Comparison of ssGBLUP and ssGTBLUP using Nordic Holstein TD data

M. Koivula[1], I. Strandén[1], G. P. Aamand[2] & E. A. Mäntysaari[1]
[1] Natural Resources Institute Finland (Luke), FI-31600 Jokioinen, Finland
[2] NAV Nordic Cattle Genetic Evaluation, 8200 Aarhus N, Denmark

Minna.Koivula@luke.fi

Summary

The aim of this study is to compare the efficiency of the two different ssGBLUP and two ssGTBLUP approaches in the analysis of Nordic Holstein (HOL) test-day (TD) data, using the official Nordic Holstein TD model. Based on the results, ssGTBLUP had similar convergence properties as the original ssGBLUP. Generally, the solutions from the two different ssGBLUP approaches and the ssGTBLUP approaches were the same. Correlations of GEBVs varied from 0.977 to 1.000. In conclusion, there were only minor differences among different studied single-step approaches in GEBVs or trends or standard deviations. The computational differences become more important as the number of genotyped animals still increase. Then, ssGTBLUP with or without eigendecomposition approach seems to offer a computationally reasonable approach for solving genomic breeding values using the single-step method.

Keywords: ssGBLUP, genomic evaluation, single-step, Holstein

Introduction

Meuwissen *et al.* (2001) introduced the concept of genome-wide marker-assisted selection, after which many alternative methods have been developed to put genomic selection into practice. Currently, genomic selection has been in wide use already for several years. Single-step genomic BLUP (ssGBLUP) is the preferred method for genomic evaluations (Aguilar *et al*., 2010; Christensen & Lund, 2010) since it is a unified approach to calculate GEBVs. The unified relationship matrix **H** used in ssGBLUP defines the relationships among genotyped and non-genotyped animals. The single-step approach as such is computationally demanding with a large number of genotyped animals. Several alternative ways to overcome some of the computational challenges have been presented. Legarra & Ducrocq (2012) presented single-step mixed-model equations (MME) where it is possible to avoid inversion of **G**, or even making it. These were implemented in Fernando *et al.* (2016). Other alternative formulations account for the genomic information through marker effects (e.g. Liu *et al*., 2014; Fernando *et al*., 2014; Taskinen *et al.,* 2017). Misztal *et al.* (2014) and Fragomeni *et al.* (2015) suggested a sparse approximation of the inverse of the **G** matrix with so-called APY (Algorithm for proven and young). Mäntysaari *et al.* (2017) proposed an exact approach named ssGTBLUP where neither the **G** matrix nor its inverse are needed. The ssGTBLUP has the same convergence properties in preconditioned conjugate gradient (PCG) iteration as the original single-step MME but is computationally less demanding.

The aim of this study is to compare the efficiency of the original ssGBLUP and ssGTBLUP approach with Nordic Holstein (HOL) test-day (TD) data, using the official Nordic Holstein TD model.

Material and methods

All analyses used the same data as are used in the official Nordic HOL milk production evaluations. The multiple trait milk production evaluation includes TD records from milk, fat and protein production. Production records from the first three lactations are in the same multiple traits model. The full routine evaluation data from September 2016 for the HOL were obtained from the Nordic Cattle Genetic Evaluation (NAV). For the production traits, the TD data included 7.6 million cows with a total of 153 million records and 9.8 million animals in the pedigree. Genotype data included 101 004 genotyped HOL animals of which 55 658 were bulls, including also Eurogenomics bulls, and 45 346 were cows and heifers. After application of exclusion criteria, 46 342 SNP markers on the 29 bovine autosomes were available for further analysis.

