Estimation of variance components in populations under genomic selection

H. Gao[1], P. Madsen[1], G.P. Aamand[2] & J. Jensen[1]

[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, DK-8830, Tjele, Denmark
hongding.gao@mbg.au.dk (Corresponding Author)
[2]Nordic Cattle Genetic Evaluation, DK-8200, Aarhus N, Denmark

Summary

Accurate estimation of variance components (VCs) is needed for genomic prediction, but until now VCs are usually estimated by restricted maximum likelihood (REML) in pedigree-based animal model (P-AM). In this study, REML in P-AM and Markov chain Monte Carlo (MCMC) procedure via Gibbs sampling in single-step Bayesian regression model (SSBR) were used to investigate the consequences of VCs estimation based on different subsets of phenotypic information. Three scenarios were analyzed: (1) phenotypes only from conventional (P-AM) part of the breeding scheme; (2) phenotypes from both conventional and genomic selection (GS) parts of the breeding scheme; (3) phenotypes only from the GS part of the breeding scheme. Unbiased estimation of VCs was obtained before GS era via P-AM. Including records from GS era led to biased estimation of VCs for P-AM. SSBR allowed utilizing data from all periods and combine information from both genotyped and non-genotyped individuals, and unbiased estimated genetic variance was achieved. With the advantage of sampling remaining genetic and marker variances separately, SSBR have the potential option to achieve unbiased VCs estimation in populations under GS.

Keywords: variance components, single-step, imputation

Introduction

Prediction of breeding values needs accurate estimates of variance components (VCs). Although there are several genomic prediction models to choose from, VCs used in these models usually rely on analysis where additive genetic effects are modeled using pedigree relationships. Before genomic information was available Van der Werf & Deboer (1990) showed that unbiased estimation of VCs can be achieved when all data that lead to selection and complete relationships were used in animal model, and ignoring data from selected ancestors led to biased estimation due to not accounting for gametic disequilibrium. However these results were obtained under the infinitesimal model and with the implementation of genomic selection (GS), firstly, gene frequencies may change faster compared to phenotypic selection. More importantly, GS introduced extra pre-selection components, hence if the populations are under GS, the estimations of VCs from pedigree-based animal model (P-AM) are expected to be biased due to not accounting the impact of GS. Secondly, phenotypes recorded under GS should also be utilized to investigate VCs instead of eliminating them. Furthermore, new traits might be continuously included in the breeding programs. Along with the above reasons, a genomic model needs to be considered for VCs estimation in populations under GS.

Lately, a single-step Bayesian regression model (SSBR) was proposed by (Fernando et al., 2014, Fernando et al., 2016), which require to explicitly impute the markers for non-genotyped individuals and then fit the marker effects in the model. Compared to single-step genomic BLUP

(SSGBLUP) (Legarra et al., 2009, Christensen & Lund, 2010), SSBR avoids brute-force matrix inversion, in addition, it provides a feature to estimate marker variance and remaining genetic variance separately. The objective of this study was to explore the VCs estimation based on simulated data using different strategies for choosing the data to be included in the analysis.

Material and methods

A Danish Jersey population was mimicked in terms of practical breeding scheme and population structure. The simulated VCs for genetic and residual effects were 3.65 and 5.05, respectively. VCs estimation was conducted using restricted maximum likelihood (REML) and Markov chain Monte Carlo (MCMC) procedure via Gibbs sampling. Five replicates for each scenario were used in this study.

Simulation design

A Danish Jersey population distributed in 100 herds was simulated using a stochastic finite locus model in consecutive 3 stages: (1) Historic population, covering 3,000 non-overlapping generations in order to create an initial linkage disequilibrium (LD) structure. A total of 3×10^8 SNP markers were evenly spaced with a mutation rate of 1.8×10^{-6} to establish mutation-drift equilibrium, and every 32nd SNP was considered to be a potential QTL. 2,000 segregating QTLs and 40,000 segregating markers were retained in the base population. (2) A conventional breeding phase was simulated for 20 years to reflect the current size of Danish Jersey population. Each year, 50 bulls were selected based on their parent average (PA) and 5 proven bulls were selected from these 50 bulls for further insemination. All these 50 bulls were genotyped and 10,000 cows in different age groups were maintained each year. Only cows in first lactation, however, were assumed to have phenotypes. In stage (3) a genomic selection based on SSGBLUP was simulated for 15 years. Each year, 500 bulls and 2,000 heifers were genotyped based on selection of their GEBVs. The simulation was performed in ADAM software (Pedersen et al., 2009).

