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

Genomic breeding values for un-genotyped individuals

B. Tier.
Animal Genetics and Breeding Unit, University of New England, Armidale NSW 2351, Australia. (AGBU is a joint unit of NSW Department of Primary Industries and the University of New England)

ABSTRACT: Genomic information is now commonly used in routine genetic evaluations. This is usually in the form of genomic breeding values (GBVs) which have a high heritability but are generally confined to those animals with genotypes. This can lead to anomalies when parents have GBVs and progeny do not. By using a single-trait genetic evaluation, GBVs can be generated for related individuals. It is most efficient to do this for genotyped individuals and their ancestors initially, and calculate mid-parent values for all other individuals. A method for approximating accuracies for the relatives' GBVs is described.
Keywords: genomic predictions; blending

Introduction

Genomic prediction equations have been developed for many traits and species from reference data sets where animals have both genotypes and phenotypes. These equations are used to generate genomic breeding values (GBVs) for young animals. It is common practice for GBVs to be combined with estimated breeding values (EBVs) produced from routine genetic evaluations based on Best Linear Unbiased prediction (BLUP, Henderson, 1984) using phenotypes and pedigrees. The resulting genomically enhanced estimated breeding values (GEBVs) are more accurate, allow selection to be more intense and are obtained earlier in life than conventional EBVs. The additional value provided by a GBV reduces as the accuracy of the EBV increases.

This process has been highly successful in the dairy industry where reference data generally includes bulls with highly accurate EBVs and the sires and maternal grand-sires of the predicted animals are included in the reference dataset. However, it has been less successful in the more extensive industries of beef cattle and sheep where there can be many breeds, many more male parents are used, recording is less intensive, less use is made of artificial insemination and consequently there are fewer bulls with highly accurate EBVs for most traits of interest. Furthermore, in these industries this process is of less value. Unlike the dairy industry, animals in the extensive industries can be recorded for many traits before selections are made and so EBVs for young beef cattle and sheep can have greater accuracy than their dairy counterparts. Unfortunately, the accuracy of genomic predictions reduces with increasing genetic distance of the predicted animals from the reference data (Habier *et al*., 2007).

Blending has generally been limited to animals that have been genotyped. This has the effect of limiting the effectiveness of genotyping as un-genotyped individuals receive no benefit from the GBVs of their relatives in their EBVs. This can produce anomalies when, for example, the effect of a sire's GBV is not inherited by any of its progeny.

In the dairy sector, GBVs have generally been developed in-house by the organizations responsible for the genetic evaluation. In the beef sector GBVs have been developed and are routinely produced by at least four separate organizations. Generally GBVs are provided by these groups, but not the genotypes used in their prediction. Consequently an approach based on exploiting the genotypes, such as the single step (Misztal *et al.*, 2009), is not available.

This paper describes a simple method for calculating GBVs for un-genotyped animals using the GBVs of the genotyped animals and the pedigree. A method for approximating the accuracy of these GBVs is also described.

Materials and Methods

GBVs are highly accurate measures of part of the breeding value (BV) of an animal and selection based solely on GBVs has no effect on the other part – the variance of the GBVs is identical to the covariance between GBVs and EBVs.

The heritability of GBVs is very high (~1), as they are generated as a linear function of an animal's genotypes (**g**) and their estimated effects (**ŝ**), (GBV=**g**'**ŝ**). Thus the corresponding prediction error variance for GBVs can be considered very low. Furthermore, as the heritability is so high, genotyped individuals receive no benefit from repeated measurements or observations on relatives.

Extending GBVs. GBVs for un-genotyped animals can be calculated readily using a single trait BLUP evaluation with a high heritability. In large populations this process can be done efficiently in two parts. First, animals with GBVs and all their ancestors are analyzed in a single BLUP, one trait at a time with a model

$$\mathbf{GBV} = \text{mean} + \mathbf{u} + \mathbf{e},$$

where **u** is a vector of breeding values, **e** a vector of residuals, $Var(e)= \mathbf{I}\sigma_e^2$ and $Var(u)=\mathbf{A}\sigma_u^2$, and σ_e^2 and σ_u^2 are the residual and genetic variances and **A** is the numerator rela-

tionship matrix. As the heritability is almost 1, σ_e^2 is a very small number relative to σ_u^2.

