Hollifield, M., Strandén, I., Heuer, C., Cheng, J., Branston, K., Tseng, M., Mäntysaari, E. (2026) UD approach and its application in variance component estimation , *Proceedings 13th World Congress on Genetics Applied to Livestock Production*: 24093
https://doi.org/10.31274/wcgalp.24093 ©2026 *Hollifield, et al.* CC BY-NC-ND

UD approach and its application in variance component estimation

M.K. Hollifield[1*], I. Strandén[2], C. Heuer[1], J. Cheng[1], K. Branston[1], M. Tseng[1], and E.A. Mäntysaari[12]

[1]Bayer Crop Science, St. Louis, U.S.A; [2]Natural Resources Institute Finland, Jokioinen, Finland; [*]Corresponding author: marykate.hollifield@bayer.com

Introduction

The estimation of variance components (VCE) traditionally relies on REML or Gibbs sampling methodologies, yet both techniques face significant computational challenges in large-scale genomic models. REML often requires repeated computation of the inverse of the mixed model equations coefficient matrix (MME), which becomes computationally demanding as genomic data increases. Meanwhile, Gibbs sampling requires a large number of samples, resulting in time-intensive analyses for large genomic data. Sub-setting the dataset for faster VCE can produce biased estimates, potentially leading to misguided selection decisions (Patry and Ducrocq, 2011; Bussiman *et al*., 2023). Applying the algorithm for proven and young to create a more sparse genomic relationship matrix (**G**; Misztal *et al.*, 2014) has shown no significant improvement in VCE computing efficiency (Junqueira *et al.*, 2022).

Previous studies have demonstrated that with large models, matrix inversions can be avoided by using the full sampling Monte Carlo (MC-REML; Matilainen *et al.*, 2012). However, in models with **G**, the MC-REML requires both the inverse of **G**, and, for the generation of samples, a factor of **G**. Because the rank (r) of the centered **G** matrix is less than the number of genotyped individuals (N) or markers, inversion requires adding a rank constraint matrix to **G** before inversion. In the presented method, named the UD approach, the genomic relationship matrix is decomposed into singular vectors (**U**) and singular values (**D**) before adding the rank constraint matrix. The objective of this study was to demonstrate that with large models, inverses can be avoided by using Monte Carlo sampling, and the models with **G** can be efficiently handled by combining singular value decomposition with the proposed UD approach.

Materials & Methods

Data. A Bayer Crop Science North America soybean dataset was used with approximately 108,000 genotyped individuals and phenotypes for 6 production traits totalling approximately 3 million observations. All individuals were genotyped or imputed to approximately 50k SNPs.

Model and analyses. A general 6-trait linear model was used:

$$\mathbf{y}=\mathbf{Xb}+\mathbf{W}_1\mathbf{p}+\mathbf{W}_2\mathbf{u}+\mathbf{e}, \quad (1)$$

where **y** is a vector of phenotypes, **b** is a vector of fixed effects, **p** is a vector of random non-genetic effects, **u** is a random vector of additive genetic effects, and **e** is a vector of random residuals. Matrices **X**, $\mathbf{W}_1$, and $\mathbf{W}_2$ are incidence matrices relating elements of **y** to **b**, **p**, and **u**, respectively. All 6 traits shared the same model. The random effects were distributed as: $\mathbf{p} \sim N(\mathbf{0}, \mathbf{I} \otimes \mathbf{P}_0)$, $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{G}_0)$, and $\mathbf{e} \sim N(\mathbf{0}, \sum \oplus \mathbf{R}_i)$, where **G** is the genomic relationship matrix as defined by VanRaden (2008), $\mathbf{P}_0$ and $\mathbf{G}_0$ are the 6×6 (co)variance matrices of non-genetic and genetic effects and $\mathbf{R}_i$ is the residual (co)variance matrix referring to i^{th} observation. An iterative algorithm, such as the preconditioned conjugate gradient method, for solving model effects can be based either on mixed model equations (Henderson., 1973), or on iteration

on data (Schaeffer and Kennedy, 1986). In both approaches, the iteration involves repeated multiplication of a search vector by the inverse of the **G** matrix. In standard REML variance component estimation, the prediction error variances of random effects are obtained from inverse of MME, while in MC-REML the prediction error variances (PEV) are obtained from repeated solution of the model with sampled observation vectors. These can be produced by first generating $\boldsymbol{p} \sim N(\mathbf{0}, \mathbf{I} \otimes \mathbf{P}_0)$, $\boldsymbol{u} \sim N(\mathbf{0}, \mathbf{G} \otimes \mathbf{G}_0)$, and $\boldsymbol{e} \sim N(\mathbf{0}, \sum \oplus \mathbf{R}_i)$, which then are summed according to Eq.1 to form $\boldsymbol{y}$.

