G-Blup without Inverting G

***J.B.C.H.M. van Kaam*[1]**
[1]*Associazone Nazionale Allevatori Frisona Italiana, Cremona, Italy*

Abstract

Regular mixed model methodology requires inversion of the variance-covariance matrix of the random effects. In the regular animal model the inverse of the numerator relationship matrix (**A**) can be easily set up directly due to its sparse nature with many simple coefficients. In G-Blup a genomic relationship matrix (**G**) is used, which is a dense matrix wherein the coefficients are determined by many loci. Inversion of the **G** matrix is a time-consuming or even a limiting factor with increasing training populations. Assuming a genomic model including SNP effects and a residual polygenic component, a derivation of G-Blup without inverting the **G** matrix is presented here. DGVs can be computed from the residual polygenic components given the relationship matrices **G** and **A** and the variance components.

Key words: genomic evaluation, G-Blup, inverse, relationship matrix

Introduction

The invention of a method to directly obtain the inverse of the numerator relationship matrix (**A**) by Henderson (1975) has been a breakthrough in the practical applicability of the animal model. Such a rapid derivation of the inverse is possible because (1) most of the entries in the inverse are zero and (2) few different values occur.

The advent of genomic selection leads to a new situation whereby the traditional numerator relationship matrix can be replaced by a more accurate genomic relationship matrix (**G**). The genomic relationship matrix contains actual relationship coefficients instead of expected values. Therefore the numerator relationship matrix can be considered as a simple approximation of the genomic relationship matrix. This matrix however is dense and since many loci contribute to the coefficients, there are many different values.

Regular mixed model methodology requires matrix inversion for the variance-covariance matrix of the random effects. However matrix inversion is computationally very costly for matrices with a dimension in the thousands and scales cubically with the matrix size.

The genomic relationship matrix contains a row/column for each unique individual used in the training population. When G-Blup is applied to estimate direct genomic values the inversion of the **G** matrix is a time-consuming or even limiting factor with increasing training populations. An easy method to obtain the inverse in a rapid manner is not likely to be found due to the dense nature and the large number of loci determining the matrix elements. An alternative approach might be to avoid the need for an inverted **G** matrix and solve G-Blup in another way.

An interesting paper has been presented by Jafarikia *et al.* (2006) where QTL solutions are obtained without inverting an IBD matrix. Their approach has been adapted here in the context of G-Blup. Instead of an IBD matrix, a genomic relationship matrix is used, whereas instead of individual QTL solutions, the interest is in direct genomic values.

Methods

Consider an additive model in which the sum of a direct genomic value (*dgv*) containing many SNPs and a residual polygenic component (*a*) is an estimate of the total genetic value (*y*) with an error effect (*e*):

$$y = a + dgv + e$$

Mixed model equations for this model:

$$\begin{bmatrix} \mathbf{Z_1'Z_1} + \boldsymbol{\lambda}_\mathbf{A}\mathbf{A}^{-1} & \mathbf{Z_1'Z_2} \\ \mathbf{Z_2'Z_1} & \mathbf{Z_2'Z_2} + \boldsymbol{\lambda}_\mathbf{G}\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{a} \\ d\hat{g}v \end{bmatrix} = \begin{bmatrix} \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \end{bmatrix}$$

where **A** and **G** contain the same animals in the same order and $\mathbf{Z_1}$ and $\mathbf{Z}_2$ are incidence matrices. Two variance ratios for the random effects are $\boldsymbol{\lambda}_\mathbf{A} = \sigma_e^2 / \sigma_a^2$ and $\boldsymbol{\lambda}_\mathbf{G} = \sigma_e^2 / \sigma_{dgv}^2$.

This can be rewritten as:

$$\begin{aligned} \left(\mathbf{Z_1'Z_1} + \boldsymbol{\lambda}_\mathbf{A}\mathbf{A}^{-1}\right)\hat{a} + \left(\mathbf{Z_1'Z_2}\right)d\hat{g}v &= \mathbf{Z_1'y} \\ \left(\mathbf{Z_2'Z_1}\right)\hat{a} + \left(\mathbf{Z_2'Z_2} + \boldsymbol{\lambda}_\mathbf{G}\mathbf{G}^{-1}\right)d\hat{g}v &= \mathbf{Z_2'y} \end{aligned}$$

