Effects of data structures on the solving of animal model equations by preconditioned conjugate gradient iteration

Masahiro SATOH

National Institute of Agrobiological Sciences, Tsukuba 305-8602, Japan.

ABSTRACT

To predict breeding values in dairy cattle, preconditioned conjugate gradient (PCG) algorithm has become of interest lately as a means of solving large sets of mixed model equations (MME). Thereby, the effects of data structure on the solving of MME by PCG algorithms were investigated by using data in a computer simulation. A large number of animals did not cause slow convergence if the data structure (*e.g.* number of generations and mating ratio) was fixed. The numbers of breeding animals and their progeny affected the structure of the **A**-inverse matrix, but selection methods did not. An increasing number of iterations seemed to be due to an imbalance of nonzero elements in the **A**-inverse matrix, which resulted in the imbalance of nonzero elements in the MME. The number of iterations until convergence was roughly constant for two-trait animal model when the heritability of a trait was constant regardless of the heritability of another trait, and it was the smallest when the heritabilities of both traits were about 0.5 to 0.6. Increasing absolute values of genetic and environmental correlations resulted in slow convergence. In particular, genetic correlation affected convergence strongly.

Key words: Data structure, Mixed model equations, Preconditioned conjugate gradient

INTRODUCTION

Calculation of predicted breeding values involves solving a large set of linear mixed-model equations (MME). For example, a single trait animal model is used in Japanese dairy cattle evaluation system and the breeding values in each trait are predicted by solving MME with over five million unknowns. Traditionally, Gauss–Seidel- and Jacobi-based iterative methods (Schaeffer and Kennedy 1986; Misztal and Gianola 1987) have been used for solving MME for predicting breeding values (Ducrocq 1992). However, a preconditioned conjugate gradient (PCG) algorithm has been evaluated recently as a means of solving huge MME systems (Strandén and Lidauer 1999; Lidauer *et al.* 1999; Tsuruta *et al.* 2001; Strandén *et al.* 2002). Although the algorithm is sensitive to rounding error, especially dense matrix, and the solutions are unreliable with single precision, theoretically the number of iterations until convergence cannot exceed the number of equations. The algorithm was simpler to implement and faster than traditional iterative methods such as Gauss–Seidel- and

Jacobi-based algorithms (Strandén and Lidauer 1999; Lidauer *et al.* 1999).

In animal breeding, data differ in types of relationship between animals, the number of traits, heritabilities of traits, and correlations among traits. The convergence by PCG algorithm seems to be affected by data structure, *i.e.* animal species (numbers of progeny per dam and dams per sire) and breeding plan (number of selected traits and genetic parameters of the traits). Tsuruta *et al.* (2001) evaluated the utility of the PCG algorithm for solving MME with different data structures and models in animal breeding applications. They suggested that the PCG was the method of choice for solving large MME if sufficient computer memory were available. However, little

Correspondence: Masahiro Satoh, Genetic Diversity Department, National Institute of Agrobiological Sciences, Tsukuba 305-8602, Japan. (e-mail: hereford@affrc.go.jp)

Received 26 April 2005; accepted for publication 2 August 2005.

is known about the relationship between data structure and convergence by PCG.

The objectives of the present study were to investigate relationships between the number of unknowns (estimates of fixed and random additive genetic effects) and convergence in MME and to compare convergence properties of the PCG algorithm in a variety of different data structures.

