Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Bayesian prediction combining genotyped and non-genotyped individuals

D.J. Garrick[1], J.C.M. Dekkers[1], B.L. Golden[2], R.L. Fernando[1]
[1]Iowa State University, Ames, Iowa, USA [2]Calpoly, San Luis Obispo, California, USA

ABSTRACT: Conventional pedigree- and performance-based national evaluations typically involve hundreds of thousands if not millions of animals. But only a small proportion of individuals with performance records have typically been genotyped to date. Bayesian methods have been widely adopted for analysis of these genotyped individuals, but implementation typically involves two-step approaches to blend genomic predictions on genotyped individuals with information from conventional analyses for non genotyped animals. Here we present a Bayesian approach that extends commonly-used methods including BayesA, BayesB, BayesC, and BayesCπ, to a single step method using observations from all genotyped and non genotyped individuals. Unlike single-step GBLUP, our approach does not require direct inversion of any matrices and is well suited to parallel computing approaches.

Keywords: Genomic prediction

Introduction

Genomic prediction uses marker genotypes to improve the accuracy of prediction. This contrasts with conventional approaches for prediction of breeding merit that only utilize pedigree and performance information. Pedigree and performance-based approaches have evolved over the last half century and typically exploit Henderson's mixed model equations (MME) (Henderson, 1984), sparse matrix computations, direct construction of the inverse relationship matrix (Henderson, 1976; Quaas, 1976), iteration on data (Schaeffer and Kennedy, 1986), preconditioned conjugate gradient (PCG) (Berger et al, 1989; Stranden and Lidauer, 1999), and approximation of diagonal elements of the inverse coefficient matrix (Harris and Johnson, 1998) to estimate prediction error variance.

Genomic prediction simultaneously fits many markers and was pioneered by Meuwissen et al. (2001) to exploit marker genotypes and individual performance information without requiring or using pedigree information. That paper marked the start of widespread adoption of Bayesian approaches that utilize Markov chain Monte Carlo (MCMC) techniques, including Gibbs sampling and Metropolis-Hastings algorithms (e.g. Fernando and Garrick, 2013). Genomic prediction could be readily extended to exploit performance information on non genotyped relatives by using deregressed breeding values in weighted analyses in place of individual information (Garrick et al., 2009), or in the special case of non genotyped offspring of genotyped parents, using reduced animal model approaches as in Wolc et al. (2012).

Routine application of genomic prediction should use all available pedigree and performance information. Van Raden et al. (2009) devised a selection index approach to combine genomic predictions (DGV) and conventional pedigree predictions (EBV) to obtain genomic enhanced breeding values (GEBV) in a two-step analysis. A single-step analysis would be operationally advantageous and avoid the need for simplifying assumptions. A single-step analysis that did not use Bayes theorem and was based on a model that fitted breeding values was proposed by Misztal et al. (2009) but required brute-force inversion of a matrix of order equal to the number of animals with performance information. Legarra et al. (2009) proposed a modification to the assumed variance-covariance among non genotyped animals and between genotyped and non genotyped animals based on genomic relationships among genotyped relatives. Aguilar et al. (2010) recognized that the Legarra et al. (2009) variance-covariance matrix could exploit the pedigree-based inverse relationship matrix, allowing the brute-force matrix inversion to be reduced to two smaller matrices of order equal to the number of genotyped individuals. Some practical problems with that implementation remain, and a strategy to overcome these based on a Bayesian framework applied to the model proposed in Legarra et al. (2009) are the subject of this paper. This strategy does not require any brute-force inversion, extends from single-step GBLUP to the entire family of Bayesian models for genomic prediction, and exposes a solution to the problem of choosing among alternative approaches to center the genomic relationship matrix.

Model

The derivation of the computing strategy begins with definition of the model. The usual model equation (e.g. Henderson, 1984) would be:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is a vector of phenotypic observations, $\boldsymbol{b}$ is a vector of unknown fixed effects, $\boldsymbol{u}$ is a random vector of breeding values, $\boldsymbol{e}$ is a vector of random residual effects, and $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices defining the particular fixed effects and breeding values that pertain to the phenotypic observations. In the simplest single trait model with homogeneous genetic and residual variances and uncorrelated residual effects, $\text{var}(\boldsymbol{u})=\boldsymbol{A}\sigma_{\boldsymbol{u}}^{\boldsymbol{2}}$, and $\text{var}(\boldsymbol{e})=\boldsymbol{I}\sigma_{\boldsymbol{e}}^{\boldsymbol{2}}$, where $\boldsymbol{A}$ is the additive or numerator relationship matrix.