Because $\mathbf{Z_1}$ equals $\mathbf{Z}_2$, they both can be replaced with $\mathbf{Z}$:

$$\begin{aligned} \left(\mathbf{Z'Z} + \boldsymbol{\lambda}_\mathbf{A}\mathbf{A}^{-1}\right)\hat{a} + \left(\mathbf{Z'Z}\right)d\hat{g}v &= \mathbf{Z'y} \\ \left(\mathbf{Z'Z}\right)\hat{a} + \left(\mathbf{Z'Z} + \boldsymbol{\lambda}_\mathbf{G}\mathbf{G}^{-1}\right)d\hat{g}v &= \mathbf{Z'y} \end{aligned}$$

The RHS of both equations is the same so the LHS must be the same as well:

$$\begin{aligned} &\left(\mathbf{Z'Z} + \boldsymbol{\lambda}_\mathbf{A}\mathbf{A}^{-1}\right)\hat{a} + \left(\mathbf{Z'Z}\right)d\hat{g}v = \\ &\left(\mathbf{Z'Z}\right)\hat{a} + \left(\mathbf{Z'Z} + \boldsymbol{\lambda}_\mathbf{G}\mathbf{G}^{-1}\right)d\hat{g}v \end{aligned}$$

Rewriting:

$$\begin{aligned} &\left(\mathbf{Z'Z}\right)\hat{a} + \left(\boldsymbol{\lambda}_\mathbf{A}\mathbf{A}^{-1}\right)\hat{a} + \left(\mathbf{Z'Z}\right)d\hat{g}v = \\ &\left(\mathbf{Z'Z}\right)\hat{a} + \left(\mathbf{Z'Z}\right)d\hat{g}v + \left(\boldsymbol{\lambda}_\mathbf{G}\mathbf{G}^{-1}\right)d\hat{g}v \end{aligned}$$

Simplifying by removing redundant terms:

$$\left(\boldsymbol{\lambda}_\mathbf{A}\mathbf{A}^{-1}\right)\hat{a} = \left(\boldsymbol{\lambda}_\mathbf{G}\mathbf{G}^{-1}\right)d\hat{g}v$$

Multiply both sides with **G**:

$$\left(\boldsymbol{\lambda}_\mathbf{A}\mathbf{A}^{-1}\right)\mathbf{G}\hat{a} = \left(\boldsymbol{\lambda}_\mathbf{G}\mathbf{G}^{-1}\right)\mathbf{G}d\hat{g}v$$

Simplifying:

$$\left(\boldsymbol{\lambda}_\mathbf{A}\mathbf{G}\mathbf{A}^{-1}\right)\hat{a} = \boldsymbol{\lambda}_\mathbf{G}d\hat{g}v$$

Solve for dgv:

$$d\hat{g}v = \frac{\left(\boldsymbol{\lambda}_\mathbf{A}\mathbf{G}\mathbf{A}^{-1}\right)}{\boldsymbol{\lambda}_\mathbf{G}}\hat{a}$$

with $\boldsymbol{\lambda}_\mathbf{A} = \sigma_e^2 / \sigma_a^2$ and $\boldsymbol{\lambda}_\mathbf{G} = \sigma_e^2 / \sigma_{dgv}^2$, which gives:

$$d\hat{g}v = \frac{\left(\sigma_{dgv}^2\mathbf{G}\mathbf{A}^{-1}\right)}{\sigma_a^2}\hat{a} = \frac{\sigma_{dgv}^2}{\sigma_a^2}\mathbf{G}\mathbf{A}^{-1}\hat{a}$$

which is:

$$d\hat{g}v = \frac{\sigma_{dgv}^2\mathbf{G}}{\sigma_a^2\mathbf{A}}\hat{a}$$

Therefore DGVs can be computed from the residual polygenic components given the relationship matrices and the variance components. Usually estimates of the total additive genetic variance are available and the division of this variance between the SNP component and the residual polygenic component is empirically determined.