METHODS

Iterative method

Consider a mixed linear model:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e},$$

where $\mathbf{y}$, $\mathbf{b}$, $\mathbf{u}$, $\mathbf{e}$ are the vectors of observations, fixed effects, random additive genetic effects, and random residual effects, respectively; and $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices relating records to the corresponding fixed effects and breeding values. Expectations of the random effects were assumed to be zero and $E(\mathbf{y}) = \mathbf{X}\mathbf{b}$. Assuming that $\text{var}(\mathbf{u}) = \mathbf{G}$, $\text{var}(\mathbf{e}) = \mathbf{R}$, and $\text{cov}(\mathbf{u}, \mathbf{e}) = \mathbf{0}$, the MME can be written as

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}, \quad (1)$$

where $\hat{\mathbf{u}}$ and $\hat{\mathbf{b}}$ are vectors of estimates of the fixed and additive genetic effects, respectively. We denote the MME (1) as

$$\mathbf{M}\mathbf{x} = \mathbf{v}, \quad (2)$$

where $\mathbf{M}$ is a coefficient matrix, $\mathbf{x}$ is a vector of solutions, and $\mathbf{v}$ is a vector on the right side of (1). In this study, the simplest preconditioner, a diagonal of the original coefficient matrix, was used as an approximation of the original matrix. The preconditioner matrix $\mathbf{D} = \text{diag}(\mathbf{M})^{-1/2}$. By substituting $\mathbf{x} = \mathbf{D}\mathbf{z}$ in equations (2),

$$\mathbf{DMDz} = \mathbf{Dv}.$$

Let $\mathbf{z}_0$ be a vector of initial values of the solution, and $\mathbf{p}_0 = \mathbf{r}_0 = \mathbf{Dv} - \mathbf{DMDz}_0$. Then continue the next iteration until convergence:

$$\begin{aligned} \mathbf{z}_{k+1} &= \mathbf{z}_k + \mathbf{p}_k(\mathbf{p}_k, \mathbf{r}_k) / (\mathbf{p}_k, \mathbf{DMDp}_k) \\ \mathbf{r}_{k+1} &= \mathbf{Dv} - \mathbf{DMDz}_{k+1} \\ \mathbf{p}_{k+1} &= \mathbf{r}_{k+1} - \mathbf{p}_k(\mathbf{r}_{k+1}, \mathbf{DMDp}_k) / (\mathbf{p}_k, \mathbf{DMDp}_k), \end{aligned}$$

where subscript k is the k^{th} round of iteration and $(\cdot, \cdot)$ represented the inner product between two vectors. The convergence indicator ($C_{r(k)}$) was based on the relative difference between consecutive solutions:

$$C_{r(k)} = \|\mathbf{b} - \mathbf{Ax}_k\| / \|\mathbf{b}\| < 10^{-9}, \quad (3)$$

where $\|\mathbf{b}\| = \sqrt{\mathbf{b}'\mathbf{b}} = \sqrt{\sum_i b_i^2}$. The genetic parameters used to solve the MME were the same as those of the following simulations.

Simulation 1

An infinitesimal model as described by Bulmer (1980) was used to generate data. Two traits of 5 or 10 separate generations excluding the base population were simulated. The base population was unselected, noninbred, unrelated, and sampled randomly from a conceptually infinite population. Each male was randomly mated to 1 or 10 females in each generation; the number of progeny per dam was constant and equal to 1 male and 1 female. Animals in the base population do not have any record, and the other animals have 1 record in each trait. The additive genetic values (g_i) of the i^{th} trait for progeny were generated by using the following equation:

$$g_i = 0.5(g_{si} + g_{di}) + [0.5 - 0.25(F_s + F_d)]^{0.5} \cdot \sum_k C_{ik} \varepsilon_k,$$

where g_{si} and g_{di} are additive genetic values of the i^{th} trait for the sire and dam, respectively; F_s and F_d are the inbreeding coefficients of the sire and dam, respectively; C_{ik} is the ik^{th} element of a lower triangular matrix for Cholesky decomposition of an additive genetic variance-covariance matrix; and ε_k is the k^{th} element of a vector for standard normal deviates (Toro and Silió 1992). Phenotypic values (p_{ijkl}) of the l^{th} animals of the i^{th} trait were simulated by using the following equation:

$$p_{ijkl} = S_{ij} + HGS_{ik} + g_{il} + \sum_m E_{im} \varepsilon_m,$$

where S_{ij} and HGS_{ik} are fixed effects of the j^{th} sex and of the k^{th} herd-generation-season, respectively; and E_{im} is the im^{th} element of a lower triangular matrix for Cholesky decomposition of environmental variance-covariance matrix. The phenotypic variance (σ_p^2) in each trait was assumed to be 100. The same heritability of 0.4 for each trait and genetic and environmental correlations of 0.1 were considered. The male constant in the sex effect was simulated to be 1 phenotypic standard deviation unit ($1.0 \sigma_p^2$).