Solving single-step evaluations

Following Mäntysaari *et al.* (2017) the genomic relationship matrix was $\mathbf{G} = \mathbf{ZZ}'+\mathbf{C}$ where $\mathbf{Z}$ is a centered and scaled genotyped matrix by the method I of VanRaden (2008), and the singularity prevention matrix $\mathbf{C}$ was varied. The centering used the estimated base population allele frequencies, which were calculated as described in McPeek *et al.* (2004). Single-step genomic evaluations were performed by four different approaches. In the first approach, the $\mathbf{G}$ matrix was formed and inverted using LAPACK (Anderson *et al.*, 1999) subroutines available in the MKL library (Intel Math Kernel Library Reference Manual, 2014). The singularity prevention was done by adding a diagonal matrix $\mathbf{C} = \mathbf{I}\varepsilon$ where a small number ε was 0.01 (later called as ssGBLUP$_{0.01}$). In the second approach, the $\mathbf{G}$ matrix was formed similarly as in the fi**T**st approach but 10 % of the polygenic variance was included into the $\mathbf{G}$ to overcome singularity (called ssGBLUP$_{\mathrm{w10}}$). In the third approach, the $\mathbf{G}^{-1}$ was replaced by $\mathbf{C}^{-1}$-$\mathbf{T}'\mathbf{T}$, where the matrix was calculated using $\mathbf{C} = \mathbf{I}\varepsilon$, and neither $\mathbf{G}$ nor its inverse was explicitly formed (ssGTBLUP). In the fourth approach, we used eigendecomposition to reduce rank of the $\mathbf{T}$ matrix which reduced the number of multiplications during the PCG. The percentage of total variance explained by the eigenvalues was 98% (method called hereinafter ssGTBLUP$_{(98)}$).

All models were solved using MiX99 software (Strandén & Lidauer, 1999) which uses PCG iteration in solving the MME. The main computational cost in the PCG method is a matrix times a vector product where within each iteration round a so-called direction vector is multiplied by the coefficient matrix. Additional genomic covariance structure was read in from a disk file in each iteration round. The inverse of the $\mathbf{A}_{22}$ matrix was not computed in advance, but instead, the method by Strandén *et al.* (2017) was implemented. There the submatrices $\mathbf{A}^{12}$ and A^{22} of $\mathbf{A}^{-1}$ are formed implicitly using pedigree information. This saves considerably memory and computing time.

Results and Discussion

Based on the results, ssGTBLUP had similar convergence properties as the original ssGBLUP (Table 1). The original ssGBLUP with 10% of polygenic variance needed the most pre-processing time and memory because the $\mathbf{A}_{22}$ matrix had to be formed and inverted when building the external genomic matrix for the MiX99. For the ssGTBLUP$_{(98)}$, the eigendecomposition is more time-consuming than inverting a matrix of the same size. However, the eigendecomposition was needed for a smaller matrix of size 46 342, i.e., number of markers, whereas the inversion in the original ssGBLUP was for the **G** inverse matrix of size 101 004. For the rank reduced ssGTBLUP$_{(98)}$, the number of rows in the T matrix was further reduced from 46 342 to 14 038.

In general, the solutions from the two different ssGBLUP approaches and the two ssGTBLUP approaches were the same. Correlations of GEBVs varied from 0.977 to 1.000 for reference bulls, so-called transition bulls having less than 100 daughters, and for young candidate bulls without offspring (Table 2). The correlation between GEBVs from the ssGTBLUP$_{w10}$ and ssGBLUP$_{0.01}$ was high in all bulls, thus it appears that addition of a small number to the diagonal of the genomic relationship matrix works as well in preventing the singularity of the matrix, but also gives similar GEBVs. As assumed, the correlation between ssGBLUP$_{0.01}$ and ssGTBLUP approaches was one, and the rank reduction employed in ssGTBLUP$_{(98)}$ did not much affect the results.

The genetic trend of protein for the genotyped Nordic Holstein bulls is presented in Figure 1. The figure shows that different single-step approaches do not affect the trends. However, more important is to see what happens to the standard deviations. Thus, Figure 2 presents SDs of the protein GEBVs for bulls. For the reference bulls or transition bulls, there were no differences among the single-step approaches, but for the young bulls, the SDs from the ssGBLUP$_{w10}$ had some differences to the other methods.

In conclusion, it seems that there were not much differences among the studied single-step approaches in GEBVs or trends or standard deviations. The computational differences become more important as the number of genotyped animals increases. Then, ssGTBLUP with or without eigendecomposition approach seems to offer a computationally reasonable approach for solving genomic breeding values using the single-step method.

List of References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta & T.J. Lawlor, 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. J. Dairy Sci. 93:743-752.

