An efficient method to calculate accuracy of estimated breeding values for individuals without phenotypes

M.H. Ferdosi[1,2]*, N. Connors*[1] *& B. Tier*[1]

[1]*Animal Genetics and Breeding Unit, University of New England, Armidale, NSW, Australia*
[2]*mferdos3@une.edu.au* (*Corresponding Author*)

Summary

Improved methodology to update the inverse of the coefficient matrix (**C**) for new individuals without phenotype is described here. Computational performance is significantly improved by re-using parts of the coefficient matrix inverse calculations that do not change from one animal to another, in combination with updated calculations for those that do change. This efficient method delivers more than 500-fold improvement in performance.

Keywords: accuracy, efficient, breeding value

Introduction

In the last decade technological advances have decreased the genotyping costs, particularly for agricultural livestock and cropping species. As a result size of the genomic data repository has been increasing exponentially. Therefore efficient methods and tools are required to analyse this data. One application of genomic data in animal breeding is to increase the accuracy of estimated breeding values by replacing the numerator relationship matrix (NRM) with the genomic relationship matrix (GRM) (VanRaden, 2008), known as Genomic Best Linear Unbiased Prediction-Estimation (GBLUP). Using genomic data to predict breeding values allows young animals to be selected thus decreases the generation interval. The first question that arises is how much accuracy can be achieved for the young animals that are not phenotyped? Henderson (1975) derived a method to estimate the accuracy of each estimated breeding value (EBV), using the inverse of the coefficient matrix ($\mathbf{C}^{-1}$) and the covariance matrix for the random genetic effects (**G**). Many young animals could have a genotype, but no phenotype. These animals can be omitted from the evaluation as they provide no information. However, their EBVs can be calculated as the index of their genotype. To calculate the accuracy we require the diagonal of $\mathbf{C}^{-1}$ for the animals with genotype and no phenotype. This paper shows how to calculate accuracy for animals with only genotype, as a function of its genotype and the $\mathbf{C}^{-1}$ one at a time.

Method

Theory

We consider a simple animal model without fixed effects. This model is

$$\mathbf{y} = \mathbf{Zu} + \mathbf{e} \tag{1}$$

where **y**, **Z**, **u** and **e** are vector of observations, design matrix, vector of solutions and vector of random residual effects, respectively. The solutions and residual variances are

$var(\mathbf{u}) = \mathbf{G}\sigma_u^2$ and $var(\mathbf{e}) = \mathbf{I}\sigma_e^2$. The mixed model equations (MME) for the above model are

$$\mathbf{Cu} = [\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{G}^{-1}]\mathbf{u} = \mathbf{r}, \tag{2}$$

where $\mathbf{C}$ is the coefficient matrix and $\alpha = \frac{\sigma_e^2}{\sigma_u^2}$. Henderson derived a method by using the diagonal of $\mathbf{C}^{-1}$ and the diagonal of $\mathbf{G}$ to calculate the accuracy of each estimated breeding value (Henderson, 1975). Accordingly, the accuracy can be calculated with this formula $\sqrt{1 - \alpha\frac{c^{ii}}{g_{ii}}}$, where c^{ii} is the diagonal element of $\mathbf{C}^{-1}$ for individual i, and g_{ii} is the diagonal element of $\mathbf{G}$ for individual i.

Updating $\mathbf{C}^{-1}$ to illustrate the proposed method

To calculate the accuracy of individuals without phenotype, each individual can be added to $\mathbf{C}^{-1}$ separately. In our case, where there is an individual without phenotype, the partitioned matrix of MME (equation 2) is

$$\begin{bmatrix} \mathbf{C}_{pp} & \mathbf{c}_{pq} \\ \mathbf{c}'_{pq} & c_{qq} \end{bmatrix} \begin{bmatrix} \mathbf{u}_p \\ u_q \end{bmatrix} = \begin{bmatrix} \mathbf{r}_p \\ 0 \end{bmatrix} \tag{3}$$

where subscript p and q are individuals with and without phenotypes respectively.