Since the rank (r) of the **G** matrix is often less than the number of genotyped individuals N, a rank constraint matrix has to be added to **G** before inversion, *i.e.*, $\mathbf{G} = \mathbf{ZZ}'+\mathbf{I}c$, where **Z** is the matrix of centered and scaled genotypes and c is a small constraint value. With a singular value decomposition of **Z**:

$$\mathbf{G} = \mathbf{UDDU}'+\mathbf{I}c, \quad (2)$$

where **U** is a matrix of r singular vectors, and **D** is a diagonal matrix of r non-zero singular values of **Z**. Note that for models with large number of genotyped individuals the **U** matrix can be compute as $\mathbf{U}=\mathbf{ZVD}^{-1}$, in where the **V** and **D** matrices are obtained from an eigenvalue (or singular value) decomposition $\mathbf{Z}'\mathbf{Z}=\mathbf{VDDV}'$.

Using the Woodbury matrix equality:

$$\mathbf{G}^{-1} = \mathbf{I}c^{-1}- \mathbf{U}(\mathbf{U}'\mathbf{U}c+ c^2 (\mathbf{DD})^{-1})^{-1}\mathbf{U}' \quad (3)$$

(Mäntysaari *et al.,* 2017). Because the columns in **U** are orthogonal and **D** is diagonal, the $\mathbf{G}^{-1}$ matrix can be stored efficiently with only r columns of **U** and a vector of the diagonal elements of **DD** (squared singular values). Using Eq.3, the multiplication of the search vector in the iterative solver can be done matrix-by-matrix without forming the $\mathbf{G}^{-1}$ matrix. The simulated breeding values $\boldsymbol{u}$ needed for MC-REML samples are easily generated using the **UD** factor and **I**c independently.

In Eq.2, the smallest singular values can be ignored to improve the condition number of **G**. We used r equal to 19,800, which accounted for 99.9% of the total sum of the eigenvalues of the **G** matrix, and c equal to 0.01 to apply the UD approach. The VCE results were compared with those from REML with the whole $\mathbf{G}^{-1}$ matrix of size 108,000, including genic correlation, which is defined as the correlation between variances attributed to the genetic effect.

Results

Genic correlation estimates between trait 1 and the remaining five traits showed negligible differences between REML with the UD approach and with the whole $\mathbf{G}^{-1}$ matrix, as the standard errors overlapped for all trait combinations (Table 1). The correlation of breeding values, calculated using variance components obtained from both methods, exceeded 0.999 for all traits. Similarly, the variances of the genomic estimated breeding values (GEBV) showed minimal differences between the two methods. For the 6-trait model analysis, the UD approach reduced computing time from nearly 10 days to less than 1.5 days and decreased average disk usage from 121.1 GB to 36.4 GB. REML with the whole $\mathbf{G}^{-1}$ matrix took 306 rounds to reach convergence while the UD approach needed 182 rounds.

Table 1. Variances of GEBV calculated using REML with the whole $\mathbf{G}^{-1}$ and UD approach.

	Trait1	Trait2	Trait3	Trait4	Trait5	Trait6
Whole G	5.42	0.83	0.14	2.32	0.23	0.25
UD	5.42	0.84	0.14	2.39	0.24	0.26

