Hierarchical Quantitative Genetic Model Using Genomic Information

***G. Gorjanc*, J. A. Woolliams[1], and J. M. Hickey[1]**
University of Edinburgh, The Roslin Institute, United Kingdom

ABSTRACT: Genomic evaluations are commonly based on models with genomic relationships. However, estimated genomic relationship matrices are often positive semi-definite and ad-hoc corrections are applied to force positive definiteness without consideration about model properties. In this contribution a hierarchical quantitative genetic model is postulated that provides a positive definite genomic relationship matrix by taking into account the amount of genetic variance captured by estimated marker effects. Based on the hierarchical formulation the proposed model also provides a system of equations to estimate marker effects and breeding values jointly without setting up and inverting covariance matrices. Further extension with pedigree information is also possible.

Keywords: genomics; hierarchical model; inverse

Introduction

Genomic information has empowered animal breeding with the ability to directly measure segregation of genomes within populations. In genetic evaluation this information provides more accurate estimates of genetic covariances among individuals than pedigree information. However, the initial proposal of genetic evaluation with genomic relationships (GBLUP; VanRaden (2008)) is based on an improper prior with a positive semi-definite covariance matrix (Rue and Held (2005)), due to the implicit assumption that genome-wide marker effects capture all additive genetic variance. Validation of predictions from these models clearly shows that current estimates of genome-wide markers do not capture all additive genetic variance (e.g., VanRaden (2008); Goddard et al. (2011)). Nonetheless, the often used ad-hoc solution to force genomic relationship matrix positive definite is to add a small value to its diagonal or to blend genomic and pedigree matrices without consideration about the model properties (Carré et al. (2013)).

In this contribution a statistical model for genetic evaluation using genomic information is postulated based on a hierarchical formulation that takes into account the amount of variance explained by markers. This formulation provides proper prior with a positive definite covariance matrix as well as an efficient computational strategy to jointly estimate marker effects and breeding values and can be extended in hierarchical manner to include pedigree information as well.

Materials and Methods

Methods. GBLUP is based on the hierarchical model involving phenotype model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b} + \boldsymbol{Z}\boldsymbol{a} + \boldsymbol{e}, \qquad (1)$$

and additive genetic model:

$$\boldsymbol{a} = \boldsymbol{W}\boldsymbol{m}, \qquad (2)$$

$$Var(\boldsymbol{a}) = \boldsymbol{V}_a = \tfrac{1}{s}\boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}\sigma_a^2, \qquad (3)$$

where $\boldsymbol{a}$ is a $n_a \times 1$ vector of breeding values, $\boldsymbol{W}$ is a $n_a \times n_m$ matrix of centered marker genotypes $-2p_j$, $1-2p_j$, and $2-2p_j$ for genotypes 0/0, 0/1, and 1/1, respectively, $\boldsymbol{m}$ is a $n_m \times 1$ vector of marker effects, $s = 2\sum_i^{n_m} p_i(1-p_i)$ is a scale factor of expected variances of marker genotypes with p_i being frequency of i-th marker allele, and σ_a^2 is additive genetic variance (VanRaden, (2008)). Alternatively marker genotypes can be scaled with marker specific standard deviations $s_i = \sqrt{2p_i(1-p_i)}$ and $\boldsymbol{V}_a = \frac{1}{n_m}\boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}\sigma_a^2$.

The formulation in the genetic model (2) implies that genome-wide markers capture all additive genetic variance. An alternative definition to (2) is:

$$\boldsymbol{a} = \boldsymbol{W}\boldsymbol{m} + \boldsymbol{r}, \qquad (4)$$

with prior assumptions:

$$\boldsymbol{a}|\boldsymbol{W},\boldsymbol{m},\sigma_r^2 \sim N(\boldsymbol{W}\boldsymbol{m},\boldsymbol{I}\sigma_r^2), \qquad (5)$$

$$\boldsymbol{m}|\sigma_m^2 \qquad \sim N(\boldsymbol{0},\boldsymbol{I}\sigma_m^2), \qquad (6)$$

where $\boldsymbol{r}$ describes the part of breeding values not captured by markers (the remainder) with the associated variance component σ_r^2 and σ_m^2 is variance of marker effects. Defining $\sigma_m^2 = \frac{c\sigma_a^2}{n_m}$ and $\sigma_r^2 = (1-c)\sigma_a^2$, where c is the proportion of additive genetic variance captured by marker effects, the covariance structure of (4) is:

$$\boldsymbol{V}_a = \left(\boldsymbol{W}\boldsymbol{W}^{\boldsymbol{T}}\tfrac{c}{n_m} + \boldsymbol{I}(1-c)\right)\sigma_a^2. \qquad (7)$$

