Genetic Evaluation using Unsymmetric Single Step Genomic Methodology with Large Number of Genotypes

***I. Aguilar*[1]*, A. Legarra,*[2] *S. Tsuruta*[3] *and I. Misztal*[3]**

[1] Instituto Nacional de Investigación Agropecuaria, INIA Las Brujas, 90200 Uruguay
[2] INRA, UR631 SAGA, BP 52627, 32326 Castanet-Tolosan, France
[3] Department of Animal and Dairy Science, University of Georgia, Athens, GA, 30602, USA

Abstract

The single step genomic methodology provides a unified framework to integrate phenotypic, pedigree and genomic information in the prediction of breeding values. Minimal modifications of current softwares are necessary in order to incorporate extra relationship matrices, however computing such matrices has a cubic cost. Recently, a system of equations relaxing the computing cost of creating the inverse of the genomic relationship matrix was presented, which creates an unsymmetric system of equations. Bi Conjugate Gradient Stabilized solvers (BiCGSTAB) were proposed to solve unsymmetric system of equations and also can be used with iteration on data programs, resulting in a good choice for solving large-scale genetic evaluations. Here we describe the implementation of a large genetic evaluation using unsymmetric solvers within the iteration on data framework. Comparison with the regular single-step methodology is presented and the effects of different preconditioners and data structures on the convergence pattern were studied. A large scale genetic evaluation was feasible, however required more rounds to get convergence compared with the regular single-step. More sophisticated preconditioners are necessary to improve the convergence for solving unsymmetric single-step genomic evaluations.

Key words: single-step, genomic selection, genetic evaluation, BiCGSTAB

Introduction

The Single Step GBLUP (SSGBLUP; Aguilar *et al.,* 2010; Christensen *et al.,* 2010) is an alternative to combine pedigree (for all individuals), marker and phenotype information in a coherent framework for genetic evaluations and marker effect estimation (Wang *et al.,* 2012).

This unified approach modifies the pedigree-based relationship matrix to include a genomic relationship matrix, and the resulting mixed model equations involve the regular inverse of the numerator relationship matrix, the inverse of the genomic relationship matrix and the inverse of the pedigree-based relationship matrix for genotyped individuals (Aguilar *et al.,* 2010; Christensen *et al.,* 2010).

Adding such extra relationship matrices to current software for genetic evaluation and variance component estimation results in the application of genomic information in a broad kind of models and species (Misztal *et al.,* 2010).

Although the computation of the inverses of such matrices has a cubic cost regarding the number of genotyped individuals, efficient methods were presented (Aguilar *et al.,* 2011). Moreover, the use of methods that exploit GPU cards make such computations feasible for hundreds of thousands of genotyped individuals (Masuda *et al.,* 2013; Coffey *et al.,* 2011).

With the current implementation of genomic computations (Aguilar *et al.,* 2011), SSGBLUP can support very large genotypes indirectly if the genotyped animals are decomposed into proven animals and an arbitrary number of young animals. In such a case, the initial computing includes only genotypes of proven animals. GEBV of these animals are subsequently converted to SNP effects (Wang *et al.,* 2012), which in turn provide DGV for young animals. The resulted DGV may have to be blended with PA, but DGV≈GEBV when the number of proven animals is high (Su *et al.,* 2012)

Recently, Legarra & Ducrocq (2012) suggested the use of an alternative system of equations to implement SSGBLUP, where there is no need of the inverse of such relationship matrices, however creates an unsymmetric system of equations which can deal with large data sets as those known in dairy cattle for some countries. These equations are, for the most general case (multiple-trait model, multiple correlated random effects as in random regression or maternal effects models) as follows:

$$\begin{bmatrix} \mathbf{X'R^{-1}X} & \mathbf{X_1'R^{-1}W_1} & \mathbf{X_2'R^{-1}W_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{W_1'R^{-1}X_1} & \mathbf{W_1'R^{-1}W_1 + A^{11} \otimes G_0^{-1}} & \mathbf{W_1'R^{-1}W_2 + A^{12} \otimes G_0^{-1}} & \mathbf{0} & \mathbf{0} \\ \mathbf{W_2'R^{-1}X_2} & \mathbf{W_2'R^{-1}W_1 + A^{21} \otimes G_0^{-1}} & \mathbf{W_2'R^{-1}W_2 + A^{22} \otimes G_0^{-1}} & \mathbf{I \otimes G_0^{-1}} & \mathbf{-I \otimes G_0^{-1}} \\ \mathbf{0} & \mathbf{0} & \mathbf{I \otimes G_0^{-1}} & \mathbf{A_{22} \otimes G_0^{-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I \otimes G_0^{-1}} & \mathbf{0} & \mathbf{G \otimes G_0^{-1}} \end{bmatrix} = \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ -\hat{\varphi} \\ -\hat{\gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{W_1'R^{-1}y_1} \\ \mathbf{W_2'R^{-1}y_2} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}$$

where G is a “genomic” relationship matrix (VanRaden, 2008), possibly after some “tuning” to refer to the same genetic base of the pedigree (Vitezica *et al.*, 2011),

$\mathbf{A} = \begin{bmatrix} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{bmatrix}$ is the pedigree-based

relationship matrix and $\mathbf{G}_0$ is the genetic covariance across traits.

