Accounting for trait-specific genomic and residual polygenic covariances in multivariate single-step genomic evaluation

Karin Meyer[*]

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

Abstract

For multivariate, single-step genomic best linear unbiased prediction analyses fitting a breeding value model, it is often assumed that the proportions of total genetic variance accounted for by genomic markers and residual polygenic effects are the same for all traits. Different covariance matrices for the two types of genetic effects are readily taken into account by fitting them separately. However, this can lead to slow convergence rates in iterative solution schemes. We propose an alternative computing strategy which – exploiting a canonical transformation – allows for trait-specific covariances whilst directly fitting total genetic effects only. Its effects on convergence rates and gains in accuracy and bias of genomic evaluation compared to analyses assuming proportionality of covariance matrices are examined using a small simulation study. Results show comparatively little improvement in accuracies but worthwhile reductions in overdispersion of predicted genetic merits for genotyped individuals without phenotypes.

Introduction

With the advent of genomic evaluation, it was quickly recognised that the single nucleotide polymorphism (SNP) markers available do not necessarily explain all genetic variation. This can be accounted for by adding a 'residual polygenic' (RPG) effect to the model of analysis fitted (e.g. Jensen *et al.*, 2012). For univariate single-step analyses fitting a breeding value model (ssGBLUP) an equivalent alternative is to replace the genomic relationship matrix, $\mathbf{G}$, with $\lambda\mathbf{G} + (1-\lambda)\mathbf{A}_{22}$, its weighted average with $\mathbf{A}_{22}$, the submatrix of the numerator relationship, $\mathbf{A}$, for genotyped animals and $\lambda \in [0,1]$ the proportion of the genetic variance attributed to SNP effects. This is readily extended to multivariate analyses if it can be assumed that $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_A$, the covariance matrices due SNP and RPG effects, are proportional.

This paper describes a computing strategy for ssGBLUP exploiting a transformation to canonical scale to avoid fitting a separate RPG effect when this does not hold, i.e. when $\mathbf{\Sigma}_G \neq c\mathbf{\Sigma}_A$ (for c a constant). In addition, a small simulation study is presented to evaluate the potential advantages of accounting for trait-specific weighting of the two sources of genetic variation.

Computing strategy

Consider a multivariate linear mixed model for t traits,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad (1)$$

with $\mathbf{y}$, $\mathbf{b}$, $\mathbf{u}$ and $\mathbf{e}$ denoting the vectors of observations, fixed effects, additive genetic effects and residuals and design matrices $\mathbf{X}$ and $\mathbf{Z}$. Let genetic effects be ordered by traits within individuals so that $\mathbf{u}$ can be partitioned into parts for n_1 non-genotyped ($\mathbf{u}_1$) and n_2 genotyped ($\mathbf{u}_2$) individuals. This gives the mixed model equations (MME)

$$\begin{bmatrix} \mathbf{X'R}^{-1}\mathbf{X} & \mathbf{X'R}^{-1}\mathbf{Z} \\ \mathbf{Z'R}^{-1}\mathbf{X} & \mathbf{Z'R}^{-1}\mathbf{Z} + \mathbf{A}^{-1} \otimes \mathbf{\Sigma}_U^{-1} + \mathbf{\Delta} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R}^{-1}\mathbf{y} \\ \mathbf{Z'R}^{-1}\mathbf{y} \end{bmatrix} \qquad (2)$$

with $\boldsymbol{\Sigma}_U = \boldsymbol{\Sigma}_G + \boldsymbol{\Sigma}_A$ the matrix of (total) genetic covariances between traits and the non-zero block of $\boldsymbol{\Delta}$ containing the 'add-on' for genotyped individuals, arising from the inverse of the joint relationship matrix for genotyped and non-genotyped individuals, $\mathbf{H}$ (Aguilar *et al.*, 2010).

