SOME ALTERNATIVE COMPUTATIONAL STRATEGIES FOR SINGLE-STEP NATIONAL GENOMIC EVALUATIONS

D.J. Garrick[1,2,3], H. Cheng[3], B.L. Golden[2] and R.L. Fernando[3]

[1]Institute of Veterinary, Animal & Biomedical Sciences, Massey University, Hamilton, NZ
[2]ThetaSolutions LLC, Atascadero, California, USA
[3]Department of Animal Science, Iowa State University, Ames, Iowa, USA

SUMMARY

Single-step genomic evaluations combine pedigree and phenotypic information on genotyped and non-genotyped individuals. Such an evaluation can be undertaken using a so-called breeding value model that fits the breeding values of the genotyped and non-genotyped animals (e.g. single-step GBLUP) or using an equivalent so-called marker effects model that directly fits the marker effects. The single-step marker-effects models allow alternatives such as BayesA and mixture models such as BayesB, BayesC or BayesR to be fitted in the context of the single-step analysis. This paper reviews alternative formulations of these equivalent models. The marker-effects formulations of the models are practical options for national genomic evaluations. The most efficient algorithm among those available depends upon the number of marker loci and the numbers of genotyped and non-genotyped animals.

INTRODUCTION

The classical model equation for genetic evaluation using best linear unbiased prediction (BLUP) describes the phenotypes for one or more traits in terms of fixed effects, random additive breeding values, and residual effects that capture that part of the phenotype that cannot be explained by the fixed effects or breeding values (Henderson, 1973). Estimation of breeding values by fitting the mixed linear model typically assumes the pedigree-based additive relationship matrix describes the variance-covariance among breeding values (Henderson, 1973). Henderson (1974) suggested the model equation might be rearranged for computational advantage as explicitly demonstrated in Henderson (1985). That concept was exploited by Quaas and Pollak (1980) in their derivation of the multiple-trait reduced animal model which allowed an animal model to be fitted with little more effort than that for fitting the sire-maternal grandsire models that were commonly used at that time. Nejati-Javeremi *et al.* (1997) showed how to compute a genomic relationship matrix and suggested that be used in place of the additive relationship matrix, a model now known as GBLUP. Meuwissen *et al.* (2001) proposed several models that explicitly fitted haplotype effects rather than breeding values. Those methods varied according to whether the variance ratio for haplotype effects was a known constant (BLUP), an unknown haplotype specific variable (BayesA, BayesB), and whether or not some haplotypes were assumed to have zero effect (BayesB). The breeding value model and BLUP marker effects models were shown to be equivalent (e.g. Stranden and Garrick, 1997). Expanding GBLUP to a more general setting with a model that appropriately accounts for a pedigree including genotyped and non-genotyped animals in a single step was introduced by Legarra *et al.* (2009). That single-step GBLUP (ss-GBLUP) model represented a major advance, and is computationally attractive when there are many more markers than genotyped animals, and all markers are weighted equally to form the genomic relationship matrix. Two marker-effects models are reviewed here which are equivalent to ss-GBLUP and practical for national evaluation. Both allow the model for marker effects to be extended when variance ratios are marker specific and unknown (like BayesA), or follow more general mixture models (BayesB, or BayesR of Erbe *et al.* 2012).

EQUIVALENT MODELS FOR JOINT USE OF GENOTYPED AND NON-GENOTYPED ANIMALS

Single-step GBLUP. Defining a vector of phenotypic records as $\boldsymbol{y_i}$, incidence matrices of fixed effects and breeding values as $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$, vectors of unknown fixed effects ($\boldsymbol{b}$), random effects ($\boldsymbol{u_i}$) and residuals $\boldsymbol{e_i}$, with the subscript *i* denoting *g*=genotyped or *n*=non-genotyped animals, the model equation can be written as

$$\begin{bmatrix}\boldsymbol{y_n}\\ \boldsymbol{y_g}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X_n}\\ \boldsymbol{X_g}\end{bmatrix}\boldsymbol{b} + \begin{bmatrix}\boldsymbol{Z_n} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{Z_g}\end{bmatrix}\begin{bmatrix}\boldsymbol{u_n}\\ \boldsymbol{u_g}\end{bmatrix} + \begin{bmatrix}\boldsymbol{e_n}\\ \boldsymbol{e_g}\end{bmatrix}, \text{ with } var\begin{bmatrix}\boldsymbol{e_n}\\ \boldsymbol{e_g}\end{bmatrix} = \begin{bmatrix}\boldsymbol{R_n} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R_g}\end{bmatrix},$$