The advantage of this formulation of the Mixed Model Equations for the SSGBLUP is that it does not require inversion of either G or A_{22}, which was the case in Aguilar *et al.* (2010) and Christensen & Lund (2010). Further, for iteration on data methods, where matrix-vector products are need, neither construction of **G** or $\mathbf{A}_{22}$ is necessary because the matrix-vectors products involving $\mathbf{A}_{22}$ and **G** can be efficiently computed using Colleau (2002) for the first or the form $\mathbf{G}x = \mathbf{Z}(\mathbf{Z}'x)/k$ for the second for any *x* vector.

Solve of the unsymmetric equations can be done by the iterative method BiCGSTAB (Van Der Vorst, 1992) as explained by Misztal *et al.* (2009) and Legarra & Ducrocq (2012). However these MME had not yet been tested for large data sets.

The objective of the present study was to implement the unsymmetric SSGBLUP and test in a large-scale genetic evaluation.

Materials and Methods

Data were US Holstein information for final score as used in Aguilar *et al.* (2010). A total of 10 466 066 records were available for 6 232 548 cows. Pedigrees were available for 9 100 106 animals. Genotypes for 6 508 bulls were generated using the Illumina BovineSNP50 BeadChip (Illumina, San Diego, CA) and DNA from semen contributed by US and Canadian AI organizations to the Cooperative Dairy DNA Repository (Beltsville, MD); genotypes were provided by the Animal Improvement Programs Laboratory, Agricultural Research Service, USDA (Beltsville, MD).

A preconditioned conjugate gradient algorithm using iteration on data program BLUP90IOD (Tsurura *et al.*, 2001) from the BLUPF90 package (Misztal *et al.*, 2002) was modified to solve the unsymmetric MME as in Legarra & Ducrocq (2012) by the BiCGSTAB method (Van der Vorst, 1992). The latter uses iteration of data and therefore has the ability to solve very large systems of equations. Multiplications products involving $\mathbf{A}_{22}$ and **G** were implemented using pre-computed **G** and $\mathbf{A}_{22}$ with optimized BLAS matrix-vector subroutines from Intel Math Kernel Library (MKL). A more memory-wise implementation would compute the matrix-vector products without explicit set up of matrices as described before.

Tests included the symmetric SSGBLUP as in Aguilar *et al.*, (2010) and also the improved BiCGSTAB algorithm (BiCGSTAB(l)), proposed by Sleijpen & Fokkema (1993) for better convergence behavior. Also different preconditioners that account for the non-symmetric system of equations were compared.

Results & Discussion

To monitor convergence the squared ratio of the norm of residual and right-hand side vectors was used, and the iterations were stopped when the criteria was below 10^{-12}. All solvers converged to the same solutions. Table 1 shows statistics of estimated breeding values

(EBV) computed using the regular SSGLUP (PCG) or the unsymmetric solvers BiCGSTAB and BiCGSTAB(l). Correlations of EBV between methods were > 0.999.

Table 1. Statistics of estimated breeding values for different solvers.

	PCG	BiCGSTAB	BiCSTAB(l)
Minimum	-6.1	-6.1	-6.1
1^{st} Quan.	3.0	3.2	3.0
Median	4.5	4.7	4.5
Mean	4.5	4.7	4.6
3^{rd} Quan.	6.2	6.4	6.2
Maximun	11.9	12.1	11.9

Figure 1 shows the convergence for the regular SSGBLUP (PCG) and for the two unsymmetric solvers (BiCGSTAB and BICGSTAB(l)). Unsymmetric solvers took more rounds to get convergence compared to the symmetric system using H^{-1}. Although the BiCGSTAB(l) took fewer iterations, the number of matrix-vector multiplications of such system is twice that of BiCGSTAB so the total computing time for both unsymmetric solvers was very similar.

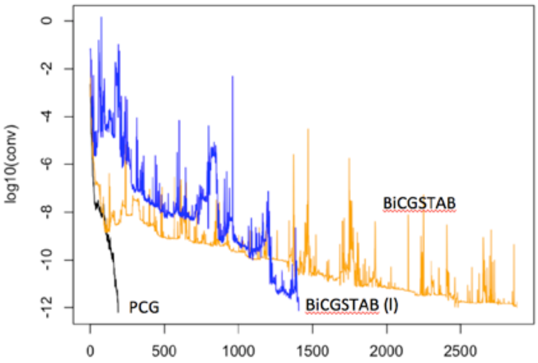

Figure 1. Convergence for different solvers.

Conjugate gradient solvers are greatly affected by a preconditioner. Figure 2 shows the effect of using different preconditioners on convergence pattern. Using an unsymmetric preconditioner improves the convergence pattern and accelerates convergence.