Statistical models

P-AM

The classic animal model (Henderson, 1984) using pedigree-based relationship was:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a} + \mathbf{e} \tag{1}$$

where $\mathbf{y}$ is the vector of phenotypes, $\boldsymbol{\beta}$ is a vector of fixed effects, i.e. herd-year-season effect, $\mathbf{a}$ is a vector of additive genetic effects $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$, where $\mathbf{A}$ is the numerator relationship matrix and σ_a^2 is the additive genetic variance. $\mathbf{X}$ and $\mathbf{Z}$ are design matrices, and $\mathbf{e}$ is a vector of residuals $\boldsymbol{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance. VCs were estimated by REML (Patterson & Thompson, 1971) with the average information algorithm (AI-REML) (Jensen et al., 1997) implemented in DMUAI (Madsen & Jensen, 2013).

SSBR

The SSBR model based on Bayes C with assumption of all markers have non-zero effects (Fernando et al., 2014) is as follows:

$$\begin{bmatrix} \mathbf{y}_n \\ \mathbf{y}_g \end{bmatrix} = \begin{bmatrix} \mathbf{X}_n \\ \mathbf{X}_g \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{Z}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_g \end{bmatrix} \begin{bmatrix} \mathbf{g}_n \\ \mathbf{g}_g \end{bmatrix} + \mathbf{e} \qquad (2)$$

where subscript n is denoted for non-genotyped individuals and subscript g is denoted for genotyped individuals. Thus, **y** represent vectors of phenotypes, **β** is a vector of fixed effects, i.e. herd-year-season effect, **X** and **Z** are design matrices, **e** is a vector of residuals, **g** is a vector of GEBVs can be written as follows:

$$\begin{bmatrix} \mathbf{g}_n \\ \mathbf{g}_g \end{bmatrix} = \begin{bmatrix} \mathbf{M}_n \boldsymbol{\alpha} + \boldsymbol{\varepsilon} \\ \mathbf{M}_g \boldsymbol{\alpha} \end{bmatrix} \qquad (3)$$

where $\mathbf{M}_g$ is a matrix contains observed marker covariates for genotyped individuals, $\mathbf{M}_n$ is a matrix contains imputed marker covariates for non-genotyped individuals via linear relationship with $\mathbf{M}_g$, **α** is the vector of random marker effects, **ε** is the vector of imputation residuals.

Gibbs sampler was used to obtain MCMC samples of the unknown parameters from their marginal distributions. The length of MCMC chain was set to 50,000 with a burn-in of 20,000 iterations. Convergence diagnostics for MCMC were assessed in R package boa (Smith, 2007) and all parameters investigated were converged.

Scenarios

Three scenarios based on choosing different subsets of available phenotypes for estimation of VCs were investigated. An overview of the subsets used and the average number of pedigree, phenotypes and genotypes for each scenario over 5 replicates are shown in Table 1.

Table 1. Scenarios based on which subsets of phenotypes used for VCs estimation and a*verage statistics of simulated datasets over 5 replicates.*

Scenario	Conventional phase	GS phase	no. in pedigree	no. of phenotypes	no. of genotypes[1]
1	✓		84,164	81,240	1,050
2	✓	✓	160,131	144,728	38,550
3		✓	106,011	63,487	38,550

[1] Genotypes were only used in SSBR.

Results and Discussion

Table 2 presents the mean and standard error (SE) of estimated VCs for each data subset over 5 replicates. As expected, genetic variance and residual variance were precisely estimated in P-AM scenario 1 (P-AM 1). This result is well known in the infinitesimal model that selection can be accounted for in the mixed model via the numerator relationship matrix (Sorensen & Kennedy, 1984). In P-AM 2 and P-AM 3, genetic variances were underestimated and residual variances were overestimated compared to the true simulated VCs. This could be mainly due to the pre-selection on genomic information, since this information was ignored in the classical animal model.

