
MODEL
 Mlk(1) = Hrd LCurve(1 t1 t2| Sea) Age Ani G(1 t1 t2| Ani)
 Prt(1) = Hrd LCurve(1 t1 t2| Sea) Age Ani G(1 t1 t2| Ani)
 Mlk(2) = Hrd LCurve(1 t1 t2| Sea) Age Ani G(1 t1 t2| Ani)
 Prt(2) = Hrd LCurve(1 t1 t2| Sea) Age Ani G(1 t1 t2| Ani)
.
```

covariable
table file

	1	2
DIM	Cv1	Cv2
5	-.319	.430
6	-.323	.429
7	.	.

Input data design

Example B: Udder Health Evaluation Model for Nordic RDC

- MT random regression model

  3 TD-SCS traits: 1., 2. & 3. lactation

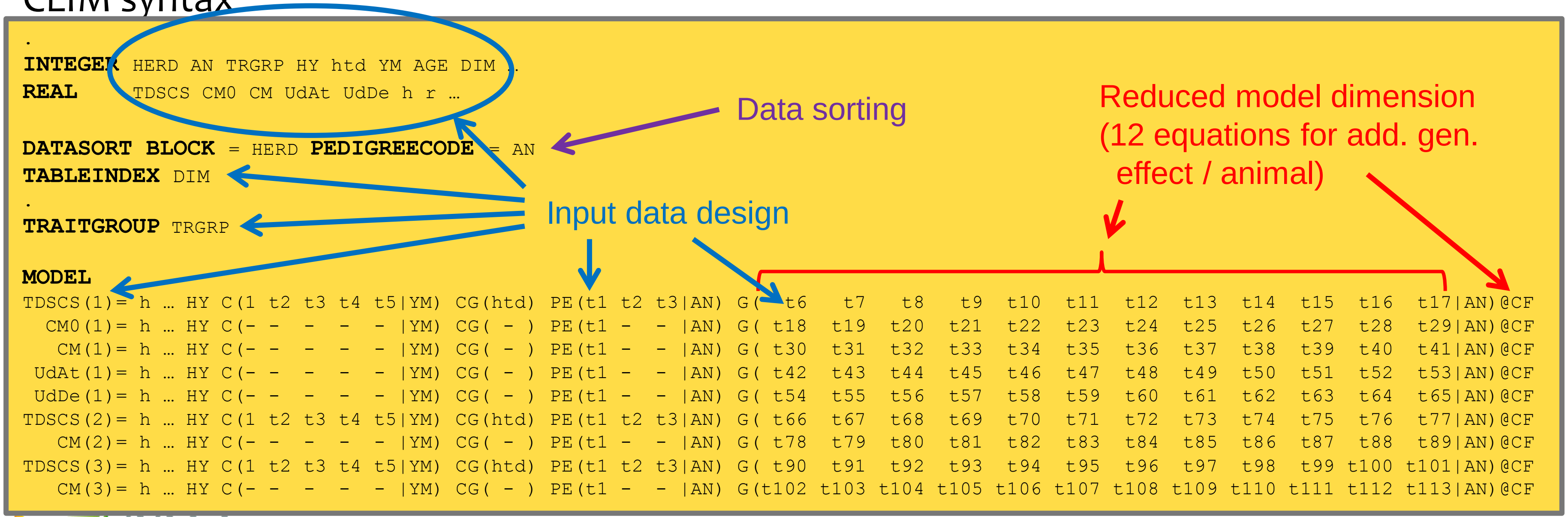
  4 clinical mastitis traits: 1. (2 traits), 2. & 3. lactation

  2 udder type traits

Combining Test Day SCS with Clinical Mastitis and Udder Type Traits: A Random Regression Model for Joint Genetic Evaluation of Udder Health in Denmark, Finland and Sweden.
Negussie et al., 20010, Interbull Bulletin 42:

- Residual correlations are modelled by VCV matrix for PE effects
- 85 million records, 5.8 million animals, 157 million equations in MME

CLIM syntax



```
.