and following Legarra *et al.* (2009) with the genetic variance being σ_u^2,

$$\boldsymbol{H} = \frac{1}{\sigma_u^2} var\begin{bmatrix}\boldsymbol{u_n}\\ \boldsymbol{u_g}\end{bmatrix} = \begin{bmatrix}\boldsymbol{A_{nn}} + \boldsymbol{A_{ng}}\boldsymbol{A_{gg}^{-1}}(\boldsymbol{G} - \boldsymbol{A_{gg}})\boldsymbol{A_{gg}^{-1}}\boldsymbol{A_{gn}} & \boldsymbol{A_{ng}}\boldsymbol{A_{gg}^{-1}}\boldsymbol{G}\\ \boldsymbol{G}\boldsymbol{A_{gg}^{-1}}\boldsymbol{A_{gn}} & \boldsymbol{G}\end{bmatrix},$$

which is somewhat formidable. However, Aguilar *et al.* (2010) showed that, for full-rank $\boldsymbol{G}$,

$$\boldsymbol{H^{-1}} = \begin{bmatrix}\boldsymbol{A^{nn}} & \boldsymbol{A^{gn}}\\ \boldsymbol{A^{ng}} & \boldsymbol{A^{gg}} + (\boldsymbol{G^{-1}} - \boldsymbol{A_{gg}^{-1}})\end{bmatrix},$$

which allows existing software used to obtain breeding values in national evaluations using PCG iteration (e.g. Tsuruta *et al.* 2001) to be relatively trivially modified by including an extra step to compute matrix-vector products for the difference matrix $(\boldsymbol{G^{-1}} - \boldsymbol{A_{gg}^{-1}})$. This ss-GBLUP approach was computationally appealing in the early days of genomic prediction, when there were fewer than 40,000 animals genotyped. As the number of genotyped animals increased, the effort to form the dense difference matrix and compute its matrix-vector products increase rapidly. Various strategies to avoid that effort have been devised and implemented, including computing matrix-vector products in parts as $(\boldsymbol{G^{-1}} - \boldsymbol{A_{gg}^{-1}})\boldsymbol{x} = \boldsymbol{G^{-1}x} - \boldsymbol{A_{gg}^{-1}x}$. Using properties of partitioned matrix inverses allows efficient computation of the product $\boldsymbol{A_{gg}^{-1}x}$ without ever forming $\boldsymbol{A_{gg}^{-1}}$ (Masuda *et al.* 2017). An approximation known as APY (Misztal *et al.* 2014) has been used to compute the matrix product $\boldsymbol{G^{-1}x}$. That approximation can in some cases give identical values as for $\boldsymbol{G^{-1}x}$ computed directly, but the lower bounds for APY in general circumstances have not been established.

Single-step GBLUP with marker effects. There are several practical alternatives for fitting the single-step model that do not require $\boldsymbol{G^{-1}}$, nor even require $\boldsymbol{G}$ to be full rank, and these equivalent models have the additional advantage that they can accommodate various priors for marker effects, allowing single-step models for marker effects akin to BayesA, BayesB and BayesR that cannot be fitted using ss-GBLUP.

Liu *et al.* (2014) rearranged the model to include equations for the marker effects, $\boldsymbol{\alpha}$, in addition to the breeding values of genotyped and non-genotyped individuals. An advantage of that representation is that it does not require the matrix $\boldsymbol{G}$, nor its inverse. However, it involves the inverse of the matrix $\boldsymbol{A_{gg}}$, which is dense. A computational strategy was proposed to avoid computing the inverse, but it requires solving a dense system of equations of order equal to the number of non-genotyped animals, and such solution is required every round of PCG or for every Gibbs sample if a model with Bayesian priors for marker effects is to be fitted. We will not consider that representation further.

Hybrid model. Fernando *et al.* (2014) wrote $\boldsymbol{u_g} = \boldsymbol{M_g}\boldsymbol{\alpha}$ as in Meuwissen *et al.* (2001) where $\boldsymbol{M_g}$ are marker covariates observed on genotyped animals, and partitioned $\boldsymbol{u_n}$ into two components, that part of the breeding values of non-genotyped animals that can be explained by the breeding values of genotyped relatives, and an independent part (imputation error, $\boldsymbol{\epsilon}$) not explained by those relatives. That is, $\boldsymbol{u_n} = \boldsymbol{M_n}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, where non-genotyped marker covariates are "imputed" using best linear prediction as $\boldsymbol{M_n} = \boldsymbol{A_{ng}}\boldsymbol{A_{gg}^{-1}}$ which can be obtained efficiently by directly solving the sparse set of equations $\boldsymbol{A^{nn}}\boldsymbol{M_n} = -\boldsymbol{A^{ng}}\boldsymbol{M_g}$ and is easily done in parallel. The resulting "hybrid" model equation is therefore written as

