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

Approximation of Reliability of Direct Genomic Breeding Values

M. Sargolzaei[1,2], L. R. Schaeffer[2], J. P. Chesnais[1], G. Kistemaker[3], G. R. Wiggans[4], F. S. Schenkel[2]
[1]The Semex Alliance, [2]Centre for Genetic Improvement of Livestock, University of Guelph, [3]Canadain Dairy Network, Guelph, ON, Canada, [4]Animal Improvement Programs Laboratory, Agricultural Research Service, US Department of Agriculture, Beltsville, MD, USA

ABSTRACT: Two methods to efficiently approximate theoretical genomic reliabilities are presented. The first relies on the direct inverse of the left hand side (LHS) of mixed model equations. It uses the genomic relationship matrix for a small subset of individuals with the highest genomic relationship with the individual of interest. The second is a ridge-regression method using the direct inverse of LHS for a small subset of SNP. The performance of the methods was evaluated for the North American genomic data set, consisting of 228,168 genotyped individuals. The ridge-regression method gives very high correlations between theoretical and estimated reliabilities for subsets of 5k SNP and above. It is easily applicable to large data sets. Both methods lead to some biases in the mean and SD of reliabilities but these can be corrected by pegging to theoretical values or values from validation studies.
Keywords: direct inverse; genomic selection; reliability approximation

Introduction

The reliability is the squared correlation between estimated and true breeding values. Reliabilities are usually used to establish the confidence interval of breeding values for marketing purposes and to determine the weights required for blending direct genomic breeding values (DGV) with pedigree-based breeding values. Accurate reliabilities are also required for proper de-regression. Reliability for each individual can be obtained by calculating direct inverse of the left hand side of mixed model equations (MME). However, while this calculation is feasible for small data set it is not for the large data sets commonly used in national genetic/genomic evaluations.

Due to the simple structure of the inverse of the pedigree-based relationship matrix several efficient reliability approximation methods have been proposed (e.g., Harris and Johnson (1998); Jamrozik et al. (2000); Tier and Meyer (2004)). However, the genomic relationship matrix (**G**) is mostly defined by genotype similarities. There is no consistent pattern of genotype similarity between relatives due to Mendelian sampling and the similarity arising from identity by state. So far, there has not been a simple and quick rule-based method to invert the **G** matrix and obtaining the direct inverse of **G** for a large population is very time-consuming. Two genomic reliability approximation methods are presented which rely on inverting a relatively small matrix for a subset of training individuals or a subset of equidistant SNP.

Materials and Methods

Approximation method based on a subset of training individuals. Traditional breeding values are usually estimated by the best linear unbiased prediction (BLUP) methods using an animal model. (Co)variances between individuals are derived from pedigree. Genomic breeding values can also be estimated by the BLUP methods (GBLUP), replacing the pedigree-based (co)variance matrix with a (co)variance matrix calculated from genotype information (VanRaden et al. (2009)). The MME for the GBLUP method is:

$$\begin{pmatrix} \mathbf{1'R^{-1}1} & \mathbf{1'R^{-1}Z} \\ \mathbf{Z'R^{-1}1} & \mathbf{Z'R^{-1}Z + G} \end{pmatrix} \begin{pmatrix} \mu \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{1'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{pmatrix},$$

where **Z** is an incidence matrix relating the vector of animal random effects ($\hat{\mathbf{u}}$) to the vector of observations (**y**), μ is the overall mean, R is a diagonal matrix with diagonal element $r_{ii} = (100/reliability_i) - 1$, and **G** is the genomic (co)variance or relationship matrix. The genomic relationship matrix is calculated as $\mathbf{WW'}/2\sum p_k(1-p_k)$, where w_{ij} is the element for the i^{th} individual and the j^{th} locus, which is genotype call (AA=2, AB=1, BB=0) minus $2p_j$, where p_j is frequency of allele A at the j^{th} locus. The observation vector contains de-regressed values which have already been adjusted for fixed effects. The reliability (r^2) of genomic breeding values is calculated as:

$$r_i^2 = 1 - \frac{PEV_i}{g_{ii}},$$

where PEV_i is the prediction error variance for the i^{th} individual, which corresponds the diagonal of the inverse of the MME, and g_{ii} is the diagonal of the genomic relationship matrix. When the number of individuals in the reference population is large obtaining the inverse of MME is computationally demanding or unfeasible. Because individuals with a higher genomic relationship with the individual of interest contribute more to its DGV reliability, one may setup for each individual the MME for the n-closest relatives (based on genomic relationship) in the training set. If n is small enough inverting such MME should be efficient.