INTEGER  HERD AN TRGRP HY htd YM AGE DIM …
REAL     TDSCS CM0 CM UdAt UdDe h r …

DATASORT BLOCK = HERD PEDIGREECODE = AN
TABLEINDEX DIM
.
TRAITGROUP TRGRP

MODEL
TDSCS(1)= h … HY C(1 t2 t3 t4 t5|YM) CG(htd) PE(t1 t2 t3|AN) G( t6   t7   t8   t9  t10  t11  t12  t13  t14  t15  t16  t17|AN)@CF
  CM0(1)= h … HY C(- -  -  -  - |YM) CG( - ) PE(t1 -  - |AN) G( t18  t19  t20  t21  t22  t23  t24  t25  t26  t27  t28  t29|AN)@CF
   CM(1)= h … HY C(- -  -  -  - |YM) CG( - ) PE(t1 -  - |AN) G( t30  t31  t32  t33  t34  t35  t36  t37  t38  t39  t40  t41|AN)@CF
 UdAt(1)= h … HY C(- -  -  -  - |YM) CG( - ) PE(t1 -  - |AN) G( t42  t43  t44  t45  t46  t47  t48  t49  t50  t51  t52  t53|AN)@CF
 UdDe(1)= h … HY C(- -  -  -  - |YM) CG( - ) PE(t1 -  - |AN) G( t54  t55  t56  t57  t58  t59  t60  t61  t62  t63  t64  t65|AN)@CF
TDSCS(2)= h … HY C(1 t2 t3 t4 t5|YM) CG(htd) PE(t1 t2 t3|AN) G( t66  t67  t68  t69  t70  t71  t72  t73  t74  t75  t76  t77|AN)@CF
   CM(2)= h … HY C(- -  -  -  - |YM) CG( - ) PE(t1 -  - |AN) G( t78  t79  t80  t81  t82  t83  t84  t85  t86  t87  t88  t89|AN)@CF
TDSCS(3)= h … HY C(1 t2 t3 t4 t5|YM) CG(htd) PE(t1 t2 t3|AN) G( t90  t91  t92  t93  t94  t95  t96  t97  t98  t99 t100 t101|AN)@CF
   CM(3)= h … HY C(- -  -  -  - |YM) CG( - ) PE(t1 -  - |AN) G(t102 t103 t104 t105 t106 t107 t108 t109 t110 t111 t112 t113|AN)@CF
```

Data sorting

Reduced model dimension (12 equations for add. gen. effect / animal)

Input data design

Parallel processing

- MiX99 has solver programs that can use several CPUs /cores at the same time
- Best speedup by means of minimizing communication
  - maximum data locality within process
  - ordering the equations in the MME to get a nearly doubly-bordered block diagonal form for the coefficient matrix
