Variance component estimation for single-step genomic BLUP for Australian terminal sire sheep

P.M. Gurman[1*], L. Li[1], A.A. Swan[1], N. Moghaddar[2] and J.H.J. van der Werf[2]

[1]AGBU, a joint venture of NSW Department of Primary Industries and University of New England, 2351, Armidale, Australia; [2]University of New England, NSW 2351, Australia; [*]pgurman@une.edu.au

Abstract

Many national genetic evaluation systems have transitioned to single-step genomic BLUP (SS-GBLUP) for breeding value estimation utilising variance components estimated from pedigree-based REML. The genomic and numerator relationship matrices are often weighted in SS-GBLUP by a parameter λ, that minimises bias and increases accuracies and is set to 0.5 in Australian sheep analyses. This study estimates variance components from genomic relationship matrices weighted by λ ranging from 0 to 1 by genomic REML. The impact of using these variances for each λ was then examined in SS-GBLUP models via cross-validation metrics. Data on terminal sire sheep in Australia were analysed using univariate models for carcase and live weight traits. We found that the maximum log-likelihood was estimated at lambda values between 0.5 and 0.875, while cross-validation results suggest that increasing accuracies can be achieved with increasing λ towards one and no significant change to bias using estimated variance components.

Introduction

National genetic evaluations have largely transitioned to single-step genomic BLUP where sufficient genotypes are available. This transition occurred for Australian sheep in 2017 (Brown *et al.*, 2018). Variance components for these analyses have often reused the same components that were derived from, and used in pedigree-based BLUP analyses. Further, it was identified that blending the genomic relationship matrix ($\boldsymbol{G}$) with part of the numerator relationship matrix ($\boldsymbol{A}_{22}$ for genotyped animals) ($\lambda G + (1-\lambda)A_{22}$) helped to increase accuracy and reduce the bias that was observed when including genomic data (McMillan & Swan, 2017; Zhang *et al.*, 2017), either correctly weighting the genomic and pedigree information, or underutilising the genomic information. Recent studies have investigated using either sampling-based methods or genomic REML (GREML) based methods to fit two effects simultaneously to examine how the variance partitions between the two effects (with the ratio between the two effects being λ), a pedigree-based genetic effect using the numerator relationship matrix and a genomic-based genetic effect using the single-step relationship matrix $\boldsymbol{H}$ (Samaraweera *et al.*, 2021; Torres-Vázquez *et al.*, 2021). This study investigates the impact of $\boldsymbol{G}$ constructed for a range of λ values on variance component estimates, log-likelihood values, and SS-GBLUP cross-validation metrics to determine what is the optimal value of λ for use in SS-GBLUP analyses.

Materials & Methods

Data for 9,688 terminal sire sheep were extracted from the Sheep Genetics LAMBPLAN terminal sire analysis, along with pedigree information on their ancestors. These animals were selected for simultaneous phenotypic recording for carcase and live weight traits including intramuscular fat (IMF, %), shear force (SF5, newtons), carcase eye muscle depth (CEMD, mm), carcase c-site fat (CCFAT, mm), carcase weight (CWT, kg) and post-weaning weight (PWT, kg). These animals are mostly from the Sheep Cooperative Research Centre Information

Nucleus Flock (van der Werf *et al.*, 2010) and the Meat and Livestock Australia Resource Flock populations which form the genomic reference population for Australian sheep, so all animals were genotyped and were recorded for all phenotypes Phenotypes were pre-adjusted as part of routine evaluation for combinations of birth type, rearing type, age of measurement, age of dam, and hot carcase weight, depending on the trait. Contemporary groups for IMF, SF5, CEMD, CCFAT, and CWT were based on kill group, while PWT was based on breed, flock, management group and sex. A breed-adjusted genomic relationship matrix was constructed (Makgahlela *et al.*, 2013), which utilised multiple sets of allele frequencies to reduce breed level misalignment between pedigree and genomic relationship matrices.

