Efficient single-step BLUP computations with MiX99 software

Ismo Strandén, Matti Taskinen, Kaarina Matilainen, Martin Lidauer & Esa Mäntysaari

Genetic Research

Natural Resources Institute Finland – Luke

Ismo.Stranden@Luke.fi
mix99@luke.fi
www.luke.fi/mix99

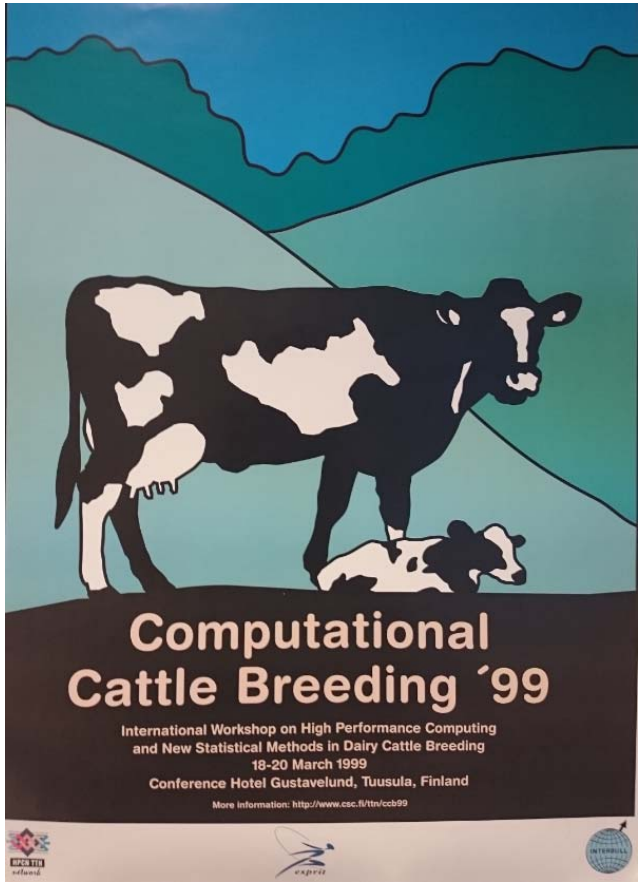
Luke
NATURAL RESOURCES INSTITUTE FINLAND

MiX99: luke.fi/mix99

- First official version in 1999
  - Breeding value estimation of large data/model
    - PCG, parallel computing using MPI
- ApaX99: Reliability approximation

- Supporting programs
  - RelaX2
    - pedigree pruning, extracting, ordering, formatting and more…
  - Hginv: genomic relationship matrix inverse
  - etc.

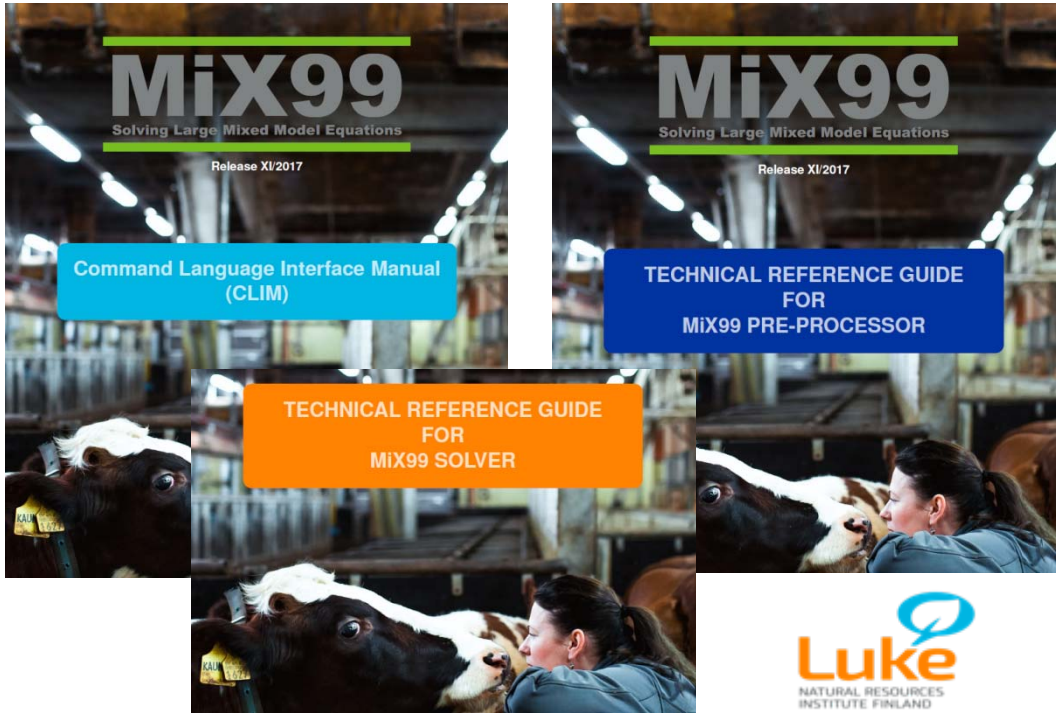
Approach and first results:



