Comparison of Some Equivalent Equations to Solve Single-Step GBLUP

I. Strandén and E.A. Mäntysaari
Biotechnology and Food Research, MTT Agrifood Research Finland, 31600 Jokioinen, Finland

ABSTRACT: We present an equivalent equation system for solving single-step genomic BLUP that does not require separate forming or inversion of the pedigree relationship matrix of genotyped animals. The equation system is contrasted with original single-step equations and two augmented equation systems. Comparison was based on a small data having 73,579 animals of which 2,885 were genotyped. Number of unknowns to solve was 73,580, 79,350, 76,465 and 144,274 by original, full augmented, partly augmented, and our approach, respectively. Numbers of non-zeros in the coefficient matrix were c. 8.9, 12.9, 12.9, and 9.2 million, respectively. Hence, our approach increased substantially number of unknowns but the coefficient matrix remained sparse. Our equivalent system needed more iterations than the original single-step but was competitive with the augmentation methods.

Keywords: breeding value; dairy cattle; genomic evaluation

Introduction

In coming years the genomic information has to be included in national breeding value evaluation along the traditional evaluation. This can be achieved through single-step genomic BLUP or ssGBLUP (Aguilar et al. (2010), Christensen and Lund (2010)). The originally introduced mixed model equations (MME) for ssGBLUP involve inverses of two dense matrices of the size of number of genotyped animals. The dense matrices are genomic relationship matrix **G** and pedigree based relationship matrix of genotyped animals $\mathbf{A}_{22}$. In practice, need for the large inverted matrices in ssGBLUP obstruct using the method when population has many genotyped animals. The inversions can be avoided by augmenting the original MME to equivalent systems of equations (Legarra and Ducrocq (2010)). However, these equations have some undesirable properties. For example, these equivalent equations tend to have slower convergence by iterative methods than with the original MME (Legarra and Ducrocq (2010), Aguilar et al. (2013)). In addition, the equivalent equations still require **G** and $\mathbf{A}_{22}$ which have a size number of genotyped animals. Another approach is to avoid building these matrices altogether. Faux and Gengler (2013) used pedigree information directly to approximate inverse of $\mathbf{A}_{22}$. Legarra and Ducrocq (2010) discussed an approach which does not augment the MME of ssGBLUP but still does not require inversion of **G** and $\mathbf{A}_{22}$. This can be achieved in iterative solving method by preconditioned conjugate gradient (PCG). Each PCG iteration requires coefficient matrix times vector product which involves solving $\mathbf{A}_{22}\,\mathbf{x} = \mathbf{d}$ and $\mathbf{G}\ \mathbf{z} = \mathbf{d}$ that can be solved iteratively by PCG. Hence, convergence characteristics of the original MME of ssGBLUP are unchanged. However, the iterative solving by PCG within PCG iteration can be computationally expensive. An alternative ssGBLUP approach is to model genetic marker effects directly for the genotyped animals and use genotype marker matrix without building **G**. Such approaches have been presented by for example Legarra and Ducrocq (2010), and Liu et al. (2013).

We will concentrate on avoiding computing $\mathbf{A}_{22}$. The objective of this study is to present ssGBLUP that does not require making matrix $\mathbf{A}_{22}$ nor its inverse. Instead, the full pedigree based $\mathbf{A}^{-1}$ is used in an equivalent system of equations. We use a small data and ssGBLUP model to illustrate performance of the approach, and compare it with the equations from original single-step and two variants presented in Legarra and Ducrocq (2010).

Materials and Methods

Equivalent equations. Consider a univariate ssGBLUP model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wa} + \mathbf{e},$$

where incidence matrix **X** relates fixed effects **b**, and incidence matrix **W** relates breeding values **a** to appropriate observation in vector **y**, and **e** is random residual vector. We assume that Var(**e**) = $\mathbf{R}\sigma_e^2$ where **R** is positive definite matrix such as **I** or diagonal matrix with weights. For the sake of presentation simplicity, we let **R**=**I**. In ssGBLUP, the covariance structure for the breeding values is Var(**u**) = $\mathbf{H}\sigma_a^2$ where σ_a^2 is the genetic variance and **H** has both pedigree (**A**) and genomic (**G**) relationship matrix information (Aguilar et al. (2010), Christensen and Lund (2010)).