The inverse of (7) $\boldsymbol{Q}_a$ is required for inference and can be obtained by inversion of a $n_a \times n_a$ matrix $\boldsymbol{V}_a$. However, by recognizing the hierarchical formulation in (5-6) this inversion can be skipped in favor of setting up joint inverse covariance (precision) matrix for breeding values and markers $(\boldsymbol{Q}_{r,m})$ directly (e.g., Rue and Held (2005)):

$$\boldsymbol{Q}_{r,m} = \begin{pmatrix} \boldsymbol{Q}_r & -\boldsymbol{Q}_r\boldsymbol{W} \\ -\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{Q}_r & \boldsymbol{Q}_m + \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{Q}_r\boldsymbol{W} \end{pmatrix}, \qquad (8)$$

$$= \begin{pmatrix} \boldsymbol{I}\frac{1}{1-c} & -\boldsymbol{W}\frac{1}{1-c} \\ -\boldsymbol{W}^{\boldsymbol{T}}\frac{1}{1-c} & \boldsymbol{I}\frac{n_m}{c} + \boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W}\frac{1}{1-c} \end{pmatrix}\frac{1}{\sigma_a^2}, \qquad (9)$$

where the subscripts of precision matrices $\boldsymbol{Q}_{r,m}$, $\boldsymbol{Q}_r$, and $\boldsymbol{Q}_m$ match the subscripts of variance components in (5-6). The mixed model equations for estimating $\boldsymbol{b}$, $\boldsymbol{a}$, and $\boldsymbol{m}$ can then be directly setup as:

$$\begin{pmatrix} \boldsymbol{X}^T\boldsymbol{Q}_e\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{Q}_e\boldsymbol{Z} & \boldsymbol{0} \\ \boldsymbol{Z}^T\boldsymbol{Q}_e\boldsymbol{X} & \boldsymbol{Q}_r + \boldsymbol{Z}^T\boldsymbol{Q}_e\boldsymbol{Z} & -\boldsymbol{Q}_r\boldsymbol{W} \\ \boldsymbol{0} & -\boldsymbol{W}^T\boldsymbol{Q}_r & \boldsymbol{Q}_m + \boldsymbol{W}^T\boldsymbol{Q}_r\boldsymbol{W} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{b}} \\ \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{m}} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}^T\boldsymbol{Q}_e\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{Q}_e\boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}, \quad (10)$$

where $\boldsymbol{Q}_e = \boldsymbol{I}\frac{1}{\sigma_e^2}$.

Data. The proposed model was tested on a simulated data of five 1 Morgan chromosomes with 2000 single nucleotide markers per chromosome, 300 quantitative trait loci per chromosome with effects sampled from a normal distribution, and quantitative trait with heritability of 0.3. These chromosomes were dropped through a pedigree of 2 generations with each having 1,000 individuals from 25 sires each mated to 20 dams producing 2 offspring per mating.

Analysis. In the first generation a random set of 500 individuals had phenotypes available for genomic prediction of the remaining 500 nominally unrelated individuals in the first generation and individuals in the second generation that were progeny of phenotyped individuals in the first generation. This design allowed predictions to be tested for unrelated individuals as well as progeny of phenotyped parents. The proposed model was used with assumed known variances σ_a^2 and σ_e^2 equal to used values in simulation, while the effect of the proportion of variance explained by markers c was quantified by performing genetic evaluations for the grid of values from 0.01 to 0.99. All analyses were performed using either 10,000 or 1,000 markers. The results were summarized with the accuracy of prediction defined as a correlation between the true and estimated breeding values, and the bias of prediction defined as a regression coefficient of true on estimated breeding values.