Now partition the vector of breeding values $\boldsymbol{u}' = [\boldsymbol{u}_{\boldsymbol{g}}' \quad \boldsymbol{u}_{\boldsymbol{n}}']$, into genotyped animals (subscript *g*) and non genotyped animals (subscript *n*), and recognize that all other model elements can be similarly partitioned.

The model for genotyped animals can be any of BayesA (Meuwissen et al., 2001), BayesB (Meuwissen et al., 2001), BayesC (Kizilkaya et al., 2010), BayesCπ (Habier et al., 2011), etc. Those models, following Falconer and Mackay (1996), define the breeding value as the sum of average effects of alleles, which in matrix notation is equivalent to $\boldsymbol{u_g = M_g \alpha}$, where $\boldsymbol{M_g}$ is the matrix of marker genotypes observed on the genotyped individuals and $\boldsymbol{\alpha}$ is the vector of allele substitution effects. Substituting this into the usual model equation gives:

$$\boldsymbol{y_g = X_g b + Z_g M_g \alpha + e_g}, \qquad [1]$$

where premultiplication of the genotype matrix by $\boldsymbol{Z_g}$ is required because some genotyped individuals may not have a phenotypic observation.

Specifying the model for non genotyped animals involves recognizing that the breeding value for non genotyped animals can be decomposed into two orthogonal components. This same concept was exploited by Quaas and Pollak (1980) in their derivation of the reduced animal model, whereby the breeding value for non parents was partitioned into a part that could be predicted from ancestral relatives, namely the parent average, and a part independent of ancestors, namely the Mendelian sampling component. In our case, the part of the breeding value of non genotyped animals that can be predicted from the breeding values of all their genotyped relatives, including both ancestors and descendants, will be denoted by the vector $\boldsymbol{\widehat{u_n}/u_g}$, representing $\boldsymbol{\widehat{u_n}}$ given $\boldsymbol{u_g}$. The part of the breeding value of non genotyped animals that cannot be explained by relatives is the residual polygenic effect or deviation of the breeding values $\boldsymbol{u_n}$ from $\boldsymbol{\widehat{u_n}/u_g}$, which we will denote as $\boldsymbol{\epsilon_n = (u_n - \widehat{u_n}/u_g)}$, so that

$$\boldsymbol{u_n = \widehat{u_n}/u_g + \epsilon_n}. \qquad [2]$$

Now $\boldsymbol{\widehat{u_n}/u_g}$ is the matrix regression of $\boldsymbol{u_n}$ on $\boldsymbol{u_g}$, namely

$$\boldsymbol{\widehat{u_n}/u_g = A_{ng} A_{gg}^{-1} u_g}. \qquad [3]$$

Substituting [2], [3], and $\boldsymbol{u_g = M_g \alpha}$ in the usual model equation for non genotyped individuals gives

$$\begin{aligned} \boldsymbol{y_n} &= \boldsymbol{X_n b + Z_n u_n + e_n} \\ &= \boldsymbol{X_n b + Z_n A_{ng} A_{gg}^{-1} u_g + Z_n \epsilon_n + e_n} \\ &= \boldsymbol{X_n b + Z_n \left(A_{ng} A_{gg}^{-1} M_g\right) \alpha + Z_n \epsilon_n + e_n} \\ &= \boldsymbol{X_n b + Z_n M_n \alpha + Z_n \epsilon_n + e_n}, \end{aligned} \qquad [4]$$

where $\boldsymbol{M_n = A_{ng} A_{gg}^{-1} M_g}$ is the matrix of genotypes on non genotyped animals, "imputed" by regression of non genotyped animals on their genotyped relatives. For the remainder of this paper we will refer to this regression as imputation, but it must be recognized our derivation will not be exact if other methods of imputation are used.