- MiX99 provides two features to meet these requirements
  - Sorting of the data by a suitable variable to get data locality (**DATASORT BLOCK**=<sorting variable>)
  - Arranging model effects within or across blocks (**WITHINBLOCKORDER** <effect names>)

Parallel processing

Example C:

milk yield is modelled by a RR model including effects for:

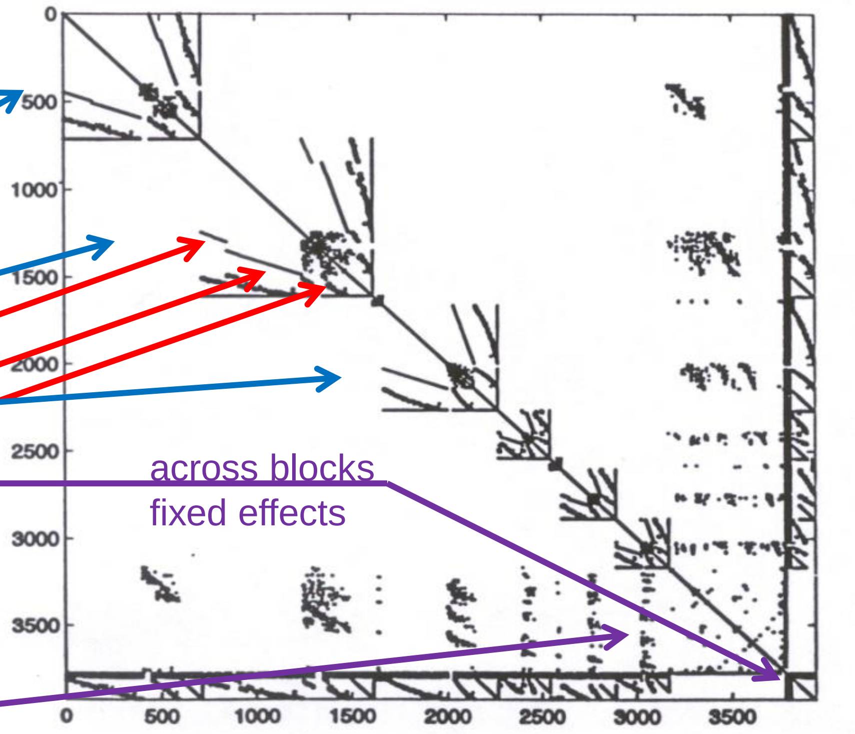
herd-test-day, age, lactation curve × year-season, PE, and animal effect

CLIM syntax

```
.
DATASORT BLOCK=HERD PEDIGREECODE=ANI

WITHINBLOCKORDER HTD PE G

MODEL
 milk = HTD AGE LC(t1 t2 t3 t4 t5| YS) &
        PE(t1 t2 t3| ANI) G(t1 t2 t3| ANI)

PARALLEL 4 1
```

4 cores   Number of common blocks

Non-zeros of coefficient matrix form doubly-bordered block diagonal matrix



Only common block equations and equations of herd-changers need to be communicated between cores

Common block equations << 10%

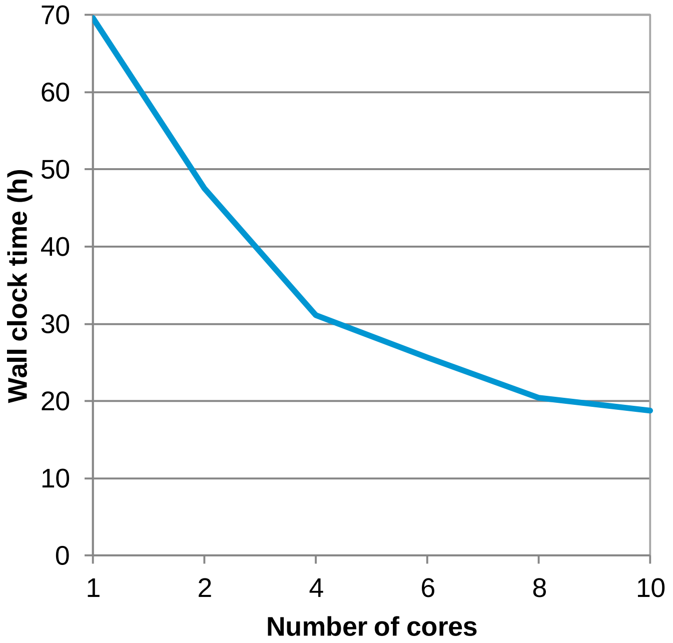
Parallel processing

Example D:

Nordic test-day yield evaluation for Holstein

- 252 million observations
- 380 million unknowns
- 2500 PCG iterations
- MiX99 parallel solver using different number of cores

Solving time (h)



Preconditioning

- PCG would converge in one step if all eigenvalues of $\mathbf{CM}^{-1}$ would be the same, i.e. the preconditioner matrix **M** would be equal to the coefficient matrix **C**
- Hence, try to find a preconditioner matrix **M** that approximates **C**
- However better preconditioner matrices often increase computations
- MiX99 allows to specify for each effect the preconditioner type
  - Diagonal
  - Block diagonal (block size equal to number of traits, or size of VCV-matrix)
  - Full block (only for across block fixed effects)