MiX99 breeding value estimation software

- Sire and animal models, repeatability model
- Multiple traits with any pattern of missing traits and different models
- Random regression models
- Reduced rank models
- Multiple residual variances
  - Heterogeneous variances by multiplicative models
- Threshold models (one categorical + several linear traits)
- Models with social effects
  - Covariance structure between an animal and its contemporary group members
- QTL effect model with external covariance matrix
- SNP-BLUP/GBLUP
- Single-step models: ssGBLUP, ssGTBLUP, APY
- Deregressed EBV, YD, DYD, …

Flexibility in models by combining model options



Luke
NATURAL RESOURCES INSTITUTE FINLAND

RelaX2

program for pedigree analysis
User's guide for version 1.65

Ismo Strandén

November 21, 2014

HGINV program
Dec 2017, version 0.85

Genomic relationship matrix builder

MiX99 program flow

Preprocessing for MiX99
- data formatting
- pedigree formatting

Model with variance components

Preprocessor: mix99i

Solver: mix99s
- mix99p for parallel

Approximate model reliabilities:
- apax99 & apax99p

Efficient ssGBLUP computations by MiX99

Single-step BLUP (ssGBLUP) allows simultaneously combining genomic information with traditional pedigree information.

Model: $\mathbf{y} = \mathbf{Xb} + \mathbf{Wa} + \mathbf{e}$ where $\mathbf{e} \sim (\mathbf{0}, \mathbf{R}\sigma_e^2)$ and $\mathbf{a} \sim (\mathbf{0}, \mathbf{H}\sigma_a^2)$

Mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \lambda\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \qquad \lambda = \frac{\sigma_e^2}{\sigma_a^2}$$

where the inverse of the relationship matrix is

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

where

- $\mathbf{A}^{-1}$ is based on the pedigree relationships (sparse & easy to compute)
- $(\mathbf{A}_{22})^{-1}$ is based on the pedigree relationships for the genotyped animals
- $\mathbf{G}^{-1}$ is based on genomic information **Z** (dense)

Computations efficiently using sparse matrices, e.g., Taskinen et al. WCGALP 2018

Increase in genotyped animals increases the dense matrix part

MiX99: hybrid parallel computing

Mixed model equations:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W} + \lambda\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \qquad \lambda = \frac{\sigma_e^2}{\sigma_a^2}$$

where the inverse of the relationship matrix is

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} \end{bmatrix}$$

Full or APY version of $\mathbf{G}^{-1}$ can be given or decomposition of **L** in $\mathbf{G}^{-1} = \mathbf{LL}'$

Sparse parts: distributed memory approach using MPI

Recent improvement in MiX99 (mix99p) is 'X'-option: faster but uses some more memory

Hybrid: MPI (distributed memory) and Cholmod library (shared memory parallel)

MPI is used when genomic matrix is not in memory,
ssGTBLUP with matrix in memory uses shared memory computing using MKL

Intel® Math Kernel Library

Reducing computations by ssGTBLUP

Assume: $\mathbf{G} = \mathbf{G_0} + \mathbf{C}$

where $\mathbf{G}_0 = \mathbf{ZZ}'$ and $\mathbf{G}_\varepsilon = \mathbf{G}_0 + \varepsilon\mathbf{I}$ ➔ $\mathbf{G}_\varepsilon^{-1} = \frac{1}{\varepsilon}\mathbf{I} - \mathbf{T}_\varepsilon'\mathbf{T}_\varepsilon$

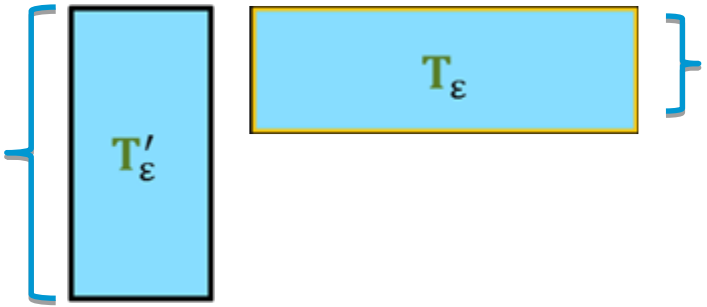
Woodbury matrix identity

where $\mathbf{T}_\varepsilon = \frac{1}{\varepsilon}\mathbf{L}_\varepsilon^{-1}\mathbf{Z}'$ and $\mathbf{L}_\varepsilon\mathbf{L}_\varepsilon' = \frac{1}{\varepsilon}\mathbf{Z}'\mathbf{Z} + \mathbf{I}$