This formulation in [4] suggests a computational approach that involves fitting a marker effects model for all animals, whether genotyped or not. First, matrix $\boldsymbol{M_n}$ must be imputed based on pedigree information and $\boldsymbol{M_g}$. This can be done separately for each locus, i.e. imputing one column of $\boldsymbol{M_n}$ from one column of $\boldsymbol{M_g}$, which is perfectly suited to parallel computing. This calculation does not require creation or storage of either $\boldsymbol{A_{ng}}$ or $\boldsymbol{A_{gg}^{-1}}$, as the i-th column $\boldsymbol{m_{i\_n}}$ of $\boldsymbol{M_n}$ can be obtained by solving $\boldsymbol{A^{ng} m_{i\_n} = -A^{ng} m_{i\_g}}$, where $\boldsymbol{m_{i\_g}}$ is the i-th column of $\boldsymbol{M_g}$. These equations only involve sparse submatrices of $\boldsymbol{A^{-1}}$.

The mixed model equations for simultaneously obtaining solutions for $\boldsymbol{\alpha}$ and $\boldsymbol{\epsilon_n}$ are

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'ZM} & \boldsymbol{X_n' Z_n} \\ \boldsymbol{M'Z'X} & \boldsymbol{M'Z'ZM + \varphi} & \boldsymbol{M_n' Z_n' Z_n} \\ \boldsymbol{Z_n' X_n} & \boldsymbol{Z_n' Z_n M_n} & \boldsymbol{Z_n' Z_n + A^{nn} \lambda} \end{bmatrix} \begin{bmatrix} \boldsymbol{\hat{b}} \\ \boldsymbol{\hat{\alpha}} \\ \boldsymbol{\widehat{\epsilon_n}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \\ \boldsymbol{M_n' y_n} \end{bmatrix} \qquad [5]$$

where $\boldsymbol{X}$, $\boldsymbol{Z}$, $\boldsymbol{M}$ and $\boldsymbol{y}$ contain the partitioned submatrices or vector for both genotyped and non genotyped individuals, λ is the scalar ratio σ_e^2/σ_u^2 and $\boldsymbol{\varphi}$ is a diagonal matrix whose elements vary according to the nature of the model assumed for marker effects; in BayesC, the elements are all $\sigma_e^2/\sigma_\alpha^2$, whereas in models like BayesA, the elements are locus specific. After solving the effects in [5], the EBV for all genotyped animals are $\boldsymbol{\widehat{u_g} = M_g \hat{\alpha}}$ and the EBV for non genotyped animals are $\boldsymbol{\widehat{u_n} = M_n \hat{\alpha} + \widehat{\epsilon_n}}$.

Inspection of the equations in [5] shows that when all animals are genotyped, the third row and third column can be deleted, leaving equations identical to those that form the basis of commonly used Bayesian approaches such as BayesA, BayesB and BayesC. When no animals are genotyped, the second row and second column can be deleted, leaving equations identical to those used in conventional genetic evaluation that incorporates only pedigree and performance information.

Matrix $\boldsymbol{M}$ is dense and will be a large matrix to store for a population of many individuals, but it will be no larger than would be the case if the entire population was genotyped. Matrix $\boldsymbol{M'Z'ZM}$ in the coefficient matrix of [5] is also dense, but its order is limited to the number of loci in $\boldsymbol{M}$.

One approach to solving the equations in [5] is to absorb the $\boldsymbol{\epsilon_n}$ equations, which can be done provided λ is known, sample or solve the marker effects, and then sample or solve $\boldsymbol{\epsilon_n}$ conditional on the marker effect MCMC samples themselves or their posterior means. This approach reduces memory requirements, as $\boldsymbol{M}$ need not be stored in its entirety, and separate solution of marker effects and $\boldsymbol{\epsilon_n}$ can be done in parallel. If this was done in a Bayesian approach using Gibbs sampling and assuming λ is unknown, the absorption would have to be repeated for each sample value of λ.

The equations could be solved conventionally if all variance parameters were known, for example using PCG, or solutions could be obtained using a Bayesian MCMC approach, which would allow simultaneous estimation of all variance components, and would allow mixture models (e.g. BayesB, BayesC) for the locus effects in $\boldsymbol{\alpha}$. The MCMC approaches include single site or block Gibbs strategies. An advantage of MCMC approaches is that the entire posterior distributions of effects can be obtained, which allows direct computation of prediction error variances (or reliabilities) for solutions or for linear functions of solutions, such as EBV.

