Proc. Assoc. Advmt. Anim. Breed. Genet. 25: 39 - 42

EFFECTS OF SELECTION AND DATA TRUNCATION ON ESTIMATES OF GENETIC PARAMETERS OBTAINED FITTING A SINGLE-STEP MODEL

Karin Meyer

Animal Genetics Breeding Unit*, University of New England, Armidale, NSW 2351 Australia

SUMMARY

Simulation was used to illustrate the effects of genomic selection on estimates of genetic parameters, comparing values when genomic relationships were ignored with those obtained accounting for the joint relationship matrix of genotyped and non-genotyped individuals. Analyses were carried out with increasing truncation of earlier records, pedigrees and genotype information. Results showed that estimates from pedigree only analyses could be markedly biased downwards as more historical data is ignored, especially with strong genomic selection, causing predicted breeding values for selection candidates in the last generation to be under-dispersed.

INTRODUCTION

Increasingly genetic evaluation schemes for livestock incorporate genomic information on a routine basis. To date, the most common method is single-step genomic best linear unbiased prediction (ssGBLUP) fitting a breeding value model. This replaces the pedigree-based inverse of the numerator relationship matrix with its counterpart which combines pedigree and genomic information (Aguilar *et al.* 2010). It is a conceptually simple extension of the classic prediction procedures using pedigree based relationships only (PBLUP). Like PBLUP, ssGBLUP requires appropriate values of genetic parameters as input. It is common practice to estimate these fitting the same – or at least a very similar – model as used for prediction of breeding values (EBV). Reviewing the status of genomic evaluation, Misztal *et al.* (2020) advocated inclusion of genomic relationships when estimating genetic parameters to counteract the bias due to genomic selection. The authors also recommended frequent re-estimation as genetic variances appeared to change quicker with ssGBLUP.

However, to date, estimates are mostly obtained considering pedigree based relationships only, and little is known about the impact of doing so on the efficacy of genomic selection. This paper presents a simple simulation study exploring the effects of accounting for genomic relationships on estimates of genetic parameters and the resulting accuracy of ssGBLUP based selection.

MATERIAL AND METHODS

Data were simulated for a trait with heritability of 0.3 and for individuals from 13 generations using the software package AlphaSim, version 1.05 (Faux *et al.* 2016). The data set contained records for 2100 and 3150 animals, respectively in generations 1 to 7 and 8 to 13, who were the progeny of 100 and 150 sires and 1000 and 1500 dams, respectively. To mimic a distribution over fixed effects subclasses, records were randomly assigned to 51 'contemporary groups' per generation. Genotypes were constructed by sampling 10 chromosomes with 4,000 single nucleotide polymorphism (SNP) and 50 quantitative trait nucleotide (QTN) each, randomly allowing for some QTN to be included among the SNP and assuming no mutation or recombination. Marker information for all individuals in generations 10 to 13 was retained, disregarding earlier genomic information.

AlphaSim provides the option to carry out selection in individual generations externally by allowing the user to select the parents and mating allocations of the next generation (Faux *et al.* 2016). This was utilised to implement three alternative selection schemes, combining random selection with selection on EBV obtained using pedigree relationships only and EBV from ssGBLUP analyses.

* A joint venture of NSW Department of Primary Industries and University of New England,

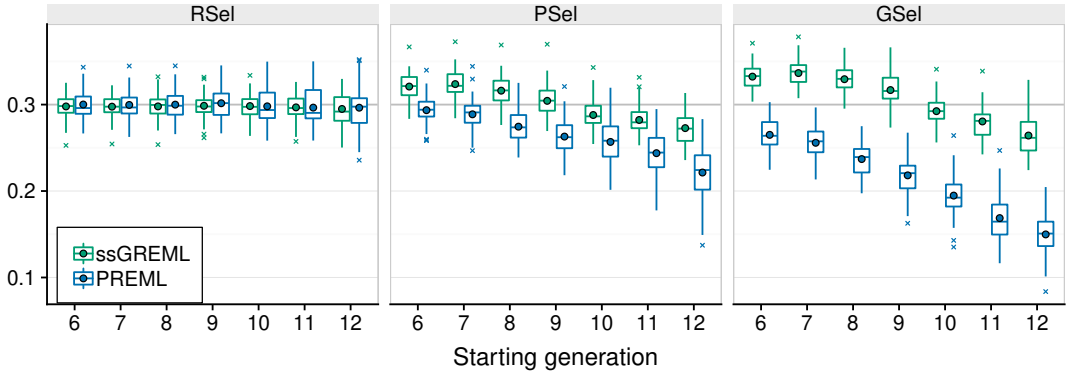

Figure 1. Distribution of heritability estimates over replicates (see text for definitions)

Discarding generations 1 to 4 as burn-in, parents of generations 5 to 7 were chosen at random. 1) For a genomic scenario (GSel) selection was on EBV from PBLUP in generations 8 to 10 and on EBV from ssGBLUP in generations 11 to 13. This was contrasted with 2) selection on EBV from PBLUP in generations 8 to 13 (PSel) and 3) random selection throughout (RSel). EBV were obtained from restricted maximum likelihood (REML) analyses at convergence. For generation i analyses utilised data and pedigree information from generation 6 to i (to select the parents of generation $i+1$) and, where applicable, all marker information from generation 10 to i.