Univariate variance component estimation was performed using MTG2 (Lee & van der Werf, 2016). The model used was $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{ZQg} + \boldsymbol{Zm} + \boldsymbol{\epsilon}$ where $\boldsymbol{y}$ is the vector of pre-adjusted phenotypes for fixed effects; $\boldsymbol{X}$ is the design matrix for the fixed effects (contemporary groups); $\boldsymbol{\beta}$ is the vector of contemporary group solutions, $\boldsymbol{Z}$ is the design matrix for the random effects associated with breeding values for individual animals (animals with records by breeding values), which in this case is $\boldsymbol{I}$; $\boldsymbol{u}$ is the vector of random additive genetic effects with $N(\boldsymbol{0}, \boldsymbol{G}_w\sigma_a^2)$; $\boldsymbol{G}_w$ is the weighted $\boldsymbol{G}$ $\boldsymbol{G}_w = \lambda\boldsymbol{G} + (1-\lambda)\boldsymbol{A}_{22}$; $\boldsymbol{Q}$ is the matrix of genetic group proportions, defined by the breed of origin from the pedigree; $\boldsymbol{g}$ is the vector of random genetic group effects with $N(\boldsymbol{0}, \boldsymbol{QQ}'\sigma_g^2)$; $\boldsymbol{M}$ is the design matrix for the random maternal permanent environment effects; $\boldsymbol{m}$ is a vector of random maternal effects only fitted for two weight traits (CWT and PWT) with $N(\boldsymbol{0}, \boldsymbol{MM}'\sigma_m^2)$ and $\boldsymbol{\epsilon}$ is the vector of residuals effects with $N(\boldsymbol{0}, \boldsymbol{I}\sigma_e^2)$. All random effects were fitted as an animals by animals matrix, to accommodate MTG2. For GREML, the value of λ was varied between 0 and 1 in steps of 0.05.

The single-step cross-validation was performed as a 5-fold analysis. The 9,688 animals were randomly allocated into one of the five folds, stratified by breed and sire family. This process was repeated five times resulting in five replicates of five-folds and 25 SS-GBLUP analyses, where the phenotypes for the validation animals were removed from the analysis and EBVs calculated. These cross-validation analyses followed the same model as the GREML analyses, except that the inverse of the appropriate joint relationship matrix, $\boldsymbol{H}^{-1}$, was used instead. Five cross-validation metrics are presented, for which we define the following for the validation animals: $\boldsymbol{y}^*$ as the phenotypes adjusted for contemporary group solutions from the model with appropriate λ; $\widehat{\boldsymbol{u}}_p$ as the EBVs from the partial analyses (data removed for the validation animals); $\widehat{\boldsymbol{u}}_w$ as the EBVs from the whole analyses (data retained for the validation animals), $\boldsymbol{K}$ as the subblock of $\boldsymbol{H}$ for the validation animals, and $\sigma_{a,\infty}^2$ as the genetic variance at equilibrium in a population under selection, assumed for simplification to be σ_a^2 and h is the square root of the heritability. Traditional accuracies were calculated as $cor(\boldsymbol{y}^*, \widehat{\boldsymbol{u}}_p)/h$ and the dispersion as the regression of phenotype on EBV. Metrics based on Legarra & Reverter (2018) using the method LR (from "linear regression") were also calculated: LR accuracies as $\sqrt{\frac{cov(\widehat{\boldsymbol{u}}_p, \widehat{\boldsymbol{u}}_w)}{(\overline{d\iota ag(\boldsymbol{K})} - \overline{\boldsymbol{K}}) \times \sigma_g^2}}$; LR bias as $(\overline{\widehat{\boldsymbol{u}}_p} - \overline{\widehat{\boldsymbol{u}}_w})/\sigma_a$ and LR dispersion as $cov(\widehat{\boldsymbol{u}}_p, \widehat{\boldsymbol{u}}_w)/var(\widehat{\boldsymbol{u}}_p)$.

Results

The variance components estimated for each value of λ are presented in Figure 1, along with a vertical dotted line at the value of λ where the maximum log-likelihood value was observed, between 0.5 and 0.85 across the six traits. The genetic variance increased from a pedigree only model (i.e. λ = 0), reaching a peak around λ = 0.25 and decreasing as λ continued to increase. The residual variance followed an inverse pattern to the genetic variance. For most traits, the maximum log-likelihood was observed at the λ value where the genetic variance was similar to that observed at λ = 0. It should also be noted that the genetic group variances were not, as is often assumed, equal to the genetic variances. For IMF, SF5 and CEMD, the ratio of genetic group variance to the genetic variance was between 0.06 and 0.6, while for the two weight traits (CWT and PWT) the ratio was between 1.97 and 4.4. The only trait where the ratio crossed one over the range of λ was CCFAT, which was between 0.76 and 1.52.

