Proc. Assoc. Advmt. Anim. Breed. Genet. 23:15-18

MORE GENOTYPES THAN MARKERS: THE SS-T-BLUP MODEL IN ACTION. AN APPLICATION STUDY IN MULTI-TRAIT AUSTRALIAN ANGUS BREEDPLAN GENETIC EVALUATION

V. Boerner and D.J. Johnston

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

SUMMARY

Multi-trait single step genetic evaluation is increasingly facing the situation of having more individuals with genotypes than an individuals' genotype has markers. This leads to an algebraically impossible inversion of the genomic relationship matrix (G). Recent derivations in single step equations called SS-T-BLUP have provided an elegant way to circumvent the inversion of the G and therefore accommodate the described situation. In this paper we examine the applicability of the SS-T-BLUP model to the multi-trait Australian Angus BREEDPLAN genetic evaluation and compare the results to applying two different ways of using G in a single step model. Results clearly show that SS-T-BLUP outperforms other single step formulations and allows users to avoid approximating the inverse of G.

INTRODUCTION

Within the last decade genotyping thousands of individuals with Single Nucleotide Polymorphism (SNP) chips at the commercial level has become common practice in many species of economic relevance. However, due to cost effectiveness these individuals are being genotyped with low to medium density SNP chips, with usually not more than 50,000 markers. To date, genetic evalua-tion systems allow for SNP marker genotypes via the so-called Single Step model (Christensen and Lund 2010). In this model most often markers are used to pre-calculate a marker based relation-ship matrix which subsequently combined with the usual pedigree derived relationship matrix to a so-called *H* matrix (SS-H-BLUP). This requires the inverse of G as well. The described situation of having thousands of individuals genotyped at medium to low density has led to the situation where G is algebraically no longer invertible due to rank deficiencies. A possible solution is to abandon G and move to a model which incorporates the markers directly (SS-SNP-BLUP). While SS-SNP-BLUP is generally equivalent to SS-H-BLUP many of its final implementations suffer from convergence problems with regard to iterative solving or demanding pre-conditioner computation. Recently an elegant intermediate model has been formulated which may be seen as a mix of SS-H-BLUP and SS-SNP-BLUP called SS-T-BLUP (Mäntysaari *et al.* 2017). SS-T-BLUP does not need G nor its inverse and fits the markers directly. As it fits G implicitly, it is algebraically equivalent to SS-H-BLUP under certain assumptions. In addition, it provides EBVs at the individual level which can be readily transformed into marker solutions. In this paper we will examine the effect of SS-T-BLUP on the computational load to a Single Step genetic evaluation of Australian Angus. We will compare the results relative to the ordinary SS-H-BLUP approach.

METHODS

Model. The "H" matrix (Christensen and Lund 2010) required for SS-H-BLUP can be written as

$$\begin{array}{c|c} A_{1,1} - A_i A_{2,2} A_i' + A_i G_w A_i' & A_i G_w \\ \hline G_w A_i' & G_w \end{array} \tag{1}$$

[*] A joint venture of NSW Department of Primary Industries and the University of New England

where 1 is a vector indexing the subset of n_{ng} non-genotyped of individuals, 2 is a vector indexing the subset of n_g genotyped individuals, A is the pedigree-based relationship matrix, $A_i = A_{1,2}A_{2,2}^{-1}$, and G_w is a genomic relationship matrix dimension $n_g \times n_g$ which is constructed from a centred and scaled marker genotypes matrix M of dimension $n_g \times n_m$ and subsequently blended. Thus $G_w = \gamma MM' + \lambda C$, where C is an arbitrary but symmetric matrix and γ and λ are arbitrary non-zero weights. For the sake of simplicity we will set $C = A_{22}$ and $1 = \gamma + \lambda, \gamma > 0,\ \lambda > 0$. H^{-1} can be written as

$$\left(\begin{array}{c|c} A^{1,1} & A^{1,2} \\ \hline A^{2,1} & A^{2,2} \end{array}\right) + \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & G_w^{-1} - A_{2,2}^{-1} \end{array}\right) \tag{2}$$

(Christensen and Lund 2010) or as $\widetilde{H}^{-1}$ (Strandén *et al.* 2017)

$$\left(\begin{array}{c|c} A^{1,1} & A^{1,2} \\ \hline A^{2,1} & A^{2,2} \end{array}\right) + \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & G_w^{-1} - (A^{2,2} - A^{2,1}(A^{1,1})^{-1}A^{2,1}) \end{array}\right), \tag{3}$$

where $A^{\cdot;\cdot}$ is a respective block of the inverse of A. However, replacing G_w with $\gamma MM' + \lambda C$ in equation 1 and inverting the resulting matrix yields matrix Ψ^{-1}

$$\left(\begin{array}{c|c} A^{1,1} & A^{1,2} \\ \hline A^{2,1} & A^{2,2} \end{array}\right) + \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & \lambda^{-1}(A^{2,2} - A^{2,1}(A^{1,1})^{-1}A^{2,1}) \end{array}\right) - \left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & M^*M^{*'} \end{array}\right) \tag{4}$$