To investigate the effects of selection bias and truncation of data on estimates of genetic parameters, analyses were carried out successively ignoring information from earlier generations, i.e. considering records, pedigrees and marker counts from generations i to 13 only where $i = 6,\ldots,12$. In the following, we refer to generation i as the 'starting generation' for an analysis. Accuracy and dispersion of EBV for selection candidates in generation 13 were measured as the correlation between and regression of true breeding values (TBV) on EBV. 50 replicates were carried out for each scenario.

REML analyses (for both the external selection steps and the data sets sampled) used either pedigree based relationships only (PREML) or pedigree and genomic relationships jointly (ssGREML), fitting a simple animal model with contemporary group as the only fixed effects. Genomic relationship matrices ($\mathbf{G}$) were built using Method 1 of Van Raden (2008), eliminating SNP with minor allele frequencies less than 2% and centering allele counts using mean frequencies in the data. These were aligned to their pedigree based counterparts ($\mathbf{A}_{22}$) following Vitezica *et al.* (2011).

RESULTS

The distribution of heritability estimates over replicates for the three selection strategies is summarised in Figure 1. In all cases, means – depicted by circles – agreed closely with the median values. As expected, for RSel, estimates from ssGREML and PREML did not differ noticeably and showed no bias, though some differences in variability across replicates were evident. For PSel and GSel, however, estimates depended strongly on the subset of data utilised. Loosely described, REML can account for selection bias, provided the information that selection decisions were based on is included in the analysis. Hence, for data starting at generations $i = 6$ or 7, no selection bias was evident for PSel, while corresponding estimates from PREML analyses for GSel were somewhat lower. The latter could be attributed to stronger selection in the last three generations for GSel, together with the fact that PREML ignored the genomic information which facilitated it. Conversely, as more and more of the generations subject to selection were omitted from the data (i.e. as the 'starting generation' increased), estimates reflected the reduced genetic variation available in what was implicitly treated as the base generation in the truncated data set.

Estimates from ssGREML were consistently higher than those from PREML for both PSel and GSel and, for analyses including data from unselected generations, were somewhat higher than the population value of 0.3 simulated. Including pedigree information for individuals in starting generation i, we would expect estimates to reflect the genetic variance in base generation $i - 1$. Presumably the overestimates might be attributed, to some extent at least, to the effects of pedigree truncation – and thus underestimates of inbreeding – resulting in imperfect alignment of $\mathbf{G}$ to $\mathbf{A}_{22}$. Limited additional analyses for GSel using data from generations 3 to 13 yielded a mean heritability estimate closer to 0.3, suggesting so.

As shown in Figure 2 for GSel, the higher heritability estimates from ssGREML analyses were mainly due to higher genetic variance estimates. Interestingly, ssGREML estimates of the residual variance did not depend strongly on the amount of data truncation, while values for PREML exhibited distinct repartitioning of genetic into residual variation.

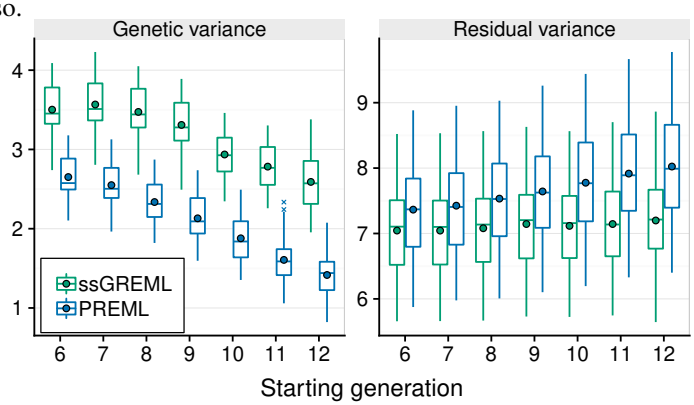

Figure 2. Distribution of variance component estimates over replicates for GSel (see text for definitions)

The distributions of correlations between TBV and EBV and regressions of TBV on EBV for selection candidates in generation 13 are shown in Figure 3. With the simulation involving strong selection and, for ssGREML, all individuals from generation 10 onward having genomic information, mean correlations for ssGREML analyses were very high and substantially exceeded those from PREML, in particular for GSel. For all three scenarios, values for PREML differed little between the subsets of data utilised. Robustness of such correlations, in particular for univariate analyses, is a well known phenomenon for PBLUP. In contrast, means for ssGREML and starting generations 11 and 12 dropped, due to the omission of marker information in these analyses. Mean regressions of TBV on EBV were close to their expected value of unity for all ssGREML analyses. Corresponding values for PREML and PSel or GSel, however, showed increasing underdispersion of EBV (i.e. regression coefficients greater than unity) with increasing starting generation, mirroring the underestimates of genetic variation reported above.