Compared to P-AM, by employing the full phenotypes and genomic information, SSBR 2 yielded unbiased estimates of genetic variance, but the residual variance was overestimated. In SSBR 1, genetic variance was overestimated although estimated residual variance was close. In SSBR 3, genetic variance was underestimated while residual variance was overestimated. With the intensive use of GS, our results confirmed the shift of genetic variance due to pre-selection for young bulls based on GEBVs. In general, all models that included only the phenotypes from GS part of the breeding scheme resulted in underestimated genetic variances and overestimated residual variances (P-AM 3 and SSBR 3).

In SSBR, the genetic variance is essentially variance of genetic effects of non-genotyped individuals conditional on the genetic effects of genotyped individuals (Legarra et al., 2009). With the assumption of multivariate normality Markers of non-genotyped individuals were imputed via a pedigree-based linear system, in which imputation quality is largely depending on the genetic relationships between genotype and non-genotyped individuals, i.e. old ancestors would expect to obtain worse imputation quality than the younger ones. However, the imputation was under the assumption of multivariate normality which is an approximation. Therefore, better methods for imputing genotypes of non-genotyped individuals conditional on the genotyped ones would be expected to yield more accurate results Alternatively, in a specific case where sires of all phenotyped daughters are genotyped, a SSBR sire model instead of the animal model could be an option to avoid the imputation process and be able to estimate the sire genetic variance based on all genomic information.

Table 2. Mean and standard error (SE) of estimated variance components for each scenario over 5 replicates.

Method	Scenario	σ_a^2	σ_e^2
True VCs		3.65	5.05
P-AM	1	3.62 (0.05)	5.02 (0.01)
	2	3.52 (0.03)	5.07 (0.01)
	3	3.19 (0.09)	5.20 (0.04)
SSBR	1	4.19 (0.04)	5.10 (0.02)
	2	3.57 (0.04)	5.24 (0.01)
	3	3.03 (0.04)	5.27 (0.03)

Conclusion

The objective of this study was to investigate VCs estimation in a simulated Danish Jersey population where GS has been conducted for several generations. P-AM and SSBR were used to estimate VCs based on different subsets of phenotypic information. SSBR shows the potential capability to yield unbiased VCs estimation in GS era whereas P-AM ignoring genomic information yields biased results.

Acknowledgements

The first author is grateful to Rohan L. Fernando and Dorian J. Garrick for the acceptance of two month visit at Iowa State University. Their patient guiding on SSBR method are fully appreciated. Discussions with Hao Cheng and Hailin Su are appreciated. We acknowledge Anders Christian Sørensen and Jørn Rind Thomasen for the help with ADAM simulation. The first author acknowledges the funding from Innovation Fund Denmark (Copenhagen), Nordic Cattle Genetic Evaluation (Aarhus, Denmark), and Aarhus University.

List of References

Christensen, O. and M. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol 42(1):2.

Fernando, R. L., H. Cheng, B. L. Golden, and D. J. Garrick. 2016. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. Genetics Selection Evolution 48.

Fernando, R. L., J. C. M. Dekkers, and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genetics Selection Evolution 46.

Henderson, C. R. 1984. Applications of linear models in animal breeding. University of Guelph, [Guelph, Ont.].

Jensen, J., E. A. Mäntysaari, P. Madsen, and R. Thompson. 1997. Residual maximum likelihood estimation of (co) variance components in multivariate mixed linear models using average information. Jour Ind Soc Ag Statistics 49:215-236.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci 92(9):4656-4663.

Madsen, P. and J. Jensen. 2013. A User's Guide to DMU, Version 6, Release 5.2. Centre for Quantitative Genetics and Genomics, Dept. of Molecular Biology and Genetics, University of Aarhus.

Patterson, H. D. and R. Thompson. 1971. Recovery of inter-block information when block sizes are unequal. biometrika 58(3):545-554.

Pedersen, L. D., A. C. Sorensen, M. Henryon, S. Ansari-Mahyari, and P. Berg. 2009. ADAM: A computer program to simulate selective breeding schemes for animals. Livest Sci 121(2-3):343-344.

Smith, B. J. 2007. boa: An R package for MCMC output convergence assessment and posterior inference. J Stat Softw 21(11):1-37.

Sorensen, D. A. and B. W. Kennedy. 1984. Estimation of Response to Selection Using Least-Squares and Mixed Model Methodology. Journal of Animal Science 58(5):1097-1106.

Van der Werf, J. H. J. and I. J. M. Deboer. 1990. Estimation of Additive Genetic Variance When Base-Populations Are Selected. Journal of Animal Science 68(10):3124-3132.