Comparison of different single-step models for (co)variance component estimation using MC AI-REML

Hongding Gao, Ismo Strandén

May 31st, 2026

@ Interbull Annual Meeting, Verona, Italy

Natural Resources Institute Finland (Luke)

Introduction

- Genomic prediction requires accurate (co)variance components estimation (VCE)
- Reliable VC are needed to estimate heritability and genetic correlations
- Ignoring genomic preselection yielded biased estimates of VC
- New traits

REML

- Restricted maximum likelihood (REML) is an important method in VCE
  - Analytical REML vs. Monte Carlo (MC) REML
  - The difference is how the quantities needed by REML are computed

Analytical REML

- Good for small and medium data
- The analytical REML-based methods usually need factorization/inversion of coefficient matrix of the MME (to form the trace terms)
- Direct inversion or factorization of a dense and large coefficient matrix is computationally challenging/infeasible

Analytical REML $\Rightarrow$ MC REML

- Approximate trace term without inverting the coefficient matrix but via MC sampling (Garcia-Cortes et al. (1992))

$$\hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\mathbf{G}^{-1}\hat{\mathbf{u}} + \mathrm{tr}(\mathbf{G}^{-1}\mathbf{C}^{uu})}{q} \quad \Rightarrow \quad \hat{\sigma}_u^2 = \frac{\hat{\mathbf{u}}'\mathbf{G}^{-1}\hat{\mathbf{u}} + \frac{1}{S}\sum_{h=1}^{S}(\tilde{\mathbf{u}}^h - \hat{\mathbf{u}}^h)'\mathbf{G}^{-1}(\tilde{\mathbf{u}}^h - \hat{\mathbf{u}}^h)}{q}$$

MC REML

- For large-scale data (1M genotyped individuals)
- For complex models (e.g. multi-trait random regression/reaction norm genomic model)

Augmented AI-REML (Strandén et al., 2024)

- The standard AI-REML needs to solve the MME for each VC
- This step can be computationally intensive for large-scale systems
- Thompson (2019) proposed an alternative approach requires solving an augmented MME only once
- It can considerably reduce the computing time within each REML iteration, particularly when using an iterative solver

The augmented MME

$$\begin{bmatrix} \mathbf{C} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{F} \\ \mathbf{F}'\mathbf{R}^{-1}\mathbf{W} & \mathbf{F}'\mathbf{R}^{-1}\mathbf{F} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{s}}^{*} \\ \boldsymbol{\Delta} \end{bmatrix} = \begin{bmatrix} \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{F}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{t} \end{bmatrix}$$

where
$\mathbf{t}$ is the trace term

$$\boldsymbol{\theta}_{new} = \boldsymbol{\theta}_{old} + \boldsymbol{\Delta}$$

Aim

- Apply Augmented MC AI-REML for VCE using different single-step models
  - ssGBLUP (Christensen et al., 2009; Aguilar et al., 2010)
  - ssGTABLUP (Vandenplas et al., 2023)
  - ssSNPBLUP (Liu et al., 2014)
- Compare the computational performance

Simulated data

- 44,280 individuals with single record
- Either 10K, 20K, or 30K youngest individuals were genotyped (9k SNP markers)
- Pedigree: 44,280

Results

Data	Model	Vg	Ve	N	Time	Peak RAM(GB)
10K	ssGBLUP	35.0±0.97	52.2±0.59	14	3m18s	2.4
	ssGTABLUP	34.1±0.95	52.6±0.59	14	12m7s	2.2
	ssSNPBLUP	34.1±0.95	52.6±0.59	18	39m54s	1.6
20K	ssGBLUP	34.7±0.93	52.9±0.52	11	7m4s	7.1
	ssGTABLUP	34.2±0.92	53.3±0.52	11	20m9s	2.9
	ssSNPBLUP	34.2±0.92	53.3±0.52	18	1h54m	2.3
30K	ssGBLUP	36.8±0.98	53.0±0.47	13	19m30s	15.5
	ssGTABLUP	36.6±0.97	53.2±0.47	12	28m59s	3.6
	ssSNPBLUP	36.6±0.97	53.2±0.47	12	1h45m	2.9

Conclusions

- ssGBLUP model was fastest but had the highest RAM usage
- ssGTABLUP and ssSNPBLUP were more memory-efficient than ssGBLUP
- ssGTABLUP and ssSNPBLUP showed better scalability than ssGBLUP as the genotyped population size increased
- ssSNPBLUP may be computationally more efficient at very large genomic scale