$$\begin{bmatrix}\boldsymbol{y_n}\\ \boldsymbol{y_g}\end{bmatrix}=\begin{bmatrix}\boldsymbol{X_n}\\ \boldsymbol{X_g}\end{bmatrix}\boldsymbol{b}+\begin{bmatrix}\boldsymbol{Z_nM_n} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{Z_gM_g}\end{bmatrix}\boldsymbol{\alpha}+\begin{bmatrix}\boldsymbol{Z_n}\\ \boldsymbol{0}\end{bmatrix}\boldsymbol{\epsilon}+\begin{bmatrix}\boldsymbol{e_n}\\ \boldsymbol{e_g}\end{bmatrix},$$

which is solved by fitting mixed model equations that explicitly include effects for $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon}$. Defining the variance of the vector of marker effects as $\boldsymbol{I}\sigma_\alpha^2$, the inverse variance-covariance matrix for these effects required to form the mixed model equations are

$$var^{-1}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\epsilon}\end{bmatrix}=\begin{bmatrix}\boldsymbol{I}\frac{1}{\sigma_\alpha^2} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{A^{nn}}\frac{1}{\sigma_u^2}\end{bmatrix}.$$

The calculations involving M_n which appears in the off-diagonal of the mixed model equations become formidable when that dense matrix is large, as is the case when there are millions of non-genotyped animals and a large number of markers, but that effort can be reduced when the number of genotyped animals is much less than the number of non-genotyped animals by exploiting the identity $\boldsymbol{A^{nn}M_n}=-\boldsymbol{A^{ng}M_g}$ and storing only $\boldsymbol{M_g}$ as detailed in Fernando *et al.* (2016a). Implementing that approach requires repeated solving of sparse equations of the form $\boldsymbol{A^{nn}x}=\boldsymbol{q}$ within each PCG iteration. This effort is akin to that required to implement the approach of Masuda *et al.* (2017) in ss-GBLUP. If the variance components σ_α^2 or σ_u^2 are assumed not to be known, and/or if mixture priors are to be used for marker effects, this hybrid model can be readily fitted using single-site Gibbs sampling, a model that does not have an equivalent ss-GBLUP form.

Super hybrid model. A further equivalent model involving marker effects can be derived as in Fernando *et al.* (2016b). In circumstances where the number of genotyped animals may be large, perhaps millions, it can be efficiently implemented for national evaluation, especially if there are more genotyped than non-genotyped animals. The model equation is written as

$$\begin{bmatrix}\boldsymbol{y_n}\\ \boldsymbol{y_g}\end{bmatrix}=\begin{bmatrix}\boldsymbol{X_n}\\ \boldsymbol{X_g}\end{bmatrix}\boldsymbol{b}+\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{Z_gM_g}\end{bmatrix}\boldsymbol{\alpha}+\begin{bmatrix}\boldsymbol{Z_n}\\ \boldsymbol{0}\end{bmatrix}\boldsymbol{u_n}+\begin{bmatrix}\boldsymbol{e_n}\\ \boldsymbol{e_g}\end{bmatrix},$$

which is solved by fitting a mixed model involving $\boldsymbol{\alpha}$, as in the hybrid model, along with $\boldsymbol{u_n}$ as in ss-GBLUP. We refer to this model here as the super-hybrid model. The inverse variance-covariance matrix for the fitted effects is given by

$$var^{-1}\begin{bmatrix}\boldsymbol{\alpha}\\ \boldsymbol{u_n}\end{bmatrix}=\begin{bmatrix}\boldsymbol{I}\frac{1}{\sigma_\alpha^2}+\boldsymbol{M_n}'\boldsymbol{A^{nn}M_n}\frac{1}{\sigma_u^2} & \boldsymbol{M_g}'\boldsymbol{A^{gn}}\frac{1}{\sigma_u^2}\\ \boldsymbol{A^{ng}M_g}\frac{1}{\sigma_u^2} & \boldsymbol{A^{nn}}\frac{1}{\sigma_u^2}\end{bmatrix},$$

which only involves the matrix of imputed marker genotypes $\boldsymbol{M_n}$ in a quadratic form on the diagonal, and that symmetric matrix has order equal to the number of markers which in national evaluations can nowadays be an order of magnitude less than the number of genotyped individuals. Comparison of the computing effort in this super-hybrid model relative to the hybrid model for a national cattle evaluation dataset is in Fernando *et al.* (2016b).