$\mathbf{T}_\varepsilon$ has size n x m

➔ Number of computations is 2nm instead of n^2

n= number of genotyped



m= number of markers/rank

Size of $\mathbf{T}_\varepsilon$ matrix is the same as the original marker matrix.

ssGTBLUP gives the same solutions as ssGBLUP with $\mathbf{G}_\varepsilon^{-1}$ (e.g., Koivula et al. WCGALP 2018)

Example: ICBF carcass evaluations

- 9 trait multiple breed carcass evaluation
  - 9.5 million data records
- 13.35 million in pedigree
- 460,152 genotyped using 50,240 markers

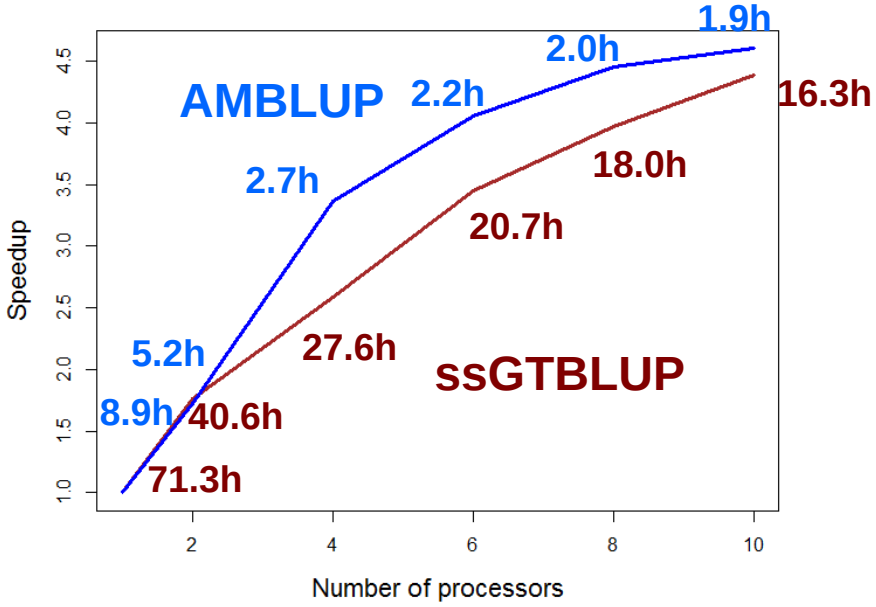
- Two evaluations:
  - Animal model BLUP → fully sparse computations using MPI
  - ssGTBLUP approach → genomic dominates 1-processor computations (88%)
    - T matrix 98% highest eigenvalues kept: T matrix rank reduced to 33,501
    - T matrix read to memory: shared memory computing (BLAS/MKL/OpenMP)

