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

Mixed model equations:

$$
\begin{bmatrix}
X'R^{-1}X & X'R^{-1}W \\
W'R^{-1}X & W'R^{-1}W + \lambda H^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{b} \\
\hat{a}
\end{bmatrix}
=
\begin{bmatrix}
X'R^{-1}y \\
W'R^{-1}y
\end{bmatrix}
$$

where the inverse of the relationship matrix is

$$
H^{-1} = A^{-1} + \begin{bmatrix}
0 & 0 \\
0 & G^{-1} - A_22^{-1}
\end{bmatrix}
$$

where

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

Increase in genotyped animals increases the dense matrix part.
MiX99: hybrid parallel computing

Mixed model equations:

\[
\begin{bmatrix}
    X' R^{-1} X & X' R^{-1} W \\
    W' R^{-1} X & W' R^{-1} W + \lambda H^{-1}
\end{bmatrix}
\begin{bmatrix}
    \hat{b} \\
    \hat{a}
\end{bmatrix}
=
\begin{bmatrix}
    X' R^{-1} y \\
    W' R^{-1} y
\end{bmatrix}
\]

where the inverse of the relationship matrix is

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

\[
\lambda = \frac{\sigma^2_e}{\sigma^2_a}
\]

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

Full or APY version of \( G^{-1} \) can be given or decomposition of \( L \) in \( G^{-1} = LL' \)
Reducing computations by ssGTBLUP

Assume: \( G = G_0 + C \)

where \( G_0 = ZZ' \) and \( G_\varepsilon = G_0 + \varepsilon I \) ⇒ \( G_\varepsilon^{-1} = \frac{1}{\varepsilon} I - T_\varepsilon' T_\varepsilon \)

where \( T_\varepsilon = \frac{1}{\varepsilon} L_\varepsilon^{-1} Z' \) and \( L_\varepsilon L_\varepsilon' = \frac{1}{\varepsilon} Z'Z + I \)

\( T_\varepsilon \) has size \( n \times m \) ⇒ Number of computations is 2nm instead of \( n^2 \)

\( T_\varepsilon' \)

\( T_\varepsilon \)

\( m = \) number of markers/rank

\( n = \) number of genotyped

Size of \( T_\varepsilon \) matrix is the same as the original marker matrix.

ssGTBLUP gives the same solutions as ssGBLUP with \( G_\varepsilon^{-1} \) (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

```
DATAFILE ../.../ICBF_1step_JAN_2016/data/data_mix99in_beeb_geno.txt

INTEGER  block anim perm dam damp byr sex twin dampar scorer wtpar culpar hyA hyB hyP hyC hyD hyE hyF hyG
REAL     hrbxb hrbxd dhrbxb dhrbxd dfrac damage aB a2B a3B ac a2C a3C a2D a3D aE a2E a3E aF a2F a3F ac a2G a3G

MISSING -99

DATASORT  BLOCK=block PEDIGREECODE=anim
WITHINBLOCK anim hysccow hysccow hycarc hycarc hylin hylin hypwpr hypwpr hypwpr hyscq hyscq damp

PEDFILE ../.../ICBF_1step_JAN_2016/ped_mix99_beeb_geno.txt
PEDIGREE  anim am

INBRFILE ../.../ICBF_1step_JAN_2016/ped_mix99_beeb_geno.inbr
INBREEDING PEDIGREECODE=1 FINBR=3

RANDOM  damp hyscq hycarc hypwpr hypwpr hylin hycarc hycarc hysccow hysccow PARFILE ../.../ICBF_1step_JAN_2016/covar_conf9t1rt.txt

PARALLEL 4 21953

TMPDIR  ./tmp

MODEL
  calTf  = AA  AU  BA  BB  CH  FR  HE  HO  JE  LM  PI  MO  PT  SA  SH  SI  hrbxb  hrbxd  dhrbxb  dhrbxd  dfrac  dmeal  -  -  -  -
  cpr1ce  = 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  -  -  -  -  -  -
  dwpr  = 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  as1  as3  -  -
  conf  = AA  AU  BA  BB  CH  FR  HE  HO  JE  LM  PI  MO  PT  SA  SH  SI  hrbxb  hrbxd  dhrbxb  dhrbxd  dfrac  -  -  -  -  -  as1au  a2s
  cullconf  = AA  AU  BA  BB  CH  FR  HE  HO  JE  LM  PI  MO  PT  SA  SH  SI  hrbxb  hrbxd  dhrbxb  dhrbxd  dfrac  -  -  -  -  -  -  as1au  a2s
  dyd_musc  = AA  AU  BA  BB  CH  FR  HE  HO  JE  LM  PI  MO  PT  SA  SH  SI  hrbxb  hrbxd  dhrbxb  dhrbxd  dfrac  -  -  -  -  -
  dyd_conf  = AA  AU  BA  BB  CH  FR  HE  HO  JE  LM  PI  MO  PT  SA  SH  SI  -  -  -  -  -  -  -  -  -
```

© Natural Resources Institute Finland
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.
Thank you!

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

MiX99 – Solving Large Mixed Model Equations

Expertise areas:
- Livestock and feed
- MiX99 – Solving Large Mixed Model Equations
- Smart plant production
- Aquaculture and water economy
- Forest bioeconomy
- Food
- Biomasses and energy
- Circular economy

MiX99 is a software suite for breeding value estimation of large-scale genetic and genomic evaluations. MiX99 can be used with a wide variety of prediction models and data sets. The software is used worldwide in national and international evaluations for cattle but also for pigs, horses, sheep, goats, fish, foxes, poultry, and barley.

Some of the most important applications are genomic evaluations with massive number of genotyped animals and high heritability regression test-day models for national dairy cattle evaluations. MiX99 software is available for Linux and Windows environments, optionally utilizing parallel computing.

MiX99 software packages

To meet the different needs of worldwide research projects and industry users, MiX99 software suite is available in three different packages: MiX99, MiX99 Pro and MiX99 Enterprise.

Further information

- mix99@luke.fi
- Presentations
  - MiX99 Workshop 2014 (program)
  - MiX99 introduction
  - MiX99 overview
  - MiX99 tutorial
  - MiX99 documentation

Solving large models with