As demonstrated in equation (2), $\mathbf{Z}'\mathbf{Z}$ becomes $\begin{pmatrix} (\mathbf{Z}'\mathbf{Z})_{pp} & 0 \\ 0 & 0 \end{pmatrix}$ where $(\mathbf{Z}'\mathbf{Z})_{pp}$ is a diagonal matrix with dimension equal to the number of animals with phenotypes, and the 0 in the lower diagonal represents an animal without phenotype. The $\mathbf{G}^{-1}$ becomes $\begin{pmatrix} \mathbf{G}_{pp} & \mathbf{g}_{pq} \\ \mathbf{g}'_{pq} & g_{qq} \end{pmatrix}^{-1}$, and therefore $\mathbf{C}^{-1}$ becomes

$$[\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{G}^{-1}]^{-1} \approx \left(\begin{bmatrix} (\mathbf{Z}'\mathbf{Z})_{pp} & 0 \\ 0 & \varepsilon \end{bmatrix} + \alpha \begin{bmatrix} \mathbf{G}_{pp} & \mathbf{g}_{pq} \\ \mathbf{g}'_{pq} & g_{qq} \end{bmatrix}^{-1} \right)^{-1} \tag{4}$$

based on equation (2) and (3) where ε is a very small number required to make $\mathbf{Z}'\mathbf{Z}$ invertible. Inverting $\mathbf{C}$ is computationally demanding, as $\mathbf{G}$ and the entire $\mathbf{C}$ should be inverted in each analysis for all individuals without phenotypes. $\mathbf{G}_{pp}^{-1}$ needs to be updated as the new individuals were added to the $\mathbf{G}$. This can be performed by the method explained in (Meyer *et al.*, 2013). However since we want to know the accuracy, we must invert $\mathbf{C}$ as well as $\mathbf{G}$. Equation (4) can be converted with the following inversion lemma which is equivalent to the Woodbury's formula (Lienart, 2017):

$$(\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{E})^{-1} = \boldsymbol{E}^{-1}\boldsymbol{D}(\boldsymbol{D} - \boldsymbol{E}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{E}\boldsymbol{A}^{-1}. \tag{5}$$

With $\mathbf{C}$ we can consider $\boldsymbol{A} = \mathbf{Z}'\mathbf{Z}$, $\boldsymbol{B} = -\mathbf{I}$, $\boldsymbol{D}^{-1} = \mathbf{G}^{-1}$ and $\boldsymbol{E} = \alpha\mathbf{I}$. Thus $\mathbf{C}^{-1}$ is

$$(\mathbf{Z}'\mathbf{Z} + \alpha\mathbf{G}^{-1})^{-1} = \alpha^{-1}\mathbf{G}(\mathbf{G} + \alpha(\mathbf{Z}'\mathbf{Z})^{-1})^{-1}\alpha(\mathbf{Z}'\mathbf{Z})^{-1} = \mathbf{G}\mathbf{M}^{-1}(\mathbf{Z}'\mathbf{Z})^{-1} \tag{6}$$

where $\mathbf{M}^{-1}$ as $(\mathbf{G} + \alpha(\mathbf{Z}'\mathbf{Z})^{-1})^{-1}$ is used for simplification and is shown below in equation (9). The partitioned matrices are

$$\left(\begin{bmatrix}\mathbf{Z}'\mathbf{Z}_{pp} & 0\\ 0 & \varepsilon\end{bmatrix}+\alpha\begin{bmatrix}\mathbf{G}_{pp} & \mathbf{g}_{pq}\\ \mathbf{g}'_{pq} & g_{qq}\end{bmatrix}^{-1}\right)^{-1}$$
$$\approx\begin{bmatrix}\mathbf{G}_{pp} & \mathbf{g}_{pq}\\ \mathbf{g}'_{pq} & g_{qq}\end{bmatrix}\left(\begin{bmatrix}\mathbf{G}_{pp} & \mathbf{g}_{pq}\\ \mathbf{g}'_{pq} & g_{qq}\end{bmatrix}+\alpha\begin{bmatrix}(\mathbf{Z}'\mathbf{Z})^{-1}_{pp} & 0\\ 0 & \frac{1}{\varepsilon}\end{bmatrix}\right)^{-1}\begin{bmatrix}(\mathbf{Z}'\mathbf{Z})^{-1}_{pp} & 0\\ 0 & \frac{1}{\varepsilon}\end{bmatrix}$$
$$\approx\begin{bmatrix}\mathbf{G}_{pp} & \mathbf{g}_{pq}\\ \mathbf{g}'_{pq} & g_{qq}\end{bmatrix}\begin{bmatrix}\mathbf{G}_{pp}+\alpha(\mathbf{Z}'\mathbf{Z})^{-1}_{pp} & \mathbf{g}_{pq}\\ \mathbf{g}'_{pq} & g_{qq}+\frac{\alpha}{\varepsilon}\end{bmatrix}^{-1}\begin{bmatrix}(\mathbf{Z}'\mathbf{Z})^{-1}_{pp} & 0\\ 0 & \frac{1}{\varepsilon}\end{bmatrix}. \quad (7)$$