- A good choice to start with:
  - **Fixed effects**: Block diagonals
  - **Random effects**: Diagonals

Preconditioning

Example E:

- Multiple trait RR test-day model with the following effects:
  - fixed effects: herd-year, age, year-season, lactation curve × year-season
  - random effects: herd-TD, RR functions for herd-curve, pe and animal
- Preconditioner alternatives
  - A: **Diagonal** for all effects
  - B: **Block diagonal** for all effects (block sizes: fixed 9; random 9, 27, 36, 36)
  - C: **Full block** for all fixed effects, **Block diagonal** for all random effects

Preconditioner alternative	Number of Iterations	Solving Time (min)	Size of Pre-conditioner (Mb)
A: Diagonals	3725	56.3	8
B: Block diagonal	584	13.6	140
C: Block diagonal + Full block	598	24.0	250

Convergence

- By definition, given **C** is positive definite, each conjugate gradient step will yield estimates which are closer to the true solutions
- However, convergence characteristics is affected by many factors
- Overall, larger and more complex models will require more iterations to reach convergence
- Very poor convergence, or even divergence, indicates that the model is ill-conditioned and requires improvements
  - Variance components: are matrices almost singular?
  - Size of pedigree *versus* phenotype information?
  - Quality of pedigree?
  - Sparseness of observations in multiple trait models?
  - Confounding of environmental and genetic effects?
  - Error in the model input instructions?

When to stop iterations?

MiX99 reports 3 convergence indicators (norms)

- **CA:** Relative difference between left-hand and right-hand side of the additive genetic effect equations