$$\boldsymbol{\Delta} \;\; = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Delta}_G - \mathbf{A}_{22}^{-1} \otimes \boldsymbol{\Sigma}_U^{-1} \end{bmatrix} \tag{3}$$

For $\boldsymbol{\Sigma}_G = c\boldsymbol{\Sigma}_A$,

$$\boldsymbol{\Delta}_G = (\lambda\mathbf{G} + (1-\lambda)\mathbf{A}_{22})^{-1} \otimes \boldsymbol{\Sigma}_U^{-1} \tag{4}$$

i.e. a single matrix inverse of size $n_2 \times n_2$ suffices to obtain $\boldsymbol{\Delta}_G$. More generally, however, when $\boldsymbol{\Sigma}_G \neq c\boldsymbol{\Sigma}_A$, $\boldsymbol{\Delta}_G$ is not separable and

$$\boldsymbol{\Delta}_G = (\mathbf{G} \otimes \boldsymbol{\Sigma}_G + \mathbf{A}_{22} \otimes \boldsymbol{\Sigma}_A)^{-1} \tag{5}$$

so that the inverse of a matrix of size $tn_2 \times tn_2$ is needed. This can impose a considerable computational burden and become prohibitive for larger numbers of traits or genotyped animals.

Canonical transformation. For two conforming, symmetric matrices, $\boldsymbol{\Sigma}_1$ positive definite and $\boldsymbol{\Sigma}_2$ positive semi-definite, there exist a matrix $\mathbf{Q}$ so that $\mathbf{Q}\boldsymbol{\Sigma}_1\mathbf{Q}' = \mathbf{I}$ and $\mathbf{Q}\boldsymbol{\Sigma}_2\mathbf{Q}' = \mathbf{D}$ with $\mathbf{D}$ a diagonal matrix with elements, d_k, equal to the eigenvalues of $\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2$ (Anderson, 1958) and $\mathbf{I}$ an identity matrix. $\mathbf{Q}$ defines the so-called transformation to canonical scale which has seen considerable use to simplify multivariate linear mixed model analyses (e.g. Ducrocq and Besbes, 1993; Meyer, 1985). For $\mathbf{Q}\boldsymbol{\Sigma}_G\mathbf{Q}' = \mathbf{I}$ and $\mathbf{Q}\boldsymbol{\Sigma}_A\mathbf{Q}' = \mathbf{D}$, we can factor $\boldsymbol{\Delta}_G$ into three terms

$$\boldsymbol{\Delta}_G = [\mathbf{I} \otimes \mathbf{Q}'](\mathbf{G} \otimes \mathbf{I} + \mathbf{A}_{22} \otimes \mathbf{D})^{-1}[\mathbf{I} \otimes \mathbf{Q}] \tag{6}$$

with the middle term in Equation 6 the inverse of a matrix comprised of n_2^2 blocks of size $t \times t$ which are diagonal. Reordering the levels of $\mathbf{u}_2$ and $\boldsymbol{\Delta}_G$ according to animals within traits, denoted by superscript '$\star$', results in

$$\boldsymbol{\Delta}_G^{\star} \;\; = [\mathbf{Q}' \otimes \mathbf{I}](\mathbf{I} \otimes \mathbf{G} + \mathbf{D} \otimes \mathbf{A}_{22})^{-1}[\mathbf{Q} \otimes \mathbf{I}] \tag{7}$$

In this formulation, the middle term is block-diagonal with t blocks $(\mathbf{G} + d_k\mathbf{A}_{22})^{-1}$. Hence the $t(t+1)/2$ submatrices of $\boldsymbol{\Delta}_G^{\star}$ ($\boldsymbol{\Delta}_{G,ij}^{\star}$ for traits i and j) can be obtained as linear combinations of $r \leq t$ inverses of the individual blocks of size $n_2 \times n_2$, with r the number of distinct values d_k. For q_{ij} denoting the $ij-$th element of $\mathbf{Q}$,

$$\boldsymbol{\Delta}_{G,ij}^{\star} = \textstyle\sum_{k=1}^{t} q_{ki}\, q_{kj}(\mathbf{G} + d_k\mathbf{A}_{22})^{-1}. \tag{8}$$