By using lemma (6) the $\mathbf{G}^{-1}$ is not required and we are only required to invert the middle matrix ($\mathbf{M}$) in equation (7). With this simplification $\mathbf{M}^{-1}$ can be updated for each new individual using Cholesky decomposition and multiplying the Cholesky factors, i.e $\mathbf{M}^{-1} = \mathbf{L}^{-T}\mathbf{L}^{-1}$ (Harville, 1997; Meyer *et al.*, 2013).

$$\mathbf{M}^{-1}=\begin{bmatrix}\mathbf{G}^{-1}_{pp}+\mathbf{L}^{-T}_{pp}\boldsymbol{\ell}'_{pq}\ell^{-1}_{qq}\ell^{-1}_{qq}\boldsymbol{\ell}_{pq}\mathbf{L}^{-1}_{pp} & -\mathbf{L}^{-T}_{pp}\boldsymbol{\ell}'_{pq}\ell^{-1}_{qq}\ell^{-1}_{qq}\\ -\ell^{-1}_{qq}\ell^{-1}_{qq}\boldsymbol{\ell}_{pq}\mathbf{L}^{-1}_{pp} & \ell^{-1}_{qq}\ell^{-1}_{qq}\end{bmatrix}=\begin{bmatrix}\mathbf{S}_1 & \mathbf{s}_2\\ \mathbf{s}_3 & s_4\end{bmatrix}, \quad (8)$$

Therefore equation (7) can be written as

$$\begin{bmatrix}\mathbf{C}_{pp} & \mathbf{c}_{pq}\\ \mathbf{c}'_{pq} & c_{qq}\end{bmatrix}\approx\begin{bmatrix}\mathbf{G}_{pp} & \mathbf{g}_{pq}\\ \mathbf{g}'_{pq} & g_{qq}\end{bmatrix}\begin{bmatrix}\mathbf{S}_1 & \mathbf{s}_2\\ \mathbf{s}_3 & s_4\end{bmatrix}\begin{bmatrix}(\mathbf{Z}'\mathbf{Z})^{-1}_{pp} & 0\\ 0 & \frac{1}{\varepsilon}\end{bmatrix}$$
$$\approx\begin{bmatrix}\mathbf{G}_{pp}\mathbf{S}_1+\mathbf{g}_{pq}\mathbf{s}_3 & \mathbf{G}_{pp}\mathbf{S}_2+\mathbf{g}_{pq}s_4\\ \mathbf{g}'_{pq}\mathbf{S}_1+g_{qq}\mathbf{s}_3 & \mathbf{g}'_{pq}\mathbf{s}_2+g_{qq}s_4\end{bmatrix}\begin{bmatrix}(\mathbf{Z}'\mathbf{Z})^{-1}_{pp} & 0\\ 0 & \frac{1}{\varepsilon}\end{bmatrix}$$
$$\approx\begin{bmatrix}(\mathbf{G}_{pp}\mathbf{S}_1+\mathbf{g}_{pq}\mathbf{s}_3)(\mathbf{Z}'\mathbf{Z})^{-1}_{pp} & \frac{1}{\varepsilon}(\mathbf{G}_{pp}\mathbf{S}_2+\mathbf{g}_{pq}s_4)\\ (\mathbf{g}'_{pq}\mathbf{S}_1+g_{qq}\mathbf{s}_3)(\mathbf{Z}'\mathbf{Z})^{-1}_{pp} & \frac{1}{\varepsilon}(\mathbf{g}'_{pq}\mathbf{s}_2+g_{qq}s_4)\end{bmatrix}, \quad (9)$$

based on equation (8) $\mathbf{s}_2 = -\mathbf{L}^{-T}_{pp}\boldsymbol{\ell}'_{pq}\ell^{-1}_{qq}\ell^{-1}_{qq}$ and $s_4 = \ell^{-1}_{qq}\ell^{-1}_{qq}$. By multiplying back the Cholesky factors of $\mathbf{M}$ the solutions for $\boldsymbol{\ell}'_{pq}$ and ℓ_{qq} are