Mixed model equations for the ssGBLUP are

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'W} \\ \mathbf{W'X} & \mathbf{W'W} + \lambda\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \end{bmatrix} \qquad [1]$$

where λ equals variance ratio σ_e^2/σ_a^2. The mixed model equations require $\mathbf{H}^{-1}$. We divide animals to two groups: non-genotyped animals are in group 1, and genotyped animals are in group 2. All vectors and matrices can be partitioned by these groups. Pedigree based relationship matrix and its inverse can be presented as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} \text{ and } \mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix}.$$

We have

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - (\mathbf{A}_{22})^{-1} \end{bmatrix}.$$

Consider the full pedigree based relationship matrix $\mathbf{A}^{-1}$. According to matrix algebra, we can use absorption of the non-genotyped animals to the genotyped animals to calculate inverse to $\mathbf{A}_{22}$:

$$\left(\mathbf{A}_{22}\right)^{-1} = \mathbf{A}^{22} - \mathbf{A}^{21}\left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12}$$

This can be rearranged to give:

$$\mathbf{A}^{22} - \left(\mathbf{A}_{22}\right)^{-1} = \mathbf{A}^{21}\left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12}$$

Substituting this equality to the equation of $\mathbf{H}^{-1}$ gives

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{G}^{-1} + \mathbf{A}^{21}\left(\mathbf{A}^{11}\right)^{-1}\mathbf{A}^{12} \end{bmatrix}.$$

We can use matrix augmentation methods as in Legarra and Ducrocq (2010) to form equations [2]:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}_1'\mathbf{W}_1 & \mathbf{X}_2'\mathbf{W}_2 & \mathbf{0} \\ \mathbf{W}_1'\mathbf{X}_1 & \mathbf{W}_1'\mathbf{W}_1 + \lambda\mathbf{A}^{11} & \lambda\mathbf{A}^{12} & \mathbf{0} \\ \mathbf{W}_2'\mathbf{X}_2 & \lambda\mathbf{A}^{21} & \mathbf{W}_2'\mathbf{W}_2 + \lambda\mathbf{G}^{-1} & -\lambda\mathbf{A}^{21} \\ \mathbf{0} & \mathbf{0} & -\lambda\mathbf{A}^{12} & -\lambda\mathbf{A}^{11} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}}_1 \\ \hat{\mathbf{a}}_2 \\ \hat{\mathbf{c}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}_1'\mathbf{y}_1 \\ \mathbf{W}_2'\mathbf{y}_2 \\ \mathbf{0} \end{bmatrix}$$

This system of equations has no inverse of $\mathbf{A}_{22}$ but instead sub-matrix of $\mathbf{A}^{-1}$. Number of unknowns has been increased by number of non-genotyped animals. Equations by Legarra and Ducrocq (2010) increase number of unknowns by number of genotyped animals and use sub-matrix of the pedigree relationship matrix $\mathbf{A}_{22}$ [3]:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}_1'\mathbf{W}_1 & \mathbf{X}_2'\mathbf{W}_2 & \mathbf{0} \\ \mathbf{W}_1'\mathbf{X}_1 & \mathbf{W}_1'\mathbf{W}_1 + \lambda\mathbf{A}^{11} & \lambda\mathbf{A}^{12} & \mathbf{0} \\ \mathbf{W}_2'\mathbf{X}_2 & \lambda\mathbf{A}^{21} & \mathbf{W}_2'\mathbf{W}_2 + \lambda\mathbf{A}^{22} + \lambda\mathbf{G}^{-1} & \lambda\mathbf{I} \\ \mathbf{0} & \mathbf{0} & \lambda\mathbf{I} & \lambda\mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}}_1 \\ \hat{\mathbf{a}}_2 \\ \hat{\mathbf{s}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}_1'\mathbf{y}_1 \\ \mathbf{W}_2'\mathbf{y}_2 \\ \mathbf{0} \end{bmatrix}$$

Data. Study data was based on production trait evaluation of Nordic Red dairy cattle. There were 73,579 animals in the pedigree of which 2,885 were genotyped. Genomic relationship matrix was computed according to method 1 in VanRaden (2008). The only fixed effect was general mean because the evaluated trait was deregressed proof. For more information on the data set see Taskinen et al. (2013).