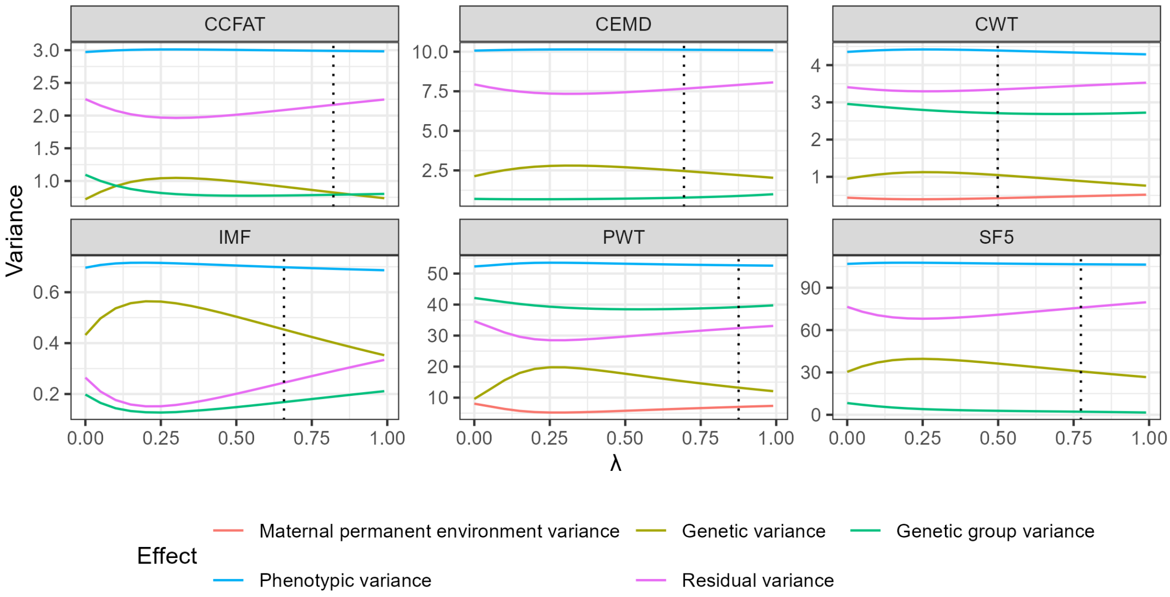

Figure 1. Variance components estimated for each value of λ. Vertical dotted lines indicate the λ value where the optimal log-likelihood value was estimated. Trait abbreviations are defined in the Materials and Methods.