larger than the female constant in each trait. The average number of animals in a herd was 100, and seasonal constants with 4 levels in a generation were assigned from a uniform distribution from a range of $\pm 2 \sigma_p^2$. For simplicity, the animals were selected randomly or using an economic selection index with the same relative weights for all traits. The total number of animals excluding the base population was 1000, 2000, 4000, 8000, 16 000, 32 000, 64 000, 128 000, 256 000, or 512 000 in each combination of number of dams per sire, number of generations, and selected methods. Two replications were performed for each of 8 different data structures in each combination.

Simulation 2

A stochastic simulation of two closed breeding herds of 200 sires and 800 dams (small population) and 2000 sires and 8000 dams (large population) per generation was conducted. For generating records and estimating breeding values, the same model as Simulation 1 was used. The animals were selected randomly or using an economic selection index with the same relative weights for all traits. Basic parameters in this simulation were as follows: 1) Number of traits = 2. 2) Number of dams per sire = 4. 3) Number of progeny per dam = 2. 4) Number of generations = 5. 5) Average number of animals in a herd = 100. 6) Range of seasonal constants assigned from a uniform distribution was $\pm 2 \sigma_p^2$. 7) Genetic and environmental correlations between

traits = 0.1. 8) Heritability of each trait = 0.4. 9) Convergence indicator, $C_{r(k)}$, was used, and the iterations were stopped when $C_{r(k)} < 10^{-9}$. Each set of parameters shown in Table 1 was simulated by holding other basic parameters constant.

RESULTS AND DISCUSSION

Table 2 shows the number of animals, rank of MME, and numbers of nonzero elements in the inverse of the additive relationship matrix and in the MME. Because the number of animals (excluding base population) in each data set was double the number in the previous data set, the numbers of nonzero elements in the inverse of the additive relationship matrix and in the MME also approximately doubled as expected. The selected methods did not affect the number of iterations until convergence (Figure 1). Five-generation data converged in fewer iterations than 10-generation data. In mating ratio, 10 dams per sire resulted in 1.8 to 2.4 times more iterations until convergence than 1 dam per sire. The number of iterations until convergence decreased slightly with increasing population size. This shows that a large number of animals and nonzero elements of MME did not cause slow convergence by PCG if the data structure (number of generations and mating ratio) was fixed.

Figure 2 shows the number of iterations until convergence for each parameter set. Because the selected

Table 1. Parameter sets used in Simulation 2

Set	Check item	Parameters ¹⁾
1	Number of traits	1, <u>2</u> , 3, 4, 5, 6
2	Number of dams per sire	1, 2, <u>4</u> , 8, 16, 32
3	Number of progeny per dam	<u>2</u> , 4, 6, 8, 10, 12
4	Number of generations	1, 3, <u>5</u> , 7, 9, 11
5	Average number of animals in a herd	50, <u>100</u> , 150, ..., 500
6	Range of seasonal constants ²⁾	0, ± 0.5 , ± 1 , ± 1.5 , <u>± 2</u> , ± 2.5 , ± 3
7	Genetic and environmental correlations	-0.9, -0.8, ..., 0.0, <u>0.1</u> , 0.2, ..., 0.9
8	Heritability	0.1, 0.2, 0.3, <u>0.4</u> , 0.5, ..., 0.9
9	Convergence indicators ³⁾ and criteria	$C_{d(k)}$, $C_{r(k)}$ 10^{-1} , 10^{-2} , ..., 10^{-8} , <u>10^{-9}</u> , 10^{-10} , 10^{-11}

¹⁾ Underlined values are basic parameters. Each set of parameters was simulated by holding other basic parameters constant.
²⁾ Unit is one phenotypic standard deviation.
³⁾ $C_{d(k)} = \|\mathbf{x}_k - \mathbf{x}_{k-1}\| / \|\mathbf{x}_k\|$, $C_{r(k)} = \|\mathbf{b} - \mathbf{A}\mathbf{x}_k\| / \|\mathbf{b}\|$.