$$\boldsymbol{\ell}'_{pq} = \mathbf{L}^{-1}_{pp}\mathbf{g}_{pq} \quad (10)$$

and

$$\ell_{qq} = \sqrt{g_{qq}+\frac{\alpha}{\varepsilon}-\mathbf{g}'_{pq}(\mathbf{G}_{pp}+\alpha(\mathbf{Z}'\mathbf{Z})^{-1}_{pp})^{-1}\mathbf{g}_{pq}}, \quad (11)$$

since ℓ_{qq} is a number $\mathbf{s}_4$ become ℓ^{-2}_{qq},

$$c^{qq}\approx\frac{1}{\varepsilon}(\mathbf{g}'_{pq}(-\mathbf{L}^{-T}_{pp}\boldsymbol{\ell}'_{pq}\ell^{-1}_{qq}\ell^{-1}_{qq})+g_{qq}(\ell^{-1}_{qq}\ell^{-1}_{qq}))\approx\frac{1}{\varepsilon}(\mathbf{g}'_{pq}(-\mathbf{L}^{-T}_{pp}\boldsymbol{\ell}'_{pq}s_4)+g_{qq}(s_4))$$
$$\approx\frac{s_4}{\varepsilon}(\mathbf{g}'_{pq}(-\mathbf{L}^{-T}_{pp}\boldsymbol{\ell}'_{pq})+g_{qq})\approx\frac{-\mathbf{g}'_{pq}\mathbf{L}^{-T}_{pp}\boldsymbol{\ell}'_{pq}+g_{qq}}{\varepsilon(g_{qq}+\frac{1}{\varepsilon}-\mathbf{g}'_{pq}(\mathbf{G}_{pp}+\alpha(\mathbf{Z}'\mathbf{Z})^{-1}_{pp})^{-1}\mathbf{g}_{pq})}$$
$$\approx\frac{-\mathbf{g}'_{pq}\mathbf{L}^{-T}_{pp}\mathbf{L}^{-1}_{pp}\mathbf{g}_{pq}+g_{qq}}{\varepsilon(g_{qq}+\frac{\alpha}{\varepsilon}-\mathbf{g}'_{pq}(\mathbf{G}_{pp}+\alpha(\mathbf{Z}'\mathbf{Z})^{-1}_{pp})^{-1}\mathbf{g}_{pq})} \quad (12)$$

Only $\mathbf{g}'_{pq}$, $\mathbf{g}_{pq}$ and g_{qq} change with each non-phenotype individual. Other variables need to be calculated only once. The denominator is a number and its limit approaches α as ε approaches zero.

$$c^{qq} = \lim_{\varepsilon \to 0} \frac{-\mathbf{g}'_{pq}\mathbf{L}_{pp}^{-T}\mathbf{L}_{pp}^{-1}\mathbf{g}_{pq} + g_{qq}}{\varepsilon(g_{qq} + \frac{\alpha}{\varepsilon} - \mathbf{g}'_{pq}(\mathbf{G}_{pp} + \alpha(\mathbf{Z}'\mathbf{Z})^{-1}_{pp})^{-1}\mathbf{g}_{pq})} = (-\mathbf{g}'_{pq}\mathbf{L}_{pp}^{-T}\mathbf{L}_{pp}^{-1}\mathbf{g}_{pq} + g_{qq})/\alpha \quad (13)$$

and $\mathbf{L}_{pp}^{-T}\mathbf{L}_{pp}^{-1} = (\mathbf{G}_{pp} + \alpha(\mathbf{Z}'\mathbf{Z})^{-1}_{pp})^{-1}$ so

$$c^{qq} = (-\mathbf{g}'_{pq}(\mathbf{G}_{pp} + \alpha(\mathbf{Z}'\mathbf{Z})^{-1}_{pp})^{-1}\mathbf{g}_{pq} + g_{qq})/\alpha \quad (14)$$

The last equation (14) was implemented as an R function to assess its performance.

Genomic data and genomic relationship matrix

This method was applied to 11,000 Brahman animals. The details of quality controls are explained in (Connors *et al.*, 2017). Sets of 500, 1000, 1500, ..., 11,000 animals in $\mathbf{G}_{pp}$ were considered as animals with phenotype and accuracies were calculated for the last individual (one in each run) in each set. The genomic relationship matrix was built using VanRaden (2008) first method.

Performance evaluation

To evaluate performance, each set was run in three steps. In the first step, the elapsed time to build the coefficient matrix by using the classic approach (i.e. inverting $\mathbf{G}$ and $\mathbf{C}$) was measured. In the second step, the time to build $(\mathbf{G}_{pp} + \alpha(\mathbf{Z}'\mathbf{Z})^{-1}_{pp})^{-1}$ – initial matrices required to update c_{qq} was measured. In the third step, the time to calculate c_{qq} by using the initial matrices was measured.