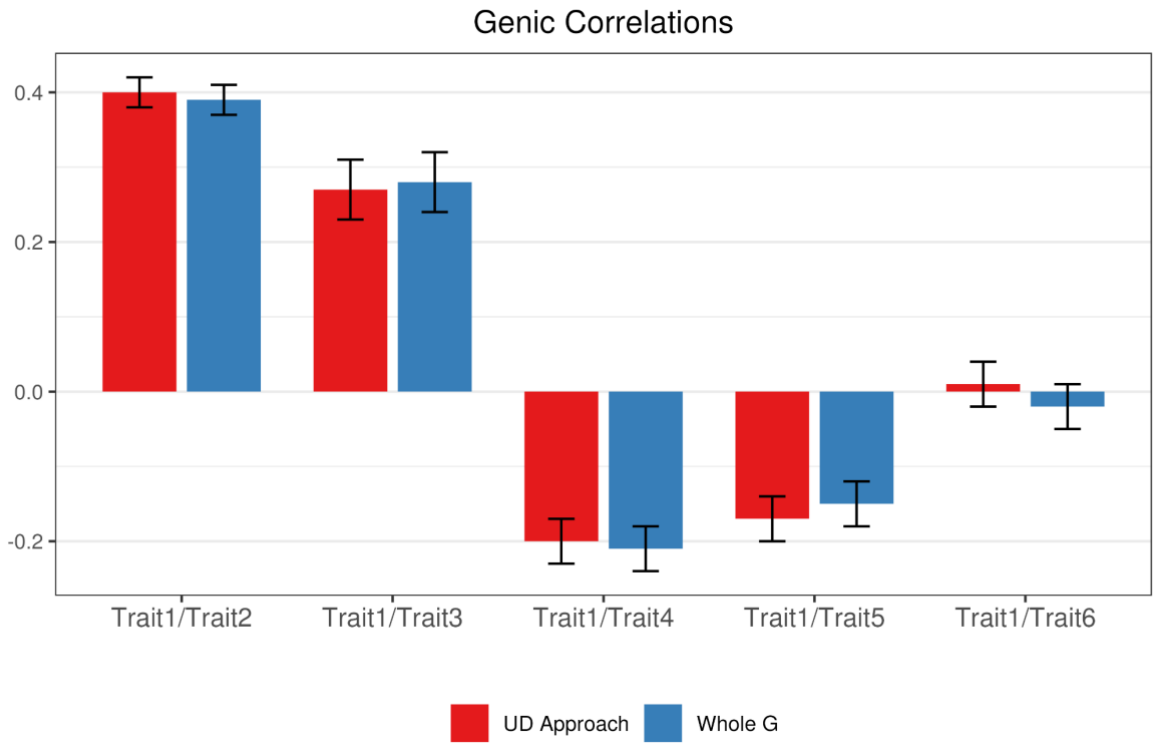

Figure 1. Genic correlations obtained using REML with the whole $\mathbf{G}^{-1}$ and UD approach.

Discussion

In VCE, each REML round requires an estimation of breeding values. In large GBLUP evaluations with many genotyped individuals, the computing time is almost directly proportional to the number of elements in the $\mathbf{G}^{-1}$ matrix. However, with the UD approach, the computations involve dense matrix $\mathbf{U}$ of size N × r. When N is two times r, GBLUP exploiting the UD approach and GBLUP with the whole $\mathbf{G}^{-1}$ matrix require similar computing time. However, when N grows relative to r, the number of computations in the UD approach grows linearly with N, while standard GBLUP computations increase quadratically. The same holds for preprocessing steps. When N grows, the computations to solve $\mathbf{G}^{-1}$ grow cubically, but the computations to solve $\mathbf{U}$ increase linearly on N.

In typical REML implementations, the advantage of using the UD approach will vanish when MME are inverted. In MC-REML, the inverse is not needed because PEV of random effects are estimated using repeated sampling. With s samples, each REML round requires s+1 model solves. Consequently, as in GBLUP, the time needed for each REML round becomes linearly proportional to the number of genotyped individuals. In our experience, the MC-REML utilizing the UD approach converged faster than the one with the full $\mathbf{G}^{-1}$ matrix, which might be because of removal of error noise from the $\mathbf{G}$ matrix. This needs to be verified in future studies.

Conclusions

The UD approach with r accounting 99.9% of the total sum of the eigenvalues offers significant computational advantages for large-scale VCE in animal and plant breeding. This method maintains accuracy, as shown by negligible differences in genic correlations, GEBV correlations, and GEBV variances compared to using the whole $\mathbf{G}^{-1}$ matrix.

References

Bussiman F., Chen C.-Y., Holl J., Bermann M., Legarra A., *et al*. (2023) Journal of Animal Science 101. doi.org/10.1093/jas/skad273

Henderson C. R., (1973) Journal of Animal Science 10-41. https://doi.org/10.1093/ansci/1973.Symposium.10

Junqueira V. S., Lourenco D., Masuda Y., Cardoso F. F., Lopes P. S., *et al.* (2022) Journal of Animal Science 100. doi.org/10.1093/jas/skac082

Mäntysaari E.A., Evans R. D., Strandén I. (2017) Journal of Animal Science 11:4728-4737. doi.org/10.2527/jas2017.1912

Matilainen K., Mäntysaari E.A., Lidauer M., Strandén I., Thompson R. (2012) Journal of Animal Breeding and Genetics 129:457-468. doi.org/10.1111/j.1439-0388.2012.01000.x

Misztal I., Legarra A., Aguilar I. (2014) Journal of Dairy Science 97:3943-3952. doi.org/10.3168/jds.2013-7752

Patry C., and Ducrocq V. (2011) Journal of Dairy Science 94:1011-1020. doi.org/10.3168/jds.2010-3804

Schaeffer, L. R., and Kennedy B. W. (1986) Journal of Dairy Science 69.2: 575-579. doi.org/10.3168/jds.S0022-0302(86)80441-6

VanRaden P. M. (2008) Journal of Dairy Science 91:4414-4423. doi.org/10.3168/jds.2007-0980