DISCUSSION

Simulation studies on ssGREML have been presented by Cesarani *et al.* (2019) and Junqueira *et al.* (2022) but involved different set-ups and questions considered. Our study attempted to mimic, in a simplified scheme, the progression from random selection to pedigree based and finally to genomic assisted selection which might occur in a livestock improvement programme. Clearly, results are at least partially specific to the scenario considered. In particular, for analyses using genomic information all individuals in the relevant generations were assumed to be genotyped. This yielded substantial differences in estimates of variance components from PREML and ssGREML. Additional ssGREML analyses retaining only genotypes for a proportion of randomly selected animals reduced estimates closer to values from PREML (not shown).

Truncation of data and pedigrees redefines the base generation. This implies that estimates of the genetic variance reflect the amount of 'usable' genetic variation in that generation. Consequently, when

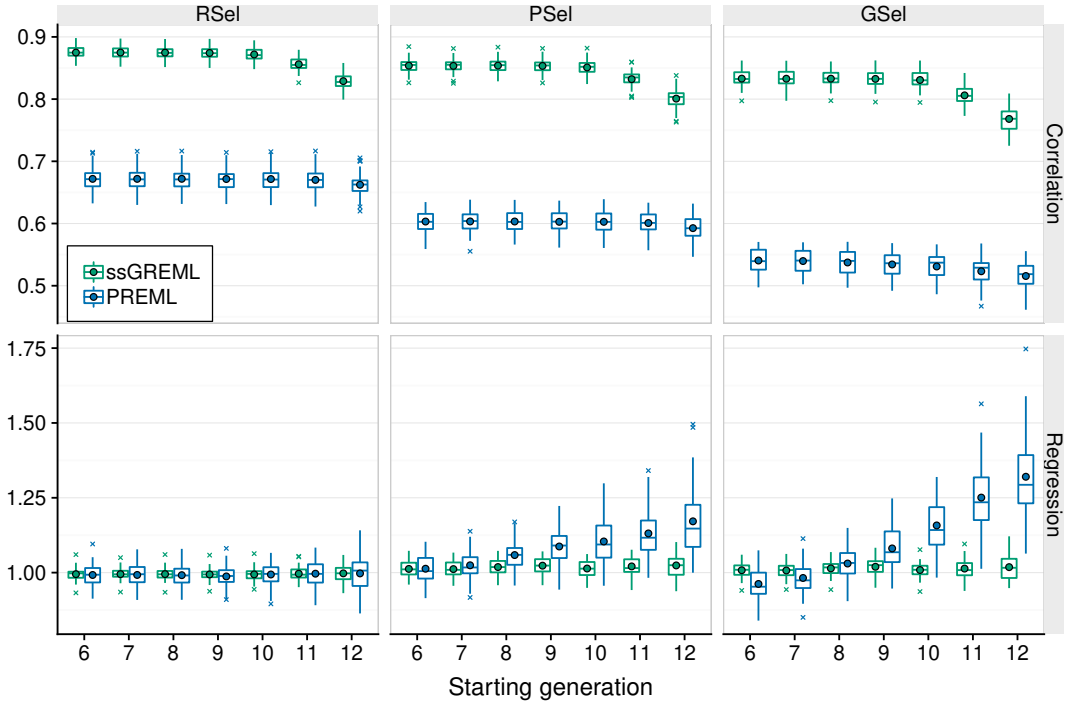

Figure 3. Distribution over replicates of correlations between true and predicted breeding values and regressions of true on predicted breeding values for animals in generation 13 (see text for definitions)

omitting information on which selection decisions have been based, estimates declined, especially for GSel. Implications thereof need to be considered when predicting response to selection or evaluating reliabilities of EBV (Gorjanc *et al.* 2015). As more and more animals are genotyped and as the emphasis on genomic selection increases, ssGREML estimation of genetic parameters will become a necessity.

ACKNOWLEDGEMENTS

This work was supported by Meat and Livestock Australia grant L.GEN.2204.

REFERENCES

Aguilar I., Misztal I., Johnson D.L., Legarra A., Tsuruta S. and Lawlor T.J. (2010) *J. Dairy Sci.* **93**:743.

Cesarani A., Pocrnic I., Macciotta N.P.P., Fragomeni B.O., Misztal I. and Lourenco D.A.L. (2019) *J. Anim. Breed. Genet.* **136**:40.

Faux A.M., Gorjanc G., Gaynor R.C., Battagin M., Edwards S.M., Wilson D.L., Hearne S.J., Gonen S. and Hickey J.M. (2016) *Plant Genome* **9**:1.

Gorjanc G., Bijma P. and Hickey J.M. (2015) *Genet. Sel. Evol.* **47**:65.

Junqueira V.S., Lourenco D., Masuda Y., Cardoso F.F., Lopes P.S., Silva F.F.E. and Misztal I. (2022) *J. Anim. Sci.* **100**:skac082.

Misztal I., Lourenco D. and Legarra A. (2020) *J. Anim. Sci.* **98**:skaa101.

Van Raden P.M. (2008) *J. Dairy Sci.* **91**:4414.

Vitezica Z.G., Aguilar I., Misztal I. and Legarra A. (2011) *Genet. Res.* **93**:357.