An iterative method, which can be used to solve these equations:

1. Set $\hat{a} = 0$, $d\hat{g}v = 0$ and set some values for the variance components
2. Compute $\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}'\left(d\hat{g}v\right)$
3. Solve these MME to obtain $\hat{a}$:
$$\left[\mathbf{Z'Z} + \frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1}\right]\left[\hat{a}\right] = \left[\mathbf{Z'}\tilde{\mathbf{y}}\right]$$
4. Compute $d\hat{g}v = \frac{\sigma_{dgv}^2}{\sigma_a^2}\mathbf{G}\mathbf{A}^{-1}\hat{a}$
5. Go to step 2 until solutions converge

This approach is an iterative approach in which another iteration, step 3, is nested. This inner iteration could be solved with a standard pre-conditioned conjugate gradient solver. The coefficient matrix in step 3 needs to be setup only once. Also the $\frac{\sigma_{dgv}^2}{\sigma_a^2}\mathbf{G}\mathbf{A}^{-1}$ matrix in step 4 needs to be computed just once. The outer iteration then requires only the computation of the $\tilde{\mathbf{y}}$ and $d\hat{g}v$ vectors, which is a matrix times vector multiplication for both, and the calculation of a convergence criterion.

Discussion

The proposed approach requires the presence of a residual polygenic component in the genetic model applied. Furthermore, if the fraction of variance assigned to the residual polygenic component is small then convergence might become more difficult. Possibly increasing the $\frac{\sigma^2_{dgv}}{\sigma^2_a}$ ratio in a few steps towards the desired level might improve convergence.

Another approach to avoid inversion of the **G** matrix has been presented by VanRaden (2012) who in turn referred to Legarra *et al.* (2011). Their algorithm appends extra equations that include the genomic relationship matrix instead of its inverse and the pedigree relationship matrix for genotyped animals instead of its inverse to the mixed model equations. The iterative strategy proposed here as well as the one of Legarra *et al.* (2011) have not yet been applied to real data sets.

A disadvantage of an approach whereby a large matrix inverse is avoided is that the inverse is often also useful for computing reliabilities. However in practice for larger populations approximate methods are developed and deployed.

Furthermore the use of a SNP-Blup model (i.e. solving SNP effects and then summing these) instead of a G-Blup model can achieve a similar effect. These two models are equivalent if SNP effects are assumed to be normally distributed (Goddard, 2009). So the G-Blup approach can also be avoided by adopting a SNP-Blup model. However the equivalence of G-Blup and SNP-Blup assumes a model without residual polygenic effect. Here I use w for the fraction of additive genetic variance assigned to the residual polygenic component. In case of a model with SNPs and a residual polygenic component in the G-Blup approach a weighted relationship matrix $\left(w\mathbf{A}+(1-w)\mathbf{G}\right)^{-1}$ is inverted, whereas with the SNP-Blup approach the inverse of each matrix is weighted so it is similar to $w\mathbf{A}^{-1}+(1-w)\mathbf{G}^{-1}$ with of course each component multiplied with the respective random vectors and incidence matrices for animals and SNPs. Note however that the inverse of the weighted matrix does not equal the weighted inverses of the two component matrices, i.e. $\left(w\mathbf{A}+(1-w)\mathbf{G}\right)^{-1} \neq w\mathbf{A}^{-1}+(1-w)\mathbf{G}^{-1}$. It matters whether the weighting is done on the relationship matrices or on the inverse of these matrices. So in case a residual polygenic component is present in the model then the equivalence of G-Blup and SNP-Blup no longer holds.

Conclusions

In a model including SNP effects and a residual polygenic effect, G-Blup can be solved without inversion of the **G** matrix. Hence the inversion of a complicated dense matrix can be avoided. Solutions can be obtained by an iterative procedure where an inner iteration is nested within an outer iteration. In case no residual polygenic effect is included then SNP-Blup can be used, while maintaining equivalence with G-Blup.

Acknowledgements

The author thanks Theo Meuwissen (Norwegian University of Life Sciences, Ås, Norway) for useful discussion.

References

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica 136*, 245-257.

Jafarikia, M., Susanto, A., Robinson, J.A.B. & Schaeffer, L.R. 2006. Method for obtaining QTL solutions without inverting the IBD matrix. *8th WCGALP*.

Legarra, A., Misztal, I. & Aguilar, I. 2011. The single step: Genomic evaluation for all. Book of Abstr. of 62nd Annu. Mtg. of Euro. Fed. of Anim. Sci. No. 17, 1. Wageningen Academic Publishers, The Netherlands.

VanRaden, P.M., 2012. Avoiding Bias From Genomic Pre-Selection in Converting Daughter Information Across Countries. *Interbull Bulletin 45.*