Table 2. Summary for each data set in Simulation 1¹⁾

Data set	Number of animals	Rank of mixed model equations	Number of nonzero elements ²⁾	
			Inverse of A ³⁾	Left hand side of mixed model equations
1	1116 ± 15	2315 ± 30	3616 ± 15	19632 ± 163
2	2233 ± 30	4627 ± 61	7233 ± 30	39261 ± 326
3	4465 ± 61	9252 ± 121	14465 ± 61	78518 ± 652
4	8930 ± 121	18502 ± 243	28930 ± 121	157033 ± 1304
5	17860 ± 243	37002 ± 486	57860 ± 243	314063 ± 2608
6	35720 ± 486	74002 ± 972	115735 ± 485	628123 ± 5216
7	71440 ± 972	148002 ± 1944	231440 ± 972	1256243 ± 10431
8	142880 ± 1944	296002 ± 3887	462880 ± 1944	2512483 ± 20863
9	285760 ± 3887	592002 ± 7775	925760 ± 3887	5047363 ± 80126
10	571520 ± 7775	1184002 ± 15550	1850020 ± 7862	10207363 ± 263118

¹⁾ Values are mean ± standard deviation. They were calculated from values of eight combinations of number of dams per sire (1 or 10), number of generations (5 or 10), and selection methods (random or index selection). Each combination had two replications.

²⁾ Upper triangular matrix.

³⁾ Additive relationship matrix.

Fig. 1. Number of iterations until convergence for each data set.

The first value indicates mating ratio and the second indicates the number of generations per data set. R = randomly selected population, I = index-selected population.

methods did not affect convergence, their means are shown. Multiple-trait data required more iterations until convergence than single-trait data. However, when the genetic and environmental correlations were equal to zero, multiple-trait data produced the same number of iterations

as single-trait data. Tsuruta *et al.* (2001) found that with low correlations among direct and maternal effects, the number of iterations required for convergence was fewer for the PCG than with high correlations among the genetic effects. In general, iterative methods are very successful for solving systems for which diagonal elements are much larger than off-diagonal elements. The existence of covariance between traits in the MME may result in slow convergence, because the correlations between traits cause larger off-diagonal elements. The number of iterations until convergence increased as the number of traits increased. Similar results were shown by Mehrabani-Yeganeh *et al.* (1999) for 4 traits and Tsuruta *et al.* (2001) for 16 traits.

Increasing the mating ratio decreased the iterations until convergence. The number of rounds at convergence increased linearly as the number of progeny per litter and the number of generations increased. However, the number of animals in a herd slightly affected convergence and the range of the fixed herd-generation-season did not affect convergence. The number of iterations until convergence was always slightly more for the small population than for the large population, but the reason is not apparent. Increasing the numbers of generations and of breeding stock per dam decreased the number of sires and dams, and increasing the mating ratio decreased the number of sires. From the rules of Quaas (1976), the number of nonzero

Fig. 2. Number of iterations until convergence for each parameter set. Symbols indicate population size.

Fig. 3. Number of iterations until convergence for different heritabilities (3a) and different genetic and environmental correlations (3b).

elements in a row (column) in an A^{-1} matrix increases as the number of parent decreases when the number of animals is constant. Therefore the numbers of breeding animals and their progeny affected the number of nonzero elements of the A^{-1} matrix, which resulted in the number of nonzero elements in the MME. The increasing number of iterations in the PCG method seems to be due to the number of nonzero elements in the MME.