Statistical analyses. We used the following 4 equations to estimate breeding values:

H: standard single-step mixed model equations [1]

LD: Legarra and Ducrocq equations where the matrix augmentation is for both **G** and $\mathbf{A}_{22}$

LDA: Legarra and Ducrocq equations [3] where the matrix augmentation is for $\mathbf{A}_{22}$, but $\mathbf{G}^{-1}$ is used

SMA: our augmented system of equations [2]

These equations were solved by two iterative methods: preconditioned conjugate gradient (PCG) and Bi conjugate gradient stabilized (BiCGSTAB) methods. These iterative methods are used in Octave with sparse matrix implementation. The Octave subroutines were from Barrett et al. (1994). Convergence statistic was the squared ratio of the norm of residual and right-hand side vectors. The four equation systems were iterated until convergence statistic reached criteria of 10^{-14}. The approaches were contrasted by the number of equations, number non-zero elements in the coefficient matrix, estimate to the condition number of the coefficient matrix and number of iterations until convergence.

Condition number of coefficient matrix is a measure for ill-condition of matrix: the larger the condition number the more ill-conditioned the matrix. In general, the more ill-conditioned the coefficient matrix the more iterations are needed for convergence, and the less reliable are the solutions to a linear system of equations. Note that condition number is a property of matrix, not iterative method. Because of large size of the coefficient matrix, condition number was estimated using Octave function condest.

Results and Discussion

Mixed model equations. Table 1 shows that the original single-step method H has the least number equations, number of non-zero elements and the lowest condition number estimate. All these numbers were as expected. Both equations of LD and LDA had slightly more equations due to the small number of genotyped animals. However, number of non-zero elements was about 46% more than in the original single-step. Even worse is that the condition number estimate is about 15 times higher. Our SMA approach had the most number of equations but number of non-zero elements was only about 4% more than in the original single-step. This will translate to reasonable computing time by iteration because number of computations in the coefficient matrix times vector is function of number of non-zero coefficients. In practice, iteration on data approaches do not build the coefficient matrix and, hence, these numbers of non-zero elements do not exactly translate to number of computations in an iteration on data algorithm. However, they still give reasonable estimate for the required number of computations. Condition number of our system of equations was almost 3 times more than by the original single-step which means worse convergence by an iterative solver. Still, this number is much less than by LD and LDA.

Table 1. Number of equations (NE), number of non-zero elements (NZ), estimated condition number (CE), number of iterations by PCG (NP), and number of iterations by BiCGSTAB (NBi) by equation system. System of equations was the original single-step (H), Legarra and Ducrocq (LD), Legarra and Ducrocq with $\mathbf{G}^{-1}$ (LDA), and our approach (SMA).

Class¥	NE	NZ	CE[1]	NP	NBi
H	73,580	8,850,442	2.8	50	39
LD	79,350	12,925,852	44.4	414	293
LDA	76,465	12,914,243	36.7	178	161
SMA	144,274	9,242,362	7.9	260	141

[1] Values divided by 10^6.

Iterative solving. PCG iteration of all equation systems converged towards correct solutions. However, this is not always guaranteed when the coefficient matrix is not positive definite which is the case for LD and SMA. PCG iteration is the preferred solving method when the coefficient matrix is symmetric and positive definite. Number of iterations to reach convergence varied much between the methods (Table 1). The original and the LDA required the lowest and the second lowest number of PCG iterations, respectively. SMA needed lower number of PCG iterations than LD. BiCGSTAB can be used for any matrix. When BiCGSTAB method was used, SMA and LDA equations needed about the same number of iterations which was almost 4 times more than by original ssGBLUP equations (Table 1). LD equations required almost 8 times more iterations to reach convergence.

Figure 1 illustrates convergence by PCG iteration. The positive definite systems show nice convergence but the two non positive definite systems LDA and SMA had more erratic convergence. BiCGTAB iteration behaves better for LDA and SMA as seen in Figure 2. SMA and LD are now quite close in terms of convergence statistic.