MiX99 instructions for the preprocessor

MiX99

```
DATAFILE ../../../ICBF_1step_JAN_2016/data/data_mix99lin_beef_geno.txt

INTEGER  block anim perm dam damp byr sex twin dampar scorer wtpar cullpar hyA hyB hyBp hyC hyCp hyD hyDp hyE
REAL     hrbxb hrbxd dhrbxb  dhrbxd  dfrac damage aB a2B a3B aC a2C a3C aD a2D a3D aE a2E a3E aF a2F a3F aG a2G a3G

MISSING -99

DATASORT    BLOCK=block PEDIGREECODE=anim
WITHINBLOCK anim hysccowp hysccow hycarcp hycarc hylin hypwpr hywpr hycpr hyscq damp

PEDFILE  ../../../ICBF_1step_JAN_2016/ped_mix99_beef_geno.txt
PEDIGREE anim am

INBRFILE   ../../../ICBF_1step_JAN_2016/ped_mix99_beef_geno.inbr
INBREEDING PEDIGREECODE=1 FINBR=3

RANDOM  damp hyscq hycpr hywpr hypwpr hylin hycarc hycarcp hysccow hysccowp
PARFILE ../../../ICBF_1step_JAN_2016/covar_conf9trt.txt

PARALLEL 4 21953

TMPDIR   ./tmp

MODEL
 calfq   = AA AU BA BB CH FR HE HO JE LM PI MO PT SA SH SI hrbxb hrbxd dhrbxb dhrbxd dfrac dmeal -    -    -    -    -
 cprice  = AA AU BA BB CH FR HE HO JE LM PI MO PT SA SH SI hrbxb hrbxd dhrbxb dhrbxd dfrac  -    -    -    -    -    -
 wprice  = AA AU BA BB CH FR HE HO JE LM PI MO PT SA SH SI hrbxb hrbxd dhrbxb dhrbxd dfrac  -    -    -    -    -    -
 pwpr    = AA AU BA BB CH FR HE HO JE LM PI MO PT SA SH SI hrbxb hrbxd dhrbxb dhrbxd dfrac  -    -    -    -    -    -
 musc    = AA AU BA BB CH FR HE HO JE LM PI MO PT SA SH SI hrbxb hrbxd dhrbxb dhrbxd dfrac  -   as0 as2 as3  -    -
 conf    = AA AU BA BB CH FR HE HO JE LM PI MO PT SA SH SI hrbxb hrbxd dhrbxb dhrbxd dfrac  -    -    -    - aslau a2s
 cullconf= AA AU BA BB CH FR HE HO JE LM PI MO PT SA SH SI hrbxb hrbxd dhrbxb dhrbxd dfrac  -    -    -    -    -    -
 dyd_musc= AA AU BA BB CH FR HE HO JE LM PI MO PT SA SH SI  -     -      -      -     - -   -    -    -    -    -    -
 dyd_conf= AA AU BA BB CH FR HE HO JE LM PI MO PT SA SH SI  -     -      -      -     - -   -    -    -    -    -    -
```

Speedups:
sparse matrix computations by MPI,
dense computations by parallel BLAS



Conclusions

- MiX99 is a flexible program that allows many models for breeding value estimation
- Data sets can be very large
  - → parallel computing can be used to reduce computing time
- Sparse and dense matrix computations can be done with different approaches:
  - Distributed or shared memory computing, or both for different matrix parts

- Even with modest number of genotyped animals, computations were dominated by the dense matrix computations due to genomic data

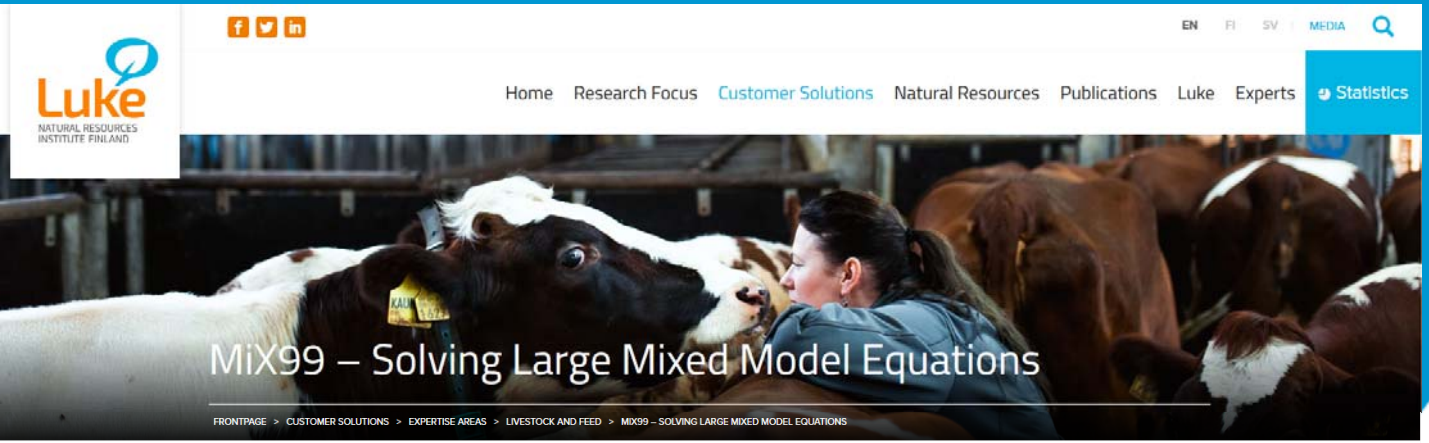
Parallel computing using dense matrix operations reduced computing time

MiX99 is kernel in MiXBLUP software (www.mixblup.eu)

Irish Cattle Breeding Federation (ICBF) is acknowledged for the data and model used in this study.

mix99@luke.fi
www.luke.fi/mix99

Thank you!



Luke
NATURAL RESOURCES INSTITUTE FINLAND