where $M^* = M^\dagger(K_u)^{-1}$, $M^\dagger = (\lambda^{-1}(A^{2,2} - A^{2,1}(A^{1,1})^{-1}A^{1,2}))M$, $(K_u)^{-1}$ is an upper triangular matrix derived from $K^{-1} = (K_u)^{-1}(K_u^{'})^{-1}$, $K = (\gamma^{-1}D^{-1} + M'M^\dagger)$ and D^{-1} is the inverse of D which is an arbitrary but symmetric and positive definite matrix of dimension $n_m \times n_m$ (Mäntysaari *et al.* 2017). Further D may contain marker specific weights, or allele frequencies if M is not scaled. Given matrices H^{-1}, $\widetilde{H}^{-1}$ and Ψ^{-1} one can define three different BLUP models, SS-H-BLUP, SS-$\widetilde{\text{H}}$-BLUP, and SS-T-BLUP, which differ solely in which formulation of the inverse of H is used (H^{-1}, $\widetilde{H}^{-1}$ or Ψ^{-1}). However, the different formulations will have consequences for solver preparation and iteration time.

Data. The SS-H-BLUP, SS-$\widetilde{\text{H}}$-BLUP and SS-T-BLUP models were applied to an Australian Angus data set currently used in commercial genetic evaluation. The data set comprised of 35 traits with a total of 9,565,814 records across all traits, and 2,621,403 individuals in the pedigree which allowed for multiple sire mating. The number of animals with genotypes was 58,705 comprising of SNP marker genotypes of various densities and panel manufacturers imputed to a common set of 56009 SNPs. To increase the computational load additional 91,295 genotypes (data set 150k) and 341,295 genotypes (data set 400k) were artificially imputed in a combined regression-sampling approach. The 400k data set was only used for SS-T-BLUP because the other models were computationally infeasible.

The multi-trait model included a single fixed factor per trait, 27 correlated genetic factors, 27 correlated genetic groups factors with 19 genetic groups each, 3 correlated maternal permanent environmental factors and 22 correlated sire-by-herd factors. The total number of equations was 76,823,378. λ and γ were set 0.05 and 0.95, respectively.

Software. The system of equations was solved with AGBU's current large scale linear mixed model library solver which uses the preconditioned gradient algorithm (PCG) for iteratively solving linear mixed models and integrates Intel(R) MKL(R), version 2017 update 8. Convergence was achieved when the L2 norm of PCG residuals scaled by the L2 norm of the mixed model equations' right hand side was $\leq 2.68e^{-9}$. All computationally relevant integer and all real numbers were represented in a 64 bit. Computations for the 150k data set were carried out on a computer with two sockets each carrying an Intel(R) Xeon(R) CPU E5-2697 v3 with 2.60GHz, a total of 28 cores, and 528GB of random access memory (RAM). Computations for the 400k data set were carried out on a computer with two sockets each carrying an Intel(R) Xeon(R) CPU E5-2697 v4 with 2.30GHz, a total of 36 cores, and 256GB of RAM.

RESULTS

Table 1: Processing time in real time seconds (hours) for various steps when iteratively solving a SS-T-BLUP, SS-H-BLUP and SS-$\widetilde{\text{H}}$-BLUP model using an Australian Angus BREEDPLAN dataset

task	SS-H-BLUP$^{1}_{150}$	SS-$\widetilde{\text{H}}$-BLUP$_{150}$	SS-T-BLUP$_{150}$	SS-T-BLUP$^{2}_{400}$
G	1,756	1,756	-	-
$A_{2,2}$	250	250	-	-
G^{-1}	9,150	9,150	-	-
${A_{2,2}}^{-1}$	3,500	-	-	-
$M^{\dagger}$ and K	-	-	3,422	4,210
K_L	-	-	352	320
M^{*}	-	-	629	1170
$A^{-1}_{2,2}$ diag[3]	-	262	262	219
preparation	14,656 (4)	11,418 (3.2)	4,665 (1.3)	5,919(1.6)
iteration	7.5	11.2	8.6	12
$\sum$ iteration	19,123 (5.3)	28,716 (7.9)	22,134 (6.1)	30,809 (8.5)
run time	33,779 (9.4)	40,134 (11.1)	26,799 (7.4)	36,728 (10.2)

1: 150,000 individuals with genotypes. 2: 400,000 individuals with genotypes. 3: sampling of diagonal elements of $A^{-1}_{2,2}$ using 10,000 samples.