Results

Accuracy of estimated breeding values increased with the increasing c in training individuals and reached maximum of 0.66 and 0.65 with the amount of genetic variance captured by marker effects c equal to 0.99 and 0.98 for 10,000 and 1,000 marker set, respectively (Figure 1). When predicting breeding values in non-phenotyped progeny maximal accuracies were 0.63 and 0.55 with substantially lower c equal to 0.23 and 0.08 for 10,000 and 1,000 marker set, respectively. When predicting breeding values in non-phenotyped unrelated individuals maximal accuracies were 0.53 and 0.48 with c equal to 0.35 for both marker sets. The difference in c between close relatives (progeny) and unrelated individuals might be due to the different importance of linkage and linkage disequilibrium information for prediction.

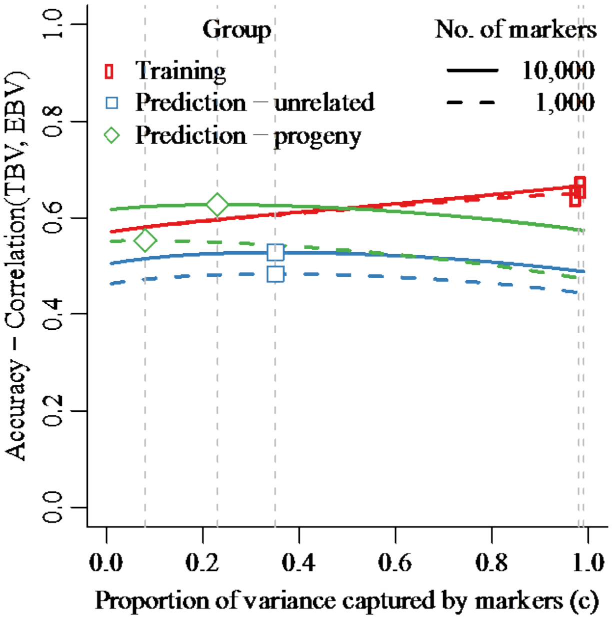

Figure 1: Accuracy of genomic evaluation for different groups of individuals and different number of markers used in relation to the proportion of variance captured by estimated marker effects – the largest values are marked with symbols

Regression coefficients of true on estimated breeding values increased with the increasing c in training individuals and reached maximum of 0.56 with c equal to 0.99 for both 10,000 and 1,000 marker sets (Figure 2). In non-phenotyped progeny and unrelated individuals bias decreased with the increasing c from values around 20 when c was close to 0 and to values around 0.5 when c was close to 1, with smaller values for the 1,000 marker set in comparison to the 10,000 marker set. In these two groups of individuals bias was close to 1 for values of c around 0.45 for the 10,000 marker set and for values of c around 0.37 for the 1,000 marker set.

Discussion

The results show that estimated marker effects did not capture all the genetic variance in the simulated data and that for prediction of non-phenotyped individuals small to intermediate values of c gave higher accuracies than values close to 1, which are commonly used. Regression coefficients of true on estimated breeding values were also close to 1 for the intermediate range of c values for this simulation. While the simulation used in this study was quite small obtained accuracies match well the theoretical expectations (Daetwyler et al. (2008); results not shown). Larger dataset would implicitly enable more accurate estimates of markers and therefore more genetic variance captured. This is exactly why the additional residual $\boldsymbol{r}$ was included in the model to allow for estimation errors and plays the same role as the Mendelian sampling residual in pedigree models. The expectation $\boldsymbol{Wm}$ is on the other hand equivalent to parent average (Thompson (1977)), though there is no recursive structure with genomic data as there is

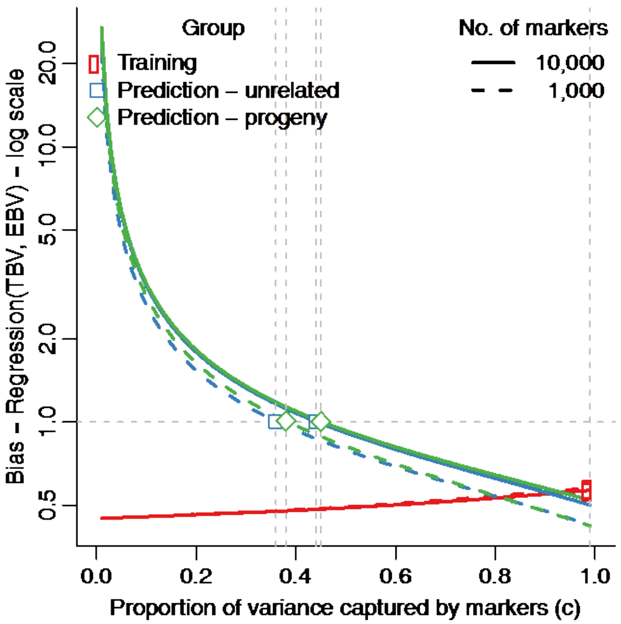

Figure 2: Bias of genomic evaluation for different groups of individuals and different number of markers used in relation to the proportion of variance captured by estimated marker effects – values the closest to 1 are marked with symbols