Following Mäntysaari *et al.* (2017), we can use the Woodbury matrix identity to compute the inverses required. For $\mathbf{G} = \mathbf{M}\mathbf{M}'/s$ with $\mathbf{M}$ the matrix of centred marker counts and s a scalar,

$$\begin{aligned}(\mathbf{G} + d_k\mathbf{A}_{22})^{-1} \quad &= [\mathbf{A}_{22}^{-1} - \mathbf{A}_{22}^{-1}\,\mathbf{M}\,(s\,d_k\mathbf{I} + \mathbf{M}'\,\mathbf{A}_{22}^{-1}\,\mathbf{M})^{-1}\mathbf{M}'\,\mathbf{A}_{22}^{-1}]/d_k \\ &= [\mathbf{A}_{22}^{-1} - \mathbf{A}_{22}^{-1}\,\mathbf{M}\,\mathbf{E}\,(s\,d_k\mathbf{I} + \mathbf{T})^{-1}\,\mathbf{E}'\,\mathbf{M}'\,\mathbf{A}_{22}^{-1}]\,/d_k\end{aligned} \tag{9}$$

with $\mathbf{T}$ and $\mathbf{E}$ the matrices of eigenvalues and eigenvectors of $\mathbf{M}'\mathbf{A}_{22}^{-1}\mathbf{M}$, respectively. Matrices $(s\,d_k\mathbf{I} + \mathbf{T})$ are diagonal and thus trivial to invert. This facilitates efficient evaluation of the r diagonal blocks required through a single eigen-decomposition of a matrix of size $m \times m$, with m the number of markers.

Calculations can be implemented using either ordering of $\mathbf{u}_2$. We assume animals within traits in the following. This allows for the direct use of highly optimised library routines for dense matrix calculations to evaluate the inverses required. For analyses forming the MME in core, multiplication with $\mathbf{Q} \otimes \mathbf{I}$ can be integrated into the design matrix $\mathbf{Z}$, in the same way as for a transformation to principal components (Meyer *et al.*, 2015). Alternatively, using an iterative scheme to solve the MME, a product $\boldsymbol{\Delta}_G^{\star}\mathbf{r}_2^{\star}$ is required for each iterate (with $\mathbf{r}_2^{\star}$ the part of the vector of directions corresponding to $\mathbf{u}_2^{\star}$). This product is readily obtained by first forming $\mathbf{s}_2^{\star} = (\mathbf{Q} \otimes \mathbf{I})\mathbf{r}_2^{\star}$, then multiplying with the inverse diagonal blocks, yielding $\mathbf{t}_2^{\star} = (\mathbf{I} \otimes \mathbf{G} + \mathbf{D} \otimes \mathbf{A}_{22})^{-1}\mathbf{s}_2^{\star}$ and finally calculating $(\mathbf{Q}' \otimes \mathbf{I})\mathbf{t}_2^{\star}$.

Material & Methods

Data for three traits recorded on 21,000 animals in eight generations with heritabilities of 0.30, 0.25 and 0.10, respectively, and genetic correlations of 0.70 were generated using AlphaSim, version 1.05 (Faux *et al.*, 2016). Genotypes were constructed sampling 125 quantitative trait loci and 32,000 SNP markers. Subsequently, RPG effects were sampled for the given pedigree structure and added to the respective records and breeding values. Variances for RPG effects were chosen to yield variance ratios of magnitude 0.90, 0.85 and 0.20. Only marker information for 10%, 30%, 40% and 50% of randomly chosen individuals in generations five to eight was retained.