- **CR:** Relative difference between left-hand and right-hand side of the MME
- **CD:** relative differences between solutions of consecutive iteration rounds
- $cd_{(k)} < 10^{-5}$ indicates convergence, often $cd_{(k)} < 10^{-4}$ is enough

$$ca_{(k)} = \sqrt{\left\{\frac{\left(\boldsymbol{r} - \boldsymbol{C}\hat{\boldsymbol{a}}^{(k)}\right)^{T}\left(\boldsymbol{r} - \boldsymbol{C}\hat{\boldsymbol{a}}^{(k)}\right)}{(\boldsymbol{r}_a)^{T}(\boldsymbol{r}_a)}\right\}}$$

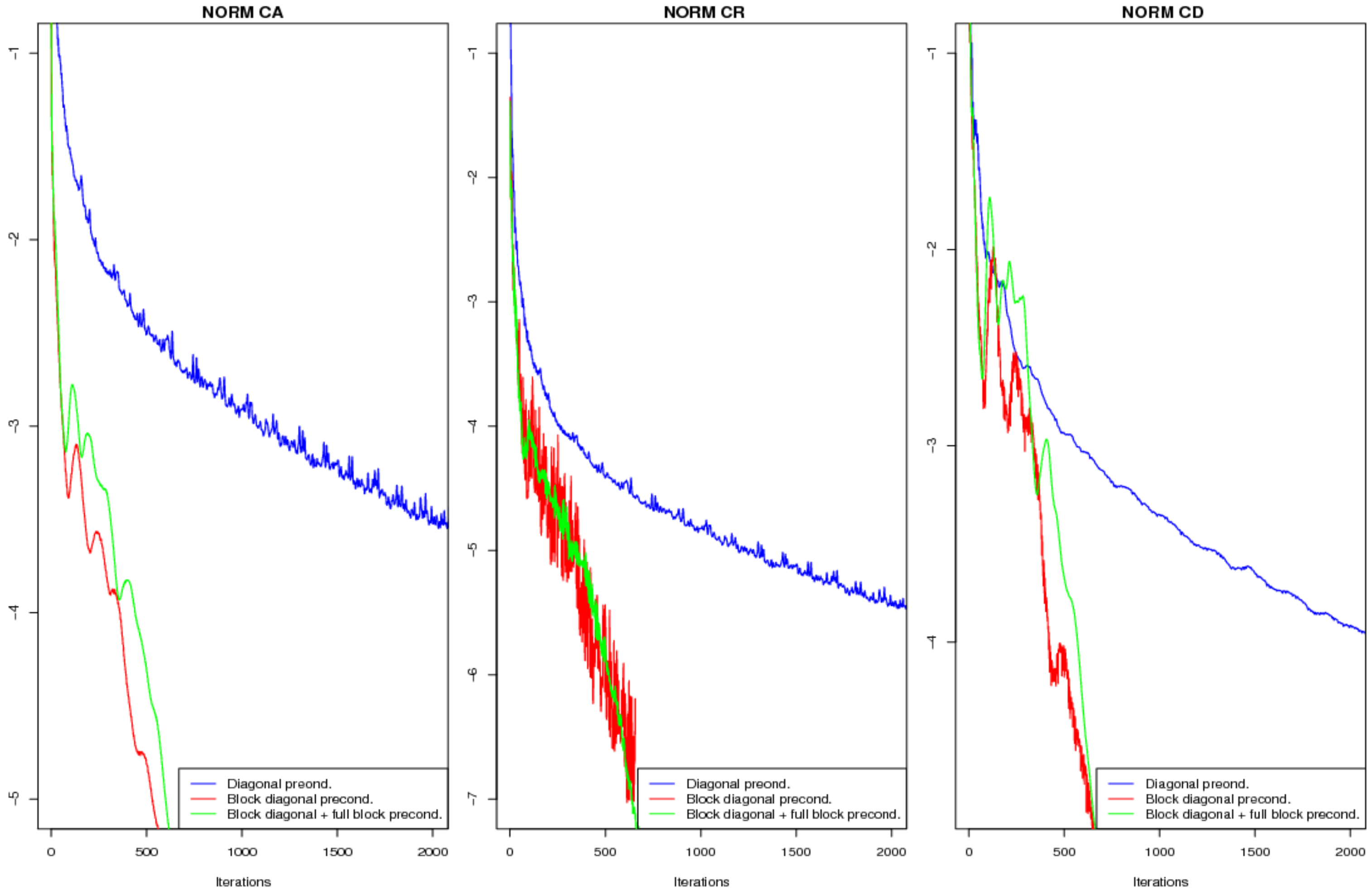
$$cr_{(k)} = \sqrt{\left\{\frac{\left(\boldsymbol{r} - \boldsymbol{C}\hat{\boldsymbol{s}}^{(k)}\right)^{T}\left(\boldsymbol{r} - \boldsymbol{C}\hat{\boldsymbol{s}}^{(k)}\right)}{(\boldsymbol{r})^{T}(\boldsymbol{r})}\right\}}$$

$$cd_{(k)} = \sqrt{\left\{\frac{\left(\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k-1)}\right)^{T}\left(\hat{\boldsymbol{s}}^{(k)} - \hat{\boldsymbol{s}}^{(k-1)}\right)}{(\hat{\boldsymbol{s}}^{(k)})^{T}(\hat{\boldsymbol{s}}^{(k)})}\right\}}$$

where **r**, ***s***, ***a*** = vector of right-hand side, solutions, add. gen. effects; **C** = coefficient matrix of MME; and *k* = iteration round

For routine evaluations, optimal stopping point depends on publishing precision of EBVs

Convergence indicators



Example E: Convergence when applying different preconditioning

Reducing model dimension

Complex multiple-trait or random regression models are often over parameterized

- Investigating the eigenvalues of applied variance components often reveals possibilities for reducing the model dimensions
- Advantage:
  - Significant reduction in memory requirements
  - Improved convergence characteristics
  - Reduced solving time

The next presentation will deal with this topic

THANK YOU