in pedigrees. The hierarchical model presented in this study is also equivalent to the work of Goddard et al. (2011) who modified genomic relationship matrix to account for errors in estimating genomic relationships, which implicitly defines model (4).

The proposed model is computationally more stable because the derived relationship matrix (7) is positive definite when c is smaller than 1. In addition this model provides a system of equations where both markers and breeding values can be estimated jointly without the need to setup and invert any covariance matrices. This is possible, due to the hierarchical formulation (5-6), which is an analog of hierarchical formulation of pedigree relationship matrix (Henderson (1976)). This system of equations is however larger and denser than usual due to dense matrix $\boldsymbol{W}$. We also attempted to derive inverse $\boldsymbol{Q}_a$ directly using matrix inversion lemma of (7) for use in a system of equations without marker effects. However, that would require inverse of a $n_m \times n_m$ matrix (results not shown), which is computationally more demanding than inverting $\boldsymbol{V}_a$ when the number of individuals is smaller than the number or markers. Until then a suggested approach is to use system (10) or to invert matrix $\boldsymbol{V}_a$.

Following the hierarchical formulation akin to a model with genetic groups Gengler et al. (2012) proposed the same type of a model as in (4), but with a combination of genomic and pedigree information, i.e., the single-step approach, and also noticed the benefit of jointly estimating breeding values and marker effects without the need to setup and invert covariance matrices. Their approach however involves the non-hierarchical formulation of joint pedigree and genomic priors that requires setting up and inverting pedigree relationship matrix for genotyped individuals. We have derived a quantitative genetic model with hierarchical formulation of both pedigree and genomic information for several common patterns of genotyped sets of individuals by merging the model presented here and the approach of Carré et al. (2013), which is beyond the scope of this contribution.

Variance components were assumed known in this work and need to be estimated in real applications, which might prove challenging as there is potential confounding between the $\boldsymbol{r}$ and $\boldsymbol{e}$ terms. This confounding is the same as confounding between the Mendelian sampling residual in pedigree model and phenotype residual if only own performance records are available and pedigree is not very informative or even missing in the limit. However, with pedigree model the amount of genetic variance captured by the expectation (parent average) and residual (Mendelian sampling) is fixed by definition, while this is a free parameter c in the proposed model. The required structure of data to be able to estimate all three variance parameters ($\sigma_e^2, \sigma_a^2,$ and c) needs to be found. An alternative is to estimate c using the theory from Goddard et al. (2011). However, in the presented simulation an estimate of c from the effective population size and genome size gives 0.98 (results not shown), which is far from the optimal c values found for prediction sets. More work is needed in this area. For pedigreed populations a combination of genomic and pedigree information might provide enough structure to estimate these variance parameters.

Conclusion

In summary this contribution presented a hierarchical quantitative genetic model using genomic information that takes into account amount of additive genetic variance explained by markers. This formulation provides proper prior with a positive definite covariance matrix and favorable computational properties to estimate marker effects and breeding values jointly.

Literature Cited

Carré, C., F. Gamboa, D. Cros, et al. (2013). *Genetica*. 141:239–246.
Daetwyler, H.D., B. Villanueva, J.A. Woolliams, et al. (2008). PLoS ONE. 3,e3395.
Gengler, N., G.J. Nieuwhof, K.V. Konstantinov, et al. (2012). EAAP.
Goddard, M.E., B.J. Hayes, and T.H.E. Meuwissen. (2011). *J. Anim. Breed. Genet.* 128:409–421.
Henderson, C.R. (1976). *Biometrics*. 32:69-83.
Rue, H., and L. Held. (2005). Gaussian Markov random fields: theory and applications. Chapman & Hall/CRC, Boca Raton.
Thompson, R. (1977). *Biometrics*. 33:497-504
VanRaden, P.M. (2008). *J. Dairy Sci.* 91:4414–4423.