In the special case where diagonal elements of $\boldsymbol{\varphi}$ are all $\sigma_e^2/\sigma_\alpha^2$, and $\sigma_\alpha^2 = \sigma_u^2/k2\overline{pq}$, where k is the number of marker loci in $\boldsymbol{M}$, $\bar{p}$ and $\bar{q}$ are the mean frequencies of the alternate alleles, and with σ_e^2 and σ_u^2 known, then the

resulting EBV are identical to those obtained from single-step GBLUP (Fernando et al., 2014), provided $\boldsymbol{M_g M_g'}$ is full rank, except that no brute-force matrix inversions are required. If there are more animals than markers, or if $\boldsymbol{M_g}$ is centered using mean allele frequencies in genotyped individuals, then $\boldsymbol{M_g}$ does not have full row rank and some approximation is required to use the computing approach of Aguilar et al. (2010). In contrast to the approach of Aguilar et al. (2010), the computing effort in [5] is reduced rather than increased as more animals are genotyped.

In conventional evaluation without group effects, a base population of founders, comprising animals with unknown parents, define the individuals for which variance components are estimated. Thus, marker covariates must be defined in relation to these founder animals in order to obtain consistency between the base defined by the numerator and the genomic relationship matrices. However, genotypes are typically not available on the founder animals and the marker covariates are centered using mean allele frequencies in the genotyped individuals.

In Bayesian regression models with all animals genotyped, the marker effect solutions are invariant to the approach used to center the marker covariates. This is easily demonstrated as follows. Centering involves subtracting some constant, say *c*, from elements of a particular column of $\boldsymbol{M_g}$. In matrix notation, a vector $\boldsymbol{c}$ comprises all these constants and the centered covariate matrix $\boldsymbol{M_g^c = M_g - 1c'}$, where $\mathbf{1}$ is a vector of 1's. The effect of using $\boldsymbol{M_g^c}$ in place of $\boldsymbol{M_g}$ can be seen by starting with model equation [1], simplified to comprise only an overall mean as fixed effect, then adding and subtracting $\boldsymbol{Z_g 1 c'\alpha}$, and defining $t = \boldsymbol{c'\alpha}$ to give

$$\begin{aligned}\boldsymbol{y_g} &= \mathbf{1}\mu + \boldsymbol{Z_g M_g \alpha} + \boldsymbol{e_g},\\ &= \mathbf{1}\mu + \boldsymbol{Z_g}\mathbf{1}(\boldsymbol{c'\alpha}) + \boldsymbol{Z_g}\left(\boldsymbol{M_g} - \mathbf{1}\boldsymbol{c'}\right)\boldsymbol{\alpha} + \boldsymbol{e_g}\\ &= \mathbf{1}(\mu + t) + \boldsymbol{Z_g}\left(\boldsymbol{M_g} - \mathbf{1}\boldsymbol{c'}\right)\boldsymbol{\alpha} + \boldsymbol{e_g}\\ &= \mathbf{1}\mu^{*} + \boldsymbol{Z_g M_g^c \alpha} + \boldsymbol{e_g}. \qquad [6]\end{aligned}$$

Model equation [6] does not have the same first and second moments as [1], so these two models are not equivalent (Henderson, 1984). However, these two models give identical rankings of EBV, provided a fixed general mean is included in the models (Stranden and Christensen, 2011).

Now consider the implication of using $\boldsymbol{M_g^c}$ in place of $\boldsymbol{M_g}$ in [4], the model equation for non genotyped animals. Adding and subtracting a vector in order to substitute $\boldsymbol{M_g}$ for $\boldsymbol{M_g - 1c'}$ results in