Genomic relationship matrices were build using Method 1 of Van Raden (2008), eliminating SNP with minor allele frequencies less than 2% and centering allele counts using 'observed' frequencies. To align $\mathbf{G}$ with $\mathbf{A}_{22}$ these were augmented by $\alpha\mathbf{J}$ with $\alpha = \mathbf{1}'(\mathbf{A}_{22} - \mathbf{G})\mathbf{1}/n_2^2$ for $\mathbf{1}$ and $\mathbf{J}$ a vector and a matrix, respectively, with all elements equal to unity (Vitezica *et al.* 2011). Estimates of variance components were obtained by restricted maximum likelihood (REML), fitting overall means as the only fixed effect. Trait-specific covariance matrices $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_A$ were estimated by splitting $\mathbf{u}$ into parts explained by markers, $\mathbf{g}$, and RPG effects, $\mathbf{a}$ and fitting these separately, assuming they were distributed as $\mathbf{H} \otimes \mathbf{\Sigma}_G$ and $\mathbf{A} \otimes \mathbf{\Sigma}_A$, respectively, with $\text{Cov}(\mathbf{g}, \mathbf{a}') = \mathbf{0}$. This was contrasted with analyses assuming $\mathbf{\Sigma}_G = c\mathbf{\Sigma}_A$ which fitted $\mathbf{u}$ with $\mathbf{\Delta}_G = (\lambda\mathbf{G} + (1 - \lambda)\mathbf{A}_{22})^{-1} \otimes \mathbf{\Sigma}_U^{-1}$ and estimated λ by quadratic approximation of its profile likelihood. Analyses were carried out both in- and excluding records for the last generation.

To evaluate the impact on convergence behaviour, MME were also solved using iteration on data. This was done fitting either both $\mathbf{g}$ and $\mathbf{a}$ (FULL) or $\mathbf{u}$ (JOINT), and considering $\mathbf{\Sigma}_G$ and $\mathbf{\Sigma}_A$ to be either proportional (SINGLE λ) or trait specific (VARIED λ). Summary statistics calculated were means and standard deviations across 30 replicates for estimates of λ, i.e the ratios of diagonal elements of $\mathbf{\Sigma}_G$ to $\mathbf{\Sigma}_U$, and correlations (ρ) between and linear regressions (β) of true on predicted total breeding values (EBV), for genotyped individuals in the last generation.

Results & Discussion

Means and standard deviations of summary statistics for VARIED λ and SINGLE λ together with corresponding values for analyses fitting single, fixed values of $\lambda = 0.50$ (FIX $\lambda = 0.50$) or $\lambda = 0.95$ (FIX $\lambda = 0.95$) are contrasted in Table 1 (analyses including all phenotypes). In spite of substantial differences in individual λ, improvements in ρ allowing for trait-specific values over those estimating a single value were modest. Similarly, deviations of β from the expected value of unity were only slightly reduced. Only for trait three and FIX $\lambda = 0.95$ was a sizeable impact on ρ evident. Omitting phenotypes for generation eight (detailed results not shown) resulted in lower levels of ρ overall but yielded only slightly bigger improvements for VARIED λ over SINGLE λ

Table 1. Means ($\bar{x}$) and standard deviations (sd) for estimates of variance ratios (λ), correlations (ρ) between and regressions (β) of true on predicted breeding values.

		VARIED λ			SINGLE λ			FIX $\lambda = 0.95$			FIX $\lambda = 0.50$		
Trait		1	2	3	1	2	3	1	2	3	1	2	3
$\hat{\lambda}$	$\bar{x}$	0.909	0.833	0.180	0.574	0.574	0.574	0.950	0.950	0.950	0.500	0.500	0.500
	sd	0.035	0.032	0.040	0.055	0.055	0.055	-	-	-	-	-	-
$\hat{\rho}$	x	0.855	0.833	0.735	0.832	0.817	0.710	0.852	0.824	0.609	0.826	0.812	0.718
	sd	0.011	0.014	0.023	0.016	0.016	0.029	0.012	0.016	0.025	0.014	0.015	0.024
$\hat{\beta}$	$\bar{x}$	0.966	0.978	0.992	0.951	0.968	1.017	0.970	0.981	0.985	0.947	0.965	1.013
	sd	0.022	0.027	0.034	0.027	0.030	0.035	0.022	0.029	0.037	0.026	0.030	0.035

than above. However this was accompanied by a drop of about 0.1 (from 0.984 to 0.874) in mean β for SINGLE λ for trait three, i.e. a marked increase in over-dispersion of EBVs.