Subsequently the GBVs calculated in the first part are used to generate mid-parent values by proceeding down the pedigree applying the equation:

$$u_a=(u_s+u_d))/2,$$

where u_i is the breeding value of the animal (a), sire(s) or dam(d)) for the rest of the population. These are subsequently called extended GBVs (XGBVs).

Approximating accuracies. The accuracy of an EBV is given by the equation: $\mathrm{Acc} = \sqrt{(1 - \mathrm{PEV(u)}/\mathrm{Var(u)})}$, where PEV is the prediction error variance of the EBV given by the diagonal element in the inverse of the coefficient matrix of the mixed model equations for that EBV. As inverting the coefficient matrix (**C**) for most large populations is generally infeasible, accuracies are commonly approximated. Ascertaining the accuracy of the XGBVs can use the population partitioned into the same two subsets as for the BLUP evaluation. Accuracies for ancestors can be approximated using effective numbers of progeny (EPN) as a measurement (Graser and Tier, 1997), and for the remaining population can be calculated directly from their parents.

Ancestors. One method for approximating the accuracy of an EBV (Graser and Tier, 1997) first calculates EPN from all correlated observations. Subsequently, the accuracy is calculated as:

$$\mathrm{Acc} = \sqrt{\mathrm{EPN}/(\mathrm{EPN} + \lambda)}$$

where $\lambda = (4 - h^2)/h^2$. As the heritability of the GBVs is almost 1, λ can be set to 3. For ease of calculation, the genetic variance of the GBVs is 1, and the residual variance as $\sim 10^{-6}$, but any small amount will do.

The accuracy of a GBV for a genotyped individual can be considered as 1, and its PEV 0. As the GBV is a direct measure there is no need to consider any loss of degrees of freedom when accumulating EPN. However, it is necessary to consider the accuracy of each un-genotyped progeny when accumulating EPN for un-genotyped ancestors, as some intermediate ancestors may not have been genotyped. The effective value of these progeny is not 1, as their accuracy is not 1. Their value as EPN is equal to their reliability:

$$\mathrm{EPN}=\mathrm{Acc}^2.$$

Contributions to ancestors are accumulated from descendants in order from youngest to oldest. Contributions from parents to an offspring, in EPN, can be calculated by first computing the accuracy of each parent in the absence of the contribution from that offspring. The accuracy of the offspring is then calculated using the equation for the accuracy of a mid-parent value:

$$\mathrm{Acc_{progeny}} = 0.5\sqrt{\mathrm{Acc_{Sire}^2} + \mathrm{Acc_{dam}^2}},$$

and the EPN for this offspring, resulting from its parents is

$$\mathrm{EPN_{parents}}=\lambda*\mathrm{Acc}^2/(1-\mathrm{Acc}^2).$$

Once the EPN from all sources has been accumulated the accuracy for each individual in this set can be determined.

Descendents. The accuracy for all other animals in the population can be approximated using the equation for calculating the accuracy of the mid-parent value (above).

Example 1. Consider the pedigree in in Figure 1, where individuals D, E and F are genotyped. Their accuracies are 1, and each is one effective progeny for their parents. Had B and C not been related, then their accuracies would have been 0.63 (=$\sqrt{(2/5)}$) and 0.50 (=$\sqrt{(1/4)}$). However, C (B) contributes 0.4 (0.25) of an EPN to A. Thus A has an accuracy of 0.42 (=$\sqrt{(0.65/3.65)}$). The value of A to B results from information from C and its descendants. The accuracy of A from C only is 0.28 (=$\sqrt{(0.25/3.25)}$), and this is worth 1/13 EPN to B, and thus B's accuracy given all the data is approximately 0.64. Similarly B has an approximate accuracy of 0.52. These values are shown in Table 1 alongside the actual accuracies calculated from C^{-1}.

Figure 1: Example Pedigree

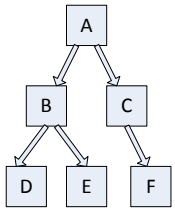

Table 1: Effective progeny numbers (EPN), Accuracies (Acc) and partial accuracies (Acc*) from descendants (desc) and all relatives without contributions from self (all).