$$\begin{aligned}\boldsymbol{y_n} &= \mathbf{1}\mu + \boldsymbol{Z_n}\left(\boldsymbol{A_{ng}A_{gg}^{-1}M_g}\right)\boldsymbol{\alpha} + \boldsymbol{Z_n\epsilon_n} + \boldsymbol{e_n}\\ &= \mathbf{1}\mu + \boldsymbol{Z_n A_{ng} A_{gg}^{-1}}\mathbf{1}(\boldsymbol{c'\alpha}) + \boldsymbol{Z_n}\left(\boldsymbol{A_{ng}A_{gg}^{-1}}\left(\boldsymbol{M_g} - \mathbf{1}\boldsymbol{c'}\right)\right)\boldsymbol{\alpha} + \boldsymbol{Z_n\epsilon_n} + \boldsymbol{e_n}\\ &= \mathbf{1}\mu + \boldsymbol{Z_n A_{ng} A_{gg}^{-1}}\mathbf{1}t + \boldsymbol{Z_n}\left(\boldsymbol{A_{ng}A_{gg}^{-1}M_g^c}\right)\boldsymbol{\alpha} + \boldsymbol{Z_n\epsilon_n} + \boldsymbol{e_n},\end{aligned}$$

which shows that the modification to the first moments is no longer the scalar $\mathbf{1}t$, but the covariate $\boldsymbol{A_{ng}A_{gg}^{-1}}\mathbf{1}t$, which can take on different values for each non genotyped animal, depending upon how closely it is related to the genotyped animals. If the solution for *t* is 0, then the centering modification to $\boldsymbol{M_g}$ will make no difference, but if $t = \boldsymbol{c'\alpha}$ has a nonzero value, then the model using $\boldsymbol{M_g}$ and the model using $\boldsymbol{M_g^c}$ could give very different results. Rather than trying to find the appropriate values to center the covariate matrix, we propose fitting an extra covariate as a fixed effect to account for the possibility that the bases of the genotyped and non genotyped animals are different (Fernando et al., 2014).

Overall Conclusions

Genomic prediction is an immature but rapidly developing technology. In concert with developments in high-density marker genotyping, there have been increased implementation of MCMC methods as computing strategies, and developments in parallel computing to exploit multi-core processors. Collectively, these developments will facilitate practical implementations of Bayesian prediction methods that combine all information from genotyped and non genotyped individuals.

Literature Cited

Aguilar, I., Misztal, I., Johnson, D.L. et al. (2010). *J Dairy Sci*, 93:743-752.

Berger, P.J., Luecke, G.R., Hoekstra, J.A. (1989). *J Dairy Sci*, 72:514-522.

Falconer, D.S., Mackay, T.F.C. (1996). Prentice-Hall.

Fernando, R.L., Dekkers, J.C.M, Garrick, D.J. (2014). *Genet Sol Evol*, (under review).

Fernando, R.L., Garrick, D.J. (2013). Berlin, Springer Series: Methods in molecular biology.

Garrick, D.J., Taylor, J.F, Fernando, R.L. (2009). *Genet Sel Evol*, 41:55.

Habier, D., Fernando, R.L., Kizilkaya, K., Garrick, D.J. (2011). *BMC Bioinformatics*, 12:186.

Harris, B.L., Johnson, D.L. (1998). *J Dairy Sci*, 81:2723-2728.

Henderson, C.R. (1976). *Biometrics*, 32:69-83.

Henderson, C.R. (1984). University of Guelph.

Kizilkaya, K., Fernando, R.L., Garrick, D.J. (2010). *J Anim Sci*, 88:544-551.

Legarra, A., Aguilar, I., Misztal, I. (2009). *J Dairy Sci*, 92:4656-4663.

Meuwissen, T.H.E, Hayes, B.J., Goddard, M.E. (2001). *Genetics*, 157:1819-1829.

Misztal, I., Legarra, A., Aguilar, I. (2009). *J Dairy Sci*, 92:4648-4655.

Quaas, R.L. (1976). *Biometrics*, 32:949-953.

Quaas, R.L., and Pollak, E.J. (1980). *J Anim Sci*, 51:1277-1287.

Schaeffer, L.R., Kennedy, B.W. (1986). *J Dairy Sci*, 69:575-579.

Stranden, I., Christensen, O.F. (2011). *Genet Sel Evol*, 43:25.

Stranden, I, Lidauer, M. (1999) *J Dairy Sci*, 82:2779-2787.

Van Raden, P.M., Van Tassell, C.P., Wiggans, G.R. et al. (2009). *J Dairy Sci*, 92:16-24.

Wolc, A., Arango, J., Settar, P. et al. (2012). *Animal Genetics*, 43:87-96.