Additional univariate analyses (not shown) evaluating the profile for ρ and β over the whole range of λ showed that values were rather robust against considerable deviations of λ from their estimated, 'optimal' values. Similarly, literature results emphasised the insensitivity in ranking of ssGBLUP predictions of genetic merit against choices of λ (e.g. McMillan and Swan, 2017).

Whilst resulting in the same predicted breeding values, attempts to split $\mathbf{u}$ into its parts, $\mathbf{g}$ and $\mathbf{a}$, (FULL) were clearly detrimental for convergence rates of iterative solution schemes. Means and ranges (in brackets) for the number of iterates required to solve the MME for VARIED λ were 1036 (631-6573) and 187 (141-263) for FULL and JOINT, respectively, i.e. JOINT reduced numbers of iterates required on average by about 80%. The highest numbers of iterates for the combination of FULL and VARIED λ were required for samples with individual λ close to unity (or zero), reflecting their effects on the condition number of the coefficient matrix in the MME. Corresponding numbers for SINGLE λ were 663 (563-818) and 152 (122-190).

Conclusions

Use of a canonical transformation facilitates trait-specific weighting of genetic covariances explained by genomic markers and RPG effects in ssGBLUP analyses, directly fitting total breeding values only. However, simulation results suggest that the resulting gains in accuracy of prediction compared to estimating a single, overall value of λ tend to be limited. Accommodating individual λ appeared to be most beneficial for a scenario with the values spanning almost the whole range of $[0,1]$ and in predicting breeding values for genotyped individuals without phenotypes. In practical applications, the trade-off between potential gains and additional computational demands need to be carefully assessed.

Acknowledgements

Work was supported by Meat and Livestock Australia grants L.GEN.1704 and L.GEN.2204.

References

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S., *et al*. (2010) J. Dairy Sci. 93 (2): 743–752. https://doi.org/10.3168/jds.2009-2730.

Anderson T. W. (1958) An Introduction to Multivariate Statistical Analysis. Wiley New York.

Ducrocq V., and Besbes B. (1993) J. Anim. Breed. Genet. 110: 81–92. https://doi.org/10.1111/j.1439-0388.1993.tb00719.x

Faux A-M, Gorjanc G., Gaynor R.C., Battagin M., Edwards S.M., *et al.* (2016) Plant Genome 9 (3): 1–14. https://doi.org/10.3835/plantgenome2016.02.0013.

Jensen J., Su G., and Madsen P. (2012) BMC Genetics 13: 44. https://doi.org/10.1186/1471-2156-13-44.

Mäntysaari E.A., Evans R.D., and Strandén I. (2017) J. Anim. Sci. 95 (11): 4728–4737. https://doi.org/10.2527/jas2017.1912.

McMillan A.J., and Swan A.A. (2017). Proc. Ass. Advan. Anim. Breed. Genet. 22: Paper no.130.

Meyer K. (1985) Biometrics 41 (1): 153–66. https://doi.org/10.2307/2530651.

Meyer K., Swan A.A., and Tier B. (2015) J. Anim. Sci. 93 (10): 4624–4628. https://doi.org/10.2527/jas.2015-9333.

Van Raden P. M. (2008) J. Dairy Sci. 91 (11): 4414–4423. https://doi.org/10.3168/jds.2007-0980.

Vitezica Z. G., Aguilar I., Misztal I., and Legarra. A. (2011). Genet. Res. 93 (5): 357–366. https://doi.org/10.1017/S001667231100022X.