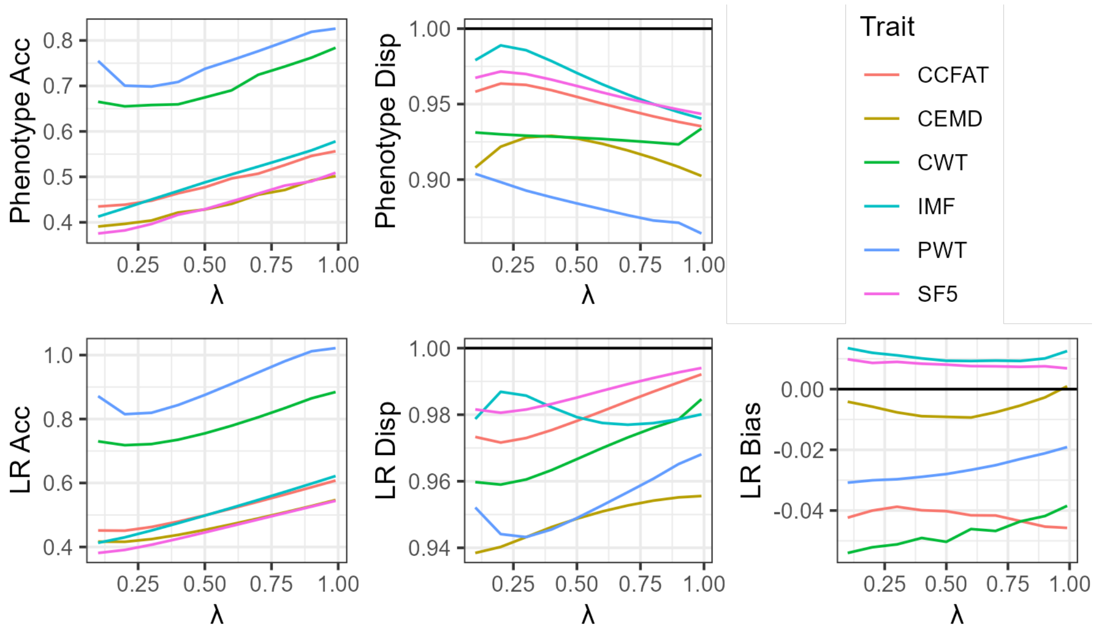

Figure 2. Cross-validation metrics calculated as the mean of the metric within each cross-validation group. Trait abbreviations are defined in the Materials and Methods.

For the cross-validation analyses (Figure 2), both the phenotypic and LR accuracies show a mostly linear increase as the value of λ increases, although for λ values < 0.3 for some traits a slight increase in accuracy is observed. For most traits except CWT, slight increases in traditional dispersion were observed as λ increased. This was not observed in the LR dispersions, with these dispersions being consistent over the range of λ. Finally, the LR biases were largest for CWT, CCFAT and PWT, though the latter decreased slightly at higher values of λ. No significant patterns concerning λ were observed.

Discussion

Of note here is that the log-likelihood results and the cross-validation results provide different indications to the optimal model, due in part to the decreasing genetic variances as λ approaches one. Methods for accommodating individual λ values per trait optimally have been developed by Meyer in a paper submitted to this congress. It is unclear why the highest genetic variances were estimated around $\lambda = 0.25$ which warrants further investigation. A more diverse set of traits would need to be studied to determine if the estimated heritabilities always decrease with high λ. These results would also need to be verified in models that utilise other methods for aligning $\boldsymbol{A}$ and $\boldsymbol{G}$, e.g. metafounders and other scaling parameters (Christensen, 2012; Legarra *et al.*, 2015) and in other datasets, including for different traits and different genomic structures. The genetic group variances estimated here are of note, suggesting that where genetic groups are fitted as random, variances should be estimated for this effect and used in SS-GBLUP.

This paper finds that while log-likelihood values presented here suggest that individual λ values per trait are optimal, cross-validation suggests that using only genomic information is optimal if variance components are used that were estimated for the $\boldsymbol{G}$ used in SS-GBLUP. Further work is warranted to validate these results in other datasets.

Acknowledgements

Work was funded by Meat and Livestock Australia project L.GEN.1815.

References

Brown, D.J., Swan, A.A., Boerner, V., Li, L., Gurman, P.M., *et al.* (2018) Proc. of the 11th WCGALP, Auckland, New Zealand.

Christensen, O.F. (2012) Genet. Sel. Evol., 44:37. https://doi.org/10.1016/S0167-5877(00)00182-3.

Lee, S.H. and van der Werf, J.H.J. (2016) Bioinformatics, 32(9):1420–1422. https://doi.org/10.1093/bioinformatics/btw012

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I. and Misztal, I. (2015) Genetics, 200(2): 455–468. https://doi.org/10.1534/genetics.115.177014

Legarra, A. and Reverter, A. (2018) Genet. Sel. Evol., 50(1):53. https://doi.org/10.1186/s12711-018-0426-6

Makgahlela, M.L., Strandén, I., Nielsen, U.S., Sillanpää, M.J. and Mäntysaari, E.A. (2013). J. Dairy Sci., 96(8):5364–5375. https://doi.org/10.3168/jds.2012-6523

McMillan, A.J., and Swan, A.A. (2017) Proc of the 22nd AAABG, Townsville, Australia.

Samaraweera, A.M., Torres-Vázquez J.A., Jeyaruban M.G., Johnston, D.J. and Boerner, V. (2021) Proc of the 24th AAABG, Adelaide, Australia.

Torres-Vázquez, J.A., Samaraweera, A.M., Jeyaruban, M.G., Johnston, D.J., and Boerner, V. (2021) Proc of the 24th AAABG, Adelaide, Australia.

van der Werf, J.H.J., Kinghorn, B.P. and Banks, R.G. (2010) Anim. Prod. Sci., 50(12):998–1003. https://doi.org/10.1071/AN10151

Zhang, Y.D., Swan, A., Johnston, D.J., and Girard, C.J. (2017) Proc of the 22nd AAABG, Townsville, Australia.