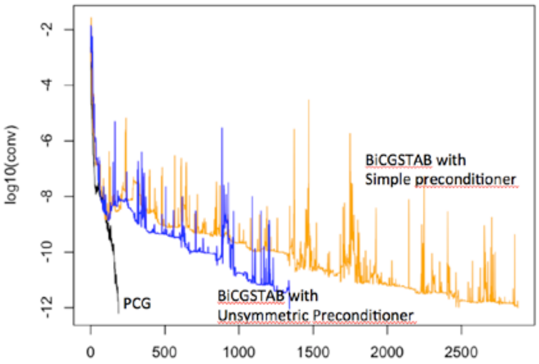

Figure 2. Effect of preconditioner in unsymmetric solvers on convergence.

A subset of the final score records was used to mimic a data structure with young sires without progeny information. Final scores records from 2005 through 2009 were removed, resulting in 2575 young sires. Figure 3 shows the effect of the data structure on convergence. For both methods addition of young sires decreased convergence. Including genotypes of young sires directly in analyses is not essential because their EBV's can be obtained from SNP solutions, as stated earlier.

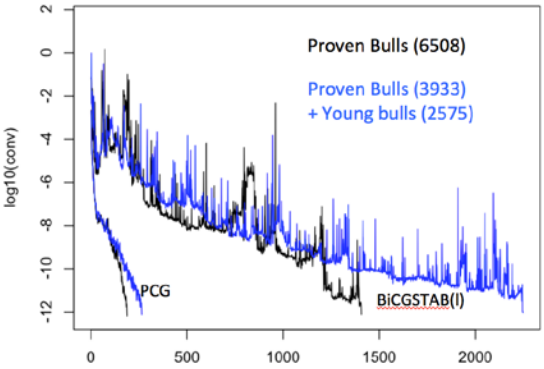

Figure 3. Effect of data structure on convergence.

Conclusions

Large scale genetic evaluation using the unsymmetric equations for a single-step GBLUP that do not require inverses of genomic relationship matrices was successfully implemented. The convergence of the unsymmetric solvers was slower than that of the regular equations, and was affected by the preconditioner and the data structure. A good understanding of convergence criterion and a more sophisticated preconditioner is necessary for the BiCGSTAB solvers for the unsymmetric SSGLUP.

Acknowledgements

The authors thank as Holstein Association USA Inc. (Brattleboro, VT) and the Cooperative Dairy DNA Repository (Beltsville, MD) for providing performance and genotypic data, respectively. A. Legarra thanks financing from FUI project Roquefort'In and SelGen metaprogram action GenSSeq. I. Misztal acknowledges support from the Holstein association, AFRI Grants no. 2010-65205-20366 from USDA NIFA, and BARD Grant no. IS-4394-11.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S. & Lawlor, T.J. 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.

Aguilar, I., Misztal, I., Legarra, A. & Tsuruta, S. 2011. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J Anim Breed Genet. 128,* 422-428.

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

Coffey, M., Mrode, R. & Krzyzelewski, T. 2011. The use of gpus in genomic data analysis. *Interbull Bulletin 44,* 114-116.

Colleau, J.J. 2002. An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol. 34,* 409–421.

Legarra, A. & Ducrocq, V. 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.

Masuda, Y. & Suzuki, M. 2013. Efficient inversion of a large genomic relationship matrix stored on a disk using a multi-core processor and graphic processing units. *J Dairy Sci. 96,* 622.

Misztal, I., Aguilar, I., Legarra, A., Tsuruta, S., Johnson, D.L. & Lawlor, T.J. 2010. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation. Commun. No. 50 in *9th WCGALP* Leipzig, Germany.

Misztal, I., Legarra, A. & Aguilar, I. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci. 92,* 4648–4655.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T. & Lee, D.H. 2002. BLUPF90 and related programs. Commun. No. 28–07 in *7th WCGLP,* Montpellier, France.

Sleijpen, G.L. & Fokkema, D.R. 1993. Bicgstab (l) for linear equations involving unsymmetric matrices with complex spectrum. *ETNA. 1,* 2000.

Su, G., Madsen, P., Nielsen, U.S., Mäntysaari, E.A., Aamand, G.P., Christensen, O.F. & Lund, M.S. 2012. Genomic prediction for nordic red cattle using one-step and selection index blending. *J. Dairy Sci. 95,* 909-917.

Tsuruta, S., Misztal, I. & Stranden, I. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim. Sci. 79,* 1166–1172.

Van der Vorst, H. 1992. Bi-CGSTAB: A fast and smoothly converging variant of bi-cg for the solution of nonsymmetric linear systems. *SIAM J. Sci. Stat. Comput. 13,* 631–644.

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

Vitezica, Z.G., Aguilar, I., Misztal, I. & Legarra, A. 2011. Bias in genomic predictions for populations under selection. *Genetics Research 93,* 357-366.

Wang, H., Misztal, I., Aguilar, I., Legarra, A. & Muir, W.M. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genetics Research 94,* 73-83.