Anderson, E., Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenny & D. Sorensen, 1999. LAPACK Users' Guide, 3rd ed. SIAM. http://www.netlib.org/lapack/lug/. Accessed July 11, 2017.

Christensen, O. & M.S. Lund, 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.

Fernando, R.L., H. Cheng, B.L. Golden & D.J. Garrick., 2016. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. Genet. Sel. Evol. 48:96.

Fernando, R.L., Dekkers, J.C.M. & Garrick D.J., 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet Sel Evol. 46:59.

Fragomeni, B. O., D.A.L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T.J. Lawlor, & I. Misztal, 2015. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. J. Dairy Sci. 98:4090-4094.

Legarra, A. & V. Ducrocq, 2012. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. J. Dairy Sci. 95:4629- 4645.

Liu, Z., M. E. Goddard, F. Reinhardt & R. Reents, 2014. A single-step genomic model with direct 442 estimation of marker effects. J. Dairy Sci. 97:5833-5850. McPeek M. S., X. Wu & C. Ober, 2004. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics 60: 359–367.

Meuwissen, T.H.E., B.J. Hayes & M.E. Goddard, 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157: 1819–1829.

Misztal, I., A. Legarra & I. Aguilar, 2014. Using recursion to compute the inverse of the genomic 455 relationship matrix. J. Dairy Sci. 97:3943–3952.

Mäntysaari, E.A., R. D. Evans, I. Strandén, 2017. Efficient single-step genomic evaluation for a multi-breed beef cattle population having many genotyped animals. J. Anim. Science, in press.

Strandén, I. & M. Lidauer, 1999. Solving large mixed models using preconditioned conjugate gradient iteration. J. Dairy Sci. 82:2779-2787.

Strandén, I, K. Matilainen, G.P. Aamand & E.A. Mäntysaari, 2017. Solving efficiently large single-step genomic best linear unbiased prediction models. J. Anim. Breed. Genet. 134: 264–274.

Taskinen, M., E.A. Mäntysaari & I. Strandén, 2017. Single-step SNP-BLUP with on-the-fly imputed genotypes and residual polygenic effects. Genet. Sel. Evol. 49:36.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414-4423.

*Table 1. Computing times and peak memory needed for the original single-step genomic model with 10% of polygenic variance (ssGBLUP$_{w10}$), the single-step using addition of 0.01 to G-diagonal (ssGBLUP$_{0.01}$), single-step using the **T** matrix approach (ssGTBLUP), or using eigendecomposition with rank reduction in ssGTBLUP$_{(98)}$ with 98% of variance explained. Peak memory needed in gigabytes in the **G/T** matrix building. Wall clock time in hours for the preprocessing (Pw) and time per 1000 iterations in hours (I), size of external matrix read by the solver in gigabytes (Matrix), number of iterations (N). In preprocessing, 10 processors were used when making the required matrices for ssGBLUP and ssGTBLUP.*

Method	Peak memory (GB)	Pw (hour)	Matrix (GB)	I (h/1000 iterations)	N (Number)
ssGBLUP$_{w10}$	152.9	5.3	20	28	4202
ssGBLUP$_{0.01}$	114.7	2.2	20	25	4891
ssGTBLUP	51.8	1.5	18	25	4881
ssGTBLUP$_{(98)}$	86.6	4	5.3	18	5576

*Table 2. Correlations among GEBVs of genotyped Nordic Holstein bulls from single-step genomic models presented separately for reference bulls, "transition bulls" with less than 100 daughters and for young bulls without daughters. The original single-step with 10% of the polygenic variance (ssGBLUP$_{w10}$), the single-step using the addition of 0.01 to diagonal (ssGBLUP$_{0.01}$), single-step using the **T** matrix approach (ssGTBLUP), eigendecomposition with a rank reduction in ssGTBLUP$_{(98)}$ with 98% of variance explained.*

Bull	Method	ssGBLUP$_{0.01}$	ssGTBLUP	ssGTBLUP$_{(98)}$
Reference bulls	ssGBLUP$_{w10}$	0.999	0.999	0.998
	ssGBLUP$_{0.01}$		1.000	0.999
	ssGTBLUP			0.999
Transition bulls	ssGBLUP$_{w10}$	0.998	0.998	0.996
	ssGBLUP$_{0.01}$		1.000	0.999
	ssGTBLUP			0.999
Young bulls	ssGBLUP$_{w10}$	0.980	0.980	0.977
	ssGBLUP$_{0.01}$		1.000	0.998
	ssGTBLUP			0.998