All of these equivalent models, namely ss-GBLUP which fits breeding values for non-genotyped and genotyped animals, the ss-GBLUP model with breeding values and marker effects, the hybrid model which fits marker effects and imputation residuals for non-genotyped animals, and the super-hybrid model which fits marker effects and breeding values for non-genotyped animals, can be extended to more complex forms of models. These include those that fit additional polygenic effects not captured by markers, those that fit maternal genetic and maternal permanent environmental effects, and those accommodating multiple traits, those with repeated measures, those including random regression polynomials, those with heterogeneous variances, in addition to breed, heterosis and group effects as required in multi-breed analyses.

DISCUSSION

The two marker effects models reviewed here are equivalent to ss-GBLUP when the genomic relationship matrix is full rank and the variance parameters are known. These marker-effects models may require greater computational effort than ss-GBLUP when the number of genotyped animals is small. The relative effort for the hybrid model that fits marker effects and imputation residuals for non-genotyped animals, compared to the super-hybrid model that fits marker effects and breeding values for non-genotyped animals, varies according to the number of markers and numbers of genotyped and non-genotyped animals. For analyses involving millions of genotyped animals, one or other or both of the marker effects models will be more efficient than ss-GBLUP. Implemented in a Gibbs sampler, these models can readily accommodate alternative priors including those representing mixture distributions, which in some situations leads to higher accuracy of prediction than ss-GBLUP (Lee et al. 2017). Furthermore, using Gibbs sampling will provide samples from the relevant posterior distributions which can be used to provide estimates of the prediction error variances and prediction error covariances, as well as the posterior means that represent the estimates of the breeding values. Both of these marker effects models have been prototyped in multi-breed multiple-trait national evaluations including maternal effects. The super-hybrid model is currently being implemented in the Pan-American Cattle Evaluation (PACE) run by ABRI for Hereford cattle, and in the North American multi-breed analysis run by International Genetic Solutions (IGS) which is the largest North American evaluation in terms of pedigree size. The super-hybrid model is also being used by global companies for pig, chicken and dairy cattle evaluation.

CONFLICT OF INTEREST

DJG and BLG are co-founders of Theta Solutions LLC, a company that licenses BOLT software which is capable of fitting all the models described in this paper.

REFERENCES

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S. and Lawlor T. (2010) *J. Dairy Sci.* **93**:743.

Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., Bowman P.J., Reich C.M., Mason B.A. and Goddard M.E. (2012) *J. Dairy Sci.* **95**:4114.

Fernando R.L., Cheng H., Golden B.L. and Garrick D.J. (2016a) *Genet. Sel. Evol.* **48**:96.

Fernando R.L., Cheng H. and Garrick D.J. (2016b) *Genet. Sel. Evol.* **48**:80.

Fernando R.L., Dekkers J.C.M. and Garrick D.J. (2014) *Genet. Sel. Evol.* **46**:50.

Henderson C.R. (1973) In *Proc of the Animal Breeding and Genetics Seminar in Honor of J.L. Lush. ASAS and ADSA Champaign, IL*

Henderson C.R. (1974) *J. Dairy Sci.* **57**:963.

Henderson C.R. (1985) *J. Dairy Sci.* **68**:2267.

Lee J., Cheng H., Garrick D., Golden B., Dekkers J., Park K., Lee D. and Fernando R. (2017) *Genet. Sel. Evol.* **49**:2.

Legarra A., Aguilar I. and Misztal I. (2009) *J. Dairy Sci.* **92**:4656.

Liu Z., Goddard M.E., Reinhardt F., and Reents R. (2014) *J. Dairy Sci.* **97**:5833.

Masuda Y., Misztal I., Legarra A., Tsuruta S., Laurenco D.A.L., Fragomeni B.O. and Aguilar I. (2017) *J. Animal Sci.* **95**:49.

Meuwissen T.H.E., Hayes B.J. and Goddard M.E. (2001) Genetics **157**:1819.

Misztal I., Legarra A. and Aguilar I. (2014)) *J. Dairy Sci.* **97**:3943.

Nejati-Javaremi A., Smith C. and Gibson J.P. (1997) *J. Animal Sci.* **75**:1738.

Quaas R.L., and Pollak E.J. (1980) *J. Animal Sci.* **51**:1277.

Stranden I. and Garrick D.J. (1997) *J. Dairy Sci.* **92**:2971.

Tsuruta S., Misztal I. and Stranden I. (2001) *J. Animal Sci.* **79**:1166.