Results for the different parts of the setup and solving steps are provided in Table 1. SS-H-BLUP$_{150}$, SS-$\widetilde{\text{H}}$-BLUP$_{150}$, SS-T-BLUP$_{150}$ and SS-T-BLUP$_{400}$ converged in equal number of rounds which was ≃2,560. The major differences between SS-H-BLUP$_{150}$, SS-$\widetilde{\text{H}}$-BLUP$_{150}$ and SS-T-BLUP$_{150}$ are the computation time for run preparation and the computation time per round of iteration. The preparation time for model specific parts for SS-T-BLUP$_{150}$ was 1.3 hours, for SS-H-BLUP$_{150}$ 4 hours and for SS-$\widetilde{\text{H}}$-BLUP$_{150}$ 3.2 hours. Thus, compared to SS-T-BLUP, SS-H-BLUP needed 3 times and SS-$\widetilde{\text{H}}$-BLUP 2.5 times more real time for all necessary pre-calculations. In terms of time per iteration SS-H-BLUP$_{150}$ took 7.5 real time seconds for a single round of the preconditioned gradient solver, followed by SS-T-BLUP$_{150}$ with 8.5 real time seconds. With 11.2 seconds per iteration SS-$\widetilde{\text{H}}$-BLUP was slowest. Due to the huge time savings for run preparation and only a slightly longer time while iterating SS-T-BLUP$_{150}$ needed only 80 % of the total processing time required by SS-H-BLUP$_{150}$ and only 66 % of SS-$\widetilde{\text{H}}$-BLUP$_{150}$. The last column in Table 1 shows the computing time for SS-T-BLUP$_{400}$.

DISCUSSION

SS-T-BLUP has been proposed as a single step model which can facilitate data sets where the number of genotyped individuals exceeds the number of markers and the G matrix is algebraically not invertible. These situations become more common in commercial plant and livestock species where individuals are genotyped with low to medium density SNP chips (Mäntysaari *et al.* 2017). This is achieved by reformulating the "H" matrix representation such that neither the G or $\text{A}_{2,2}$ matrices nor their inverses need to be built or approximated. As shown by the results, SS-T-BLUP clearly outperforms SS-H-BLUP in terms of total processing time which is mainly due to the huge computational cost for setting up G, $\text{A}_{2,2}$ and inverting both as the inversion cost grows cubicly with n_g, whereas at a constant n_m the cost for generating $M^{\dagger}$ grows less than linearly and the cost for K grow $(n_m \times n_m + 1)/2 \times n_g$. In terms of seconds per iteration the main difference between SS-T-BLUP, SS-H-BLUP and SS-$\widetilde{\text{H}}$-BLUP is caused by the operations of Ψ^{-1}, H^{-1} and $\widetilde{H^{-1}}$ times a vector y. This can be narrowed down further to a single matrix vector operation $\Delta H_{2,2}^{-1} y = (G_w^{-1} - A_{2,2}^{-1})y$ in SS-H-BLUP, or one matrix vector operation $\Delta H_{2,2}^{-1} y = G_w^{-1} y$ and one solver operation $y = (A^{2,2} - A^{2,1}(A^{1,1})^{-1}A^{1,2})x$ in SS-$\widetilde{\text{H}}$-BLUP, or two matrix vector operations $M^{\star'} M^{\star} y$ and one solver operation $y = (A^{2,2} - A^{2,1}(A^{1,1})^{-1}A^{1,2})x$ in SS-T-BLUP. In the example given here operations $\Delta H_{2,2}^{-1} y$ and $G_w^{-1} y$ required $\approx 2.25e10$ floating point operations (FLOPs), whereas operation $M^{\dagger'} M^{\dagger} y$ required $\approx 1.5e10$ FLOPs. SS-T-BLUP and SS-$\widetilde{\text{H}}$-BLUP have additional cost for solving $y = (A^{2,2} - A^{2,1}(A^{1,1})^{-1}A^{1,2})x$ which offsets the FLOP advantage of SS-T-BLUP and produce an additional overhead for SS-$\widetilde{\text{H}}$-BLUP. For SS-$\widetilde{\text{H}}$-BLUP these disadvantages whilst iterating are not balanced due to not inverting $\text{A}_{2,2}$, because its inverse can be calculated much quicker than the inverse of G, resulting in almost 20% more total processing time compared to SS-H-BLUP. For SS-T-BLUP the combination of an advantage in terms of FLOPs, extra burden for solving and huge saving in preparation time resulted in a 20% and 33% decrease in processing time compared to SS-H-BLUP and SS-$\widetilde{\text{H}}$-BLUP, respectively.

CONCLUSION

These results support the conclusion that SS-T-BLUP provides a feasible algorithm to calculate exact solutions for estimated breeding values when the number of genotyped individuals exceeds the number of markers. A limitation to the number of genotyped individuals is solely set by the available RAM. Therefore SS-T-BLUP allows solving Single Step equation systems iteratively without generating G or $\text{A}_{2,2}$ or their inverse matrices or any approximation of these matrices.

ACKNOWLEDGEMENTS

The authors thank Angus Australia for providing access to their data. This work was funded by Meat and Livestock Australia (Project L.GEN.0174).

REFERENCES

Christensen O. F. and Lund M. S. (2010) *Genet. Sel. Evol.* **42**:2.

Mäntysaari E. A., Evans R. and Strandén I. (2017) *J. Anim. Sci.* **95**(11):4728.

Strandén I., Matilainen K., Aamand G. P. and Mäntysaari E. A. (2017) *J. Anim. Breed. Genet.* **134**:264.