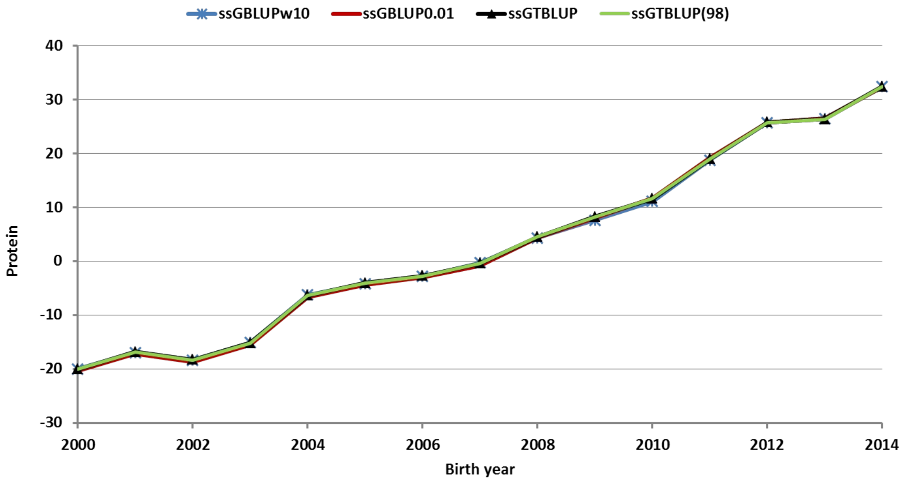

*Figure 1. Genetic trends by birth year for protein from different single-step approaches. Single-step genomic model with 10% of polygenic variance (ssGBLUP$_{w10}$), the single-step using addition of 0.01 to diagonal (ssGBLUP$_{0.01}$), single-step using the **T** matrix approach (ssGTBLUP), eigendecomposition with rank reduction in ssGTBLUP$_{(98)}$ with 98% of variance explained.*

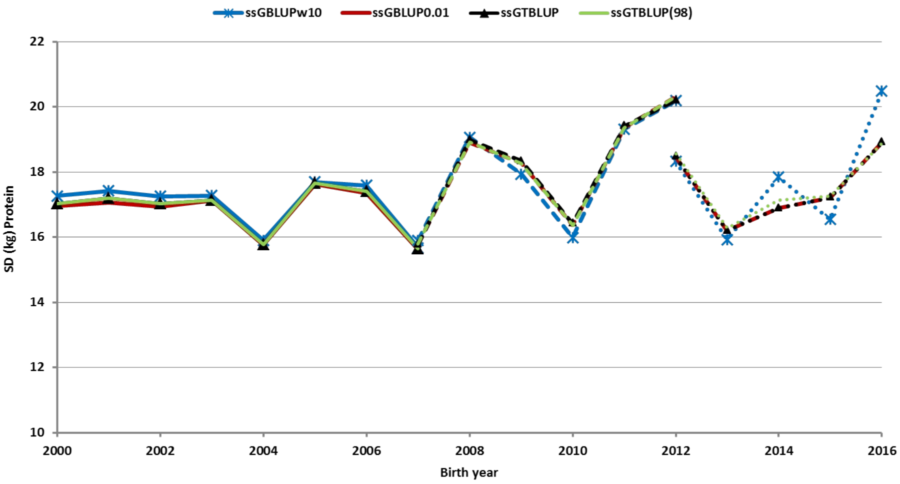

Figure 2. The standard deviation of protein GEBVs by birth year from different single-step approaches. Single-step genomic model with 10% of polygenic variance (ssGBLUP$_{w10}$), the single-step using addition of 0.01 to diagonal (ssGBLUP$_{0.01}$), single-step using the ***T*** *matrix approach (ssGTBLUP), eigendecomposition with rank reduction in ssGTBLUP$_{(98)}$ with 98% of variance explained. Solid lines for reference bulls, dashed lines for "transition bulls" with less than 100 daughters and dotted lines for young bulls without daughters.*