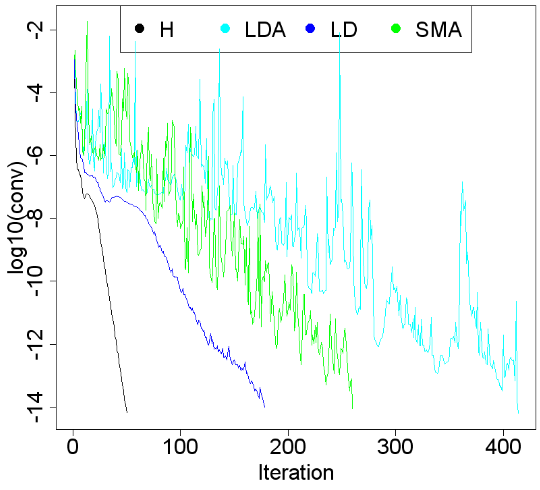

Figure 1: Convergence statistics of the equations systems by PCG iteration.

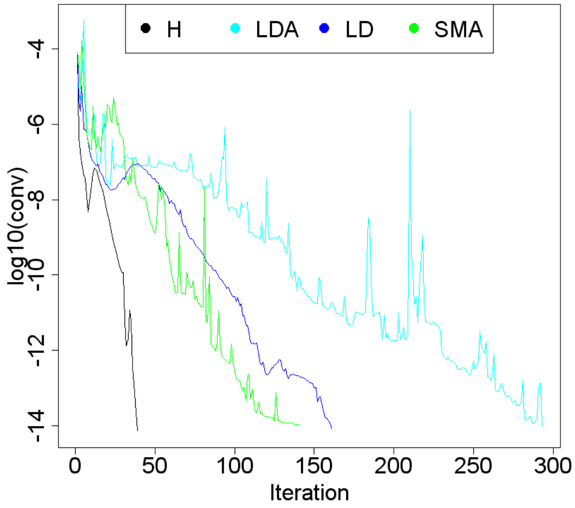

Figure 2: Convergence statistics of the equations systems by BiCGSTAB iteration.

Discussion. Number of iterations to convergence by SMA and LD were about the same by BiCGSTAB. However, SMA had almost twice as many unknowns to solve than LDA. Increase in number of unknowns often predict well increase in number of iterations but in this case the good structure of LDA coefficient matrix was able to diminish this effect. Moreover, when number of genotyped animals increases, number of unknowns in SMA will decrease but in LD and LDA it will increase. In addition, the SMA equations are easier to solve by iteration on data algorithm because the familiar rules used to make $\mathbf{A}^{-1}$ can be used in the augmented part of the coefficient matrix.

In order to avoid the inverse of **G**, it is possible to transform mixed model equations such that that the transformed equations do not need $\mathbf{G}^{-1}$ but **G** (Henderson (1984)). There is a symmetric coefficient matrix version that is based on left and right multiplication of the MME by a matrix having **G**, and a non-symmetric version. We tested the symmetric version. However, the iterative methods failed to reach the quite strict convergence criterion within 1000 iterations although solutions were quite close to true solutions. Hence, this system of equations may have poor convergence properties.

Conclusion

We showed an equivalent equation system for solving ssGBLUP that does not require separate forming or inversion of the pedigree relationship matrix for genotyped animals. In analysis of a small data, the system of equations had more unknowns to solve but had reasonable convergence properties. Coefficient matrix by the equivalent model had nice sparseness which means fast computing time in iteration on data algorithms.

Literature Cited

Aguilar, I., Legarra, A., Tsuruta, S., et al. (2013). Interbull Bulletin No. 47: 222-225.

Aguilar, I., Misztal, I., Johnson, D.L., et al. (2010). J. Dairy Sci. 93:743-752.

Barrett, R., Berry, M., Chan, T.F., et al. (1994). Templates for the solution of linear systems. SIAM. Philadelphia, PA.

Christensen, O., Lund, M.S. (2010). Genet. Sel. Evol. 42:2.

Faux, P., and Gengler, N. (2013). Genet. Sel. Evol. 45:45.

Henderson, C.R. (1984). Applications of Linear Models in Animal Breeding. University of Guelph. CA.

Legarra, A., Ducrocq, V. (2010). J. Dairy Sci. 95:1-17.

Liu, Z., Goddard, M., Reinhardt et al. (2013). In: Book of abstracts of the 64th annual meeting of EAAP, Nantes, France, August 26-30, No. 19:452.

Misztal, I., Legarra, A., and Aguilar, I. (2009). J. Dairy Sci. 92: 4648-4655

Taskinen, M., Mäntysaari, E.A., Lidauer, M.H. et al. (2013) Interbull Bulletin No. 47: 246-251.

VanRaden, P.M. (2008). J. Dairy Sci. 91:4414-4423.