Result and discussion

By calculating the accuracy of young individuals using equation (14) computational times have been significantly reduced. Through using this equation computational performance has been improved by more than 500 times with only negligible differences in accuracy due to rounding errors (less than 8.88×10^{-16}). As described above, since the ε approaches zero, the denominator becomes α. Thus c_{qq} is the exact prediction error variance (c^{qq}) not an approximation. Because the proposed approach used $\mathbf{Z}'\mathbf{Z}$ (a diagonal matrix), the time required to build the matrices used to update c^{qq} was shorter compared to when using the classic approach to calculate accuracies. This method can be extended in order to accommodate fixed effects and dense $\mathbf{Z}'\mathbf{Z}$ when c^{qq} is updated. Furthermore, the part of $\mathbf{C}$ for animals with phenotype ($\mathbf{C}^{pp}$) must be updated as more individuals are phenotyped.

Conclusion

Updating the inverse of $\mathbf{C}$ for new individuals without phenotype, using the method here, is shown to reduce the computational effort significantly. With increasing numbers of genotyped animals in genetic evaluations, computational efficiency is essential for frequent and timely evaluations. This method provides an improved and efficient method to deliver accurate and fast evaluations when few new young individuals are genotyped but have no phenotypes.

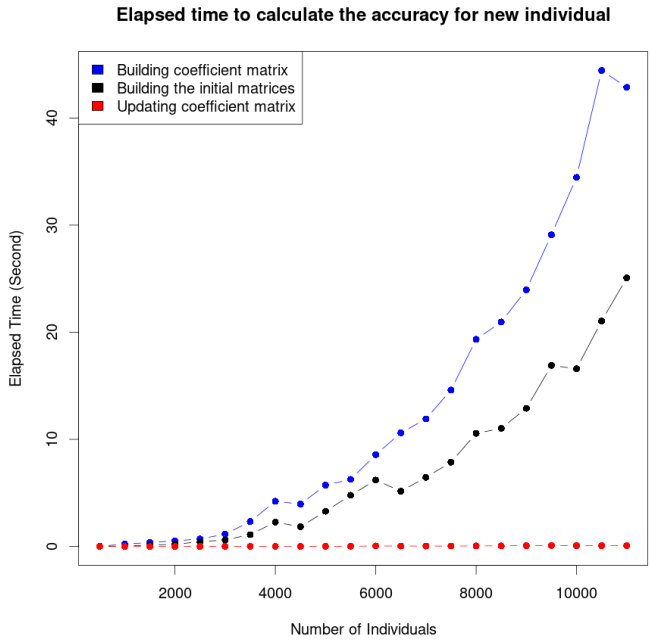

Figure 1. The graph shows the elapsed time required to calculate c^{qq}. c^{qq} is the diagonal of inverted coefficient matrix for an individual without phenotype, Building coefficient matrix: The elapsed time to calculate c^{qq} by rebuilding the coefficient matrix. Building the initial matrices: The elapsed time to build the matrices required to update the c^{qq}. Updating coefficient matrix: The elapsed time to update coefficient matrix and calculate c^{qq}.

Acknowledgments

MHF, NC and BT were supported by Meat and Livestock Australia. The authors wish to thank Cooperative Research Centre for Beef Genetic Technologies (BeefCRC) and Brahman Society for providing real genotypes used in this work.

List of References

Connors, N. K., J. Cook, C. Girard, B. Tier, K. Gore, D. Johnston & M. Ferdosi, 2017. Development of the beef genomic pipeline for breedplan single step evaluation. Australian Association of Animal Breeding and Genetics Proceeding :72.

Harville, D. A., 1997. *Matrix algebra from a statistician's perspective*, volume 1. Springer.

Henderson, C. R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics :423–447.

Lienart, T., 2017. Matrix inversion lemmas. `https://www.stats.ox.ac.uk/~lienart/blog_linalg_invlemmas.html`. Accessed: 20-08-20170.

Meyer, K., B. Tier & H.-U. Graser, 2013. Updating the inverse of the genomic relationship matrix. Journal of animal science 91(6):2583–2586.

VanRaden, P. M., 2008. Efficient methods to compute genomic predictions. Journal of dairy science 91(11):4414–4423.