Figure 3 illustrate the distributions of the number of

iterations until convergence for different heritabilities and for different genetic and environmental correlations of two traits for the large population. The number of iterations until convergence was roughly constant when the heritability of a trait was constant regardless of the heritability of another trait, and it was the smallest when the heritabilities of both traits were about 0.5 to 0.6 (Fig. 3a). When the heritability was <0.35 or >0.80 , the number of iterations increased rapidly. The number of iterations until convergence

increased with increasing absolute values of genetic and environmental correlations (Fig. 3b). In particular, genetic correlation affected convergence strongly. Genetic and environmental correlations with different signs converged in slightly fewer iterations than those with the same signs. This genetic correlation result affected convergence agreed well with the results of Strandén *et al.* (2002). The lower of environmental correlation was rather unexpected because environmental correlation affects the structure of the MME by affecting dependencies between the equations.

The number of rounds at convergence increased linearly with decreasing values of C_d and C_r , as expected (Figure 4). The line for C_d was almost parallel to that for C_r . This agrees with the result of Lidauer *et al.* (1999).

The results obtained in this study show that a large number of animals did not cause slow convergence in the PCG algorithm if the data structure was fixed, *i.e.* the same numbers of traits, dams per sire, progeny per dam, *etc.* On the other hand, the data structure, especially numbers of progeny per litter, dams per sire, and generations affected the number of iterations until convergence. These results indicate that flexible programs for animal breeding are contrary to efficient programs.

Fig. 4. Number of iterations until convergence for each indicator and its criteria.
 C_r = relative average difference between left- and right-hand sides; C_d = relative average difference between consecutive solutions.

Although the PCG algorithm is sensitive to rounding error, the programs are simpler than those by traditional iterative algorithms such as Gauss–Seidel and Jacobi-based methods. The PCG algorithm generally converges in less iterations until convergence than the traditional iterative algorithms and the accuracy of solutions of MME was high (Strandén and Lidauer 1999; Satoh 2005). Although the data structure affected the number of iterations until convergence, the PCG algorithm is a good choice for solving breeding values from large data sets because a large number of animals did not cause slow convergence by PCG algorithm when the data structure was fixed.

REFERENCES

- Bulmer MG. 1980. The Mathematical Theory of Quantitative Genetics. 144–163. Clarendon Press, Oxford.
- Ducrocq V. 1992. Solving animal model equations through an approximate incomplete Cholesky decomposition. *Genetics Selection Evolution*, 24: 193–209.
- Lidauer M, Strandén I, Mäntysaari EA, Pösö J, Kettunen A. 1999. Solving large test-day models by iteration on data and preconditioned conjugate gradient. *Journal of Dairy Science*, 82: 2788–2796.
- Mehrabani-Yeganeh H, Gibson JP, Schaeffer LR. 1999. Effect of more stringent convergence criterion of estimated breeding values on response to selection. *Journal of Animal Breeding and Genetics*, 116: 339–349.
- Misztal I, Gianola D. 1987. Indirect solution of mixed model equations. *Journal of Dairy Science*, 70: 716–723.
- Quaas RL. 1976. Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics*, 32: 949–953.
- Satoh M. 2005. Comparison of different solving strategies in multiple-trait mixed model equations with iterative algorithms. *Journal of Animal Genetics*, 33: 11–18.
- Schaeffer LR, Kennedy BW. 1986. Computing strategies for solving mixed model equations. *Journal of Dairy Science*, 69: 575–579.
- Strandén I, Lidauer M. 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. *Journal of Dairy Science*, 82: 2779–2787.
- Strandén I, Tsuruta S, Misztal I. 2002. Simple

preconditioners for the conjugate gradient method: experience with test day models. *Journal of Animal Breeding and Genetics*, 119: 166–174.

Toro MA, Silió, L. 1992. An alternative to restricted BLUP based on mate selection. *Livestock Production Science