	EPN (desc)	Acc* (desc)	Acc* of individual A without self	EPN (all)	Acc (all)	Acc from C^{-1}
A	0.65[+]	0.42	n.a.	0.65	0.42	0.39
B	2	0.63	0.28	2.08	0.64	0.64
C	1	0.50	0.34	1.12	0.52	0.51

[+]=0.4+0.25

Example 2. Populations of 10 sires and 100 dams were simulated for 10 generations. In each generation, half the sires and one third of the dams were replaced randomly. Dams mated randomly to sires and all dams had a single offspring. GBVs for all animals were simulated using 1000 SNP with 'effects' drawn randomly from a normal distribution. Parents of individuals born in the last two years were 'genotyped' and their GBVs calculated from the SNP 'effects'. XGBVs and their approximate accuracies were calculated using the method described above. Correlations

between GBVs and XGBVs were calculated for each of three un-genotyped groups – ancestors, young animals and others – for each of 10 replicates. Accuracies were calculated using $\mathbf{C}^{-1}$ and approximated using the method described. Correlations between true and approximate accuracies were calculated for the same groups. The averages of the correlations were used to estimate the accuracy of the method.

Blending. In its simplest form blending of a GBV with an EBV is based on the selection index. More complex forms have also been developed for cases where the data used to generate the GBV are not independent of those used to generate the EBV. For the extensive industries in Australia, GBVs have generally been developed from data that is independent from the national recording schemes. For such cases the variance a GBV or XGBV is the same as the covariance between it and the EBV and is given by the product of its reliability with the variance of the XGBVs.

Once the XGBVs and their accuracies have been calculated they can be blended with the EBVs. Note that when there are multiple traits with GBVs, and all animals have all the GBVs, then the accuracies for the XGBVs need only be determined once, as all the GBVs have the same heritability. That the XGBVs themselves are of different value for different traits is accommodated during blending and is defined by their covariance with the EBV.

Results and Discussion

A simple method for extending GBVs to un-genotyped individuals, where there genotypes are unavailable, has been presented. Calculation of XGBVs is straightforward as is the algorithm which approximates their accuracies.

Approximate accuracies were very close to the expected accuracies for the example population (Table 1). This was also the case for any subset of the simulated population in example 2 where correlations between approximate and 'true' accuracies were 0.99, althought the approximations were slightly higher (0.00-0.06) than the exact accuracies for all groups. Correlations between XGBVs and 'true' GBVs averaged at 0.36 for ancestors, 0.63 for progeny and 0.31 for other animals and were slightly less than their average accuracies of 0.48, 0.71 and 0.44, respectively.

An alternative method to determine the EBVs for ancestors could simply have been based on the selection index $\mathbf{u_x} = \mathbf{A_{xg}A_{gg}^{-1}u_g}$, where XGBVs ($\mathbf{u_x}$) are computed directly from the numerator relationship matrix ($\mathbf{A}$) and the GBVs ($\mathbf{u_g}$). However, this would be computationally much more demanding than the method described, as it would require dealing with dense matrices $\mathbf{A_{xg}}$ and $\mathbf{A_{gg}}$. The density of $\mathbf{A}$ depends upon relatedness amongst the individuals in the population but when founders are not included in the $\mathbf{A_{gg}}$ partition its inverse will be dense.

This method is routinely applied to a very large population of beef cattle. It is likely that blending will soon be replaced by the joint evaluation of animals' phenotypes and genotypes (the so-called 'single step'). However there may still be species where third parties provide GBVs and the need for blending may remain. In the dairy cattle and sheep industries genomic prediction equations are generally determined by the same organization responsible for their genetic evaluation. However, in beef cattle, GBVs are often provided by third parties, and the original data used to generate the equations may not be available for inclusion in routine genetic evaluations.

Conclusion

A simple and efficient method for generating GBVs and their approximate accuracies for un-genotyped individuals from their genotyped relatives has been presented. Estimated XGBVs are well correlated with the simulated GBVs, and the approximate accuracies are highly correlated with theoretical accuracies.

Literature Cited

Graser H.-U. and Tier B. (1997) Proc Assoc. Adv. Anim. Breed Genet. 12: 547-551.

Habier D., Fernando R.L. and Dekkers J.C.M. (2007) Genetics 177: 2389-2397.

Henderson C.R. (1984) Applications of Linear Models in Animal Breeding. U. Guelph.

Johnston D.J., Tier B. and Graser H.-U. (2009) Proc Assoc. Adv. Anim. Breed Genet. 17:30-33.

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