Approximation method based on a subset of equidistant SNP. An equivalent model to GBLUP is a ridge-regression with a shrinkage factor (λ) of 1:

$$\begin{pmatrix} \mathbf{1'R^{-1}1} & \mathbf{1'R^{-1}T'} \\ \mathbf{TR^{-1}1} & \mathbf{TR^{-1}T'} + \lambda\mathbf{I} \end{pmatrix} \begin{pmatrix} \mu \\ \hat{\mathbf{v}} \end{pmatrix} = \begin{pmatrix} \mathbf{1'R^{-1}y} \\ \mathbf{TR^{-1}y} \end{pmatrix},$$

where $\mathbf{T}$ is $\mathbf{W}(2\sum p_i(1-p_i))^{-0.5}$ and $\hat{\mathbf{v}}$ is a vector of SNP effects. Prediction error variance can then be obtained as:

$$PEV_i = \mathrm{t_i}\mathbf{C}\mathrm{t_i}',$$

where $\mathbf{C} = (\mathbf{TR^{-1}T'} + \mathbf{I})^{-1}$ (Strandén and Christensen (2011)). The size of matrix $\mathbf{C}$ is the number of SNP and therefore computation of $\mathbf{C}$ is also time-consuming. However, if we assume that a subset of SNP can capture most of genomic relationships between individuals then obtaining $\mathbf{C}^*$ for a subset of SNP should be computationally affordable.

In this study the possibility of approximating reliabilities based on a subset of training individuals or a subset of SNP was investigated. Both methods were examined on a large Holstein genomic data set.

Data. Genomic data set was provided by the Canadian Dairy Network. The data consisted of 22,856 training individuals with either domestic or across country (MACE) proofs, and 205,312 prediction individuals. Each individual had 45,187 original or imputed SNP genotypes from the Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA). These SNP are being used for official genomic evaluation in Canada (Canadian Dairy Network, Guelph, ON) and all have passed quality control measures used in national genomic evaluation. In this study, deregressed proofs and corresponding reliabilities for protein yield were used.

Statistical analyses. Reliabilities were approximated using either GBLUP (subset of training individuals) or the ridge-regression method (subset of SNP) for different numbers of individuals and SNP. The **G** matrix was not adjusted and no pedigree-based relationships were added to the **G** matrix. The correlation between approximated and theoretical reliabilities was calculated in order to measure the accuracy of the approximations.

Results and Discussion

Table 1 shows the correlation between approximated and theoretical reliabilities, mean and standard deviation (SD) of the approximated values for different numbers of training individual. Correlation between approximated and theoretical reliabilities increased as the number of training individuals with the highest genomic relationship increased. The accuracy of approximated reliabilities was lower for the prediction set compared to the training set. Although the correlations were moderate, reliabilities were severely underestimated compared to theoretical reliabilities and had substantially less variation. There was more underestimation for the prediction group. Computing time increased exponentially as the number of animals in the subset increased mainly because one MME must be inverted for each individual. A much larger number of training individuals than what was considered in this study is required to obtain a more accurate approximation, but then such approximation is no longer computationally justified.

Table 1. Correlation between approximated and theoretical reliabilities using GBLUP method.

	Training			Prediction			
No.[1]	r	Mean	SD	r	Mean	SD	Time[2]
40	0.858	4.79	0.46	0.636	3.33	0.44	0.068
80	0.863	5.06	0.47	0.665	3.58	0.46	0.075
180	0.872	5.40	0.48	0.702	3.86	0.50	0.092
360	0.882	6.59	0.55	0.741	4.93	0.58	0.221
640	0.896	7.70	0.59	0.783	5.88	0.67	1.121
Theo.	1.000	93.74	2.21	1.000	87.66	2.95	1.000

[1]Number of training animals with the highest genomic relationship with the individual of interest, for which the reliability needs to be approximated
[2]Computing time relative to the time required to obtain theoretical (theo.) reliabilities from GBLUP
SD: Standard deviation

Table 2. Correlation between approximated and theoretical reliabilities using ridge-regression method.

	Training			Prediction			
No.[1]	r	Mean	SD	r	Mean	SD	Time[2]
1k	0.839	99.54	0.06	0.833	99.44	0.05	0.004
3k	0.942	98.34	0.36	0.945	97.70	0.33	0.022
5k	0.973	97.29	0.71	0.976	95.88	0.72	0.054
10k	0.994	95.71	1.34	0.995	92.63	1.54	0.217
15k	0.998	94.98	1.66	0.998	90.91	2.01	0.496
20k	0.999	94.55	1.85	0.999	89.81	2.33	0.908
Theo.	1.000	93.74	2.21	1.000	87.66	2.95	1.000

[1]Number of equidistant SNP
[2]Computing time relative to the time required to obtain theoretical (theo.) reliabilities from GBLUP
SD: Standard deviation

The results for reliability approximations based on a subset of equidistant SNP are given in Table 2. For subset of 5k and above, there was good agreement between approximated and theoretical reliabilities, and correlations were very high. Accuracies for training and prediction sets were almost the same. There was less bias in the mean and SD of approximated reliabilities compared to those from GBLUP using a subset of training individuals. However, these biases were still significant enough that they would require correction.

In most livestock species, the size of the training set increases over time due to phenotyping and genotyping of new individuals. This increase is not expected to impact the mean and SD of approximated reliabilities from the GBLUP method because the size of the subset is fixed.

With ridge-regression method, the increase in size of the training set is taken into account, which makes the method robust to the changes in the size of the training set.

For the ridge-regression method, several different selection strategies for the subset of SNP were investigated: a) n equidistant SNP, b) n randomly selected SNP, c) n SNP with the largest absolute effects, d) n SNP with a similar distribution of SNP effects as the full set, and e) n SNP with a similar allele frequency distribution as the full set. Approximation accuracies were very close for all above scenarios when the number of SNP in the subset was larger than 3k (results not shown).

For both approximation methods, the bias in the mean and SD could be adjusted periodically (for example once a year between official evaluations). This adjustment should be done through equivalent daughter contributions (EDC) rather than directly on reliabilities. If a direct inverse is still feasible with the full set of SNP or individuals, one can establish a prediction equation by regressing EDC based on the full set of SNP or individuals on EDC based on the reduced set. Then reliabilities in subsequent evaluations can be adjusted based on the change in EDC resulting from this prediction equation. If a direct inverse is not feasible or too time-consuming, prediction equations based on SNP or individuals subsets of increasing size can be used to obtain a prediction for the full SNP set. Given that the number of SNP in genomic evaluation is more stable over time than the number of training individuals, this adjustment is more practical for the ridge-regression method with a subset of SNP. The mean of approximated reliabilities could also be equated to the average reliability derived from a genomic validation study for the same trait, rather than to the mean of theoretical reliabilities. This would have the merit of forcing the reliabilities of individual animals to be more in line with those found from validation. If need be, an upward adjustment to this mean could be made to account for the larger number of reference animals in official evaluations compared to validation. However, reliability increases that result from adding new animals to current reference populations appear to be lower than theoretical predictions, likely because our statistical models do not account for all biological effects (e.g. non-additive effects, gene interactions).

Conclusion

Estimated reliabilities were substantially more accurate with the ridge-regression method based on a subset of SNP than with the GBLUP method based on a subset of closely related training individuals. The correlation between approximated and theoretical reliabilities was very high for the method based on a subset of SNP. Both methods lead to biases in the mean and SD of approximated reliabilities but these can be adjusted periodically by pegging to theoretical EDC values obtained from a data set where the computation is feasible. For the mean, an alternative is to peg the approximated reliability to the average reliability from validation studies. The ridge-regression method was more robust across different scenarios and computationally more efficient than the GBLUP method.

Acknowledgement

The Canadian Dairy Network is gratefully acknowledged for providing the genomic data set for this work.

Literature Cited

Harris, B., and Johnson, D. (1998). J. Dairy Sci. 81: 2723–2728.

Jamrozik, J., Schaeffer, L., and Jansen, G. B. (2000). Livest. Prod. Sci. 66: 85–92.

Strandén, I., and Christensen, O. F. (2011). Genet. Sel. Evol. 43: 25.

Tier, B., and Meyer, K. (2004). J. Anim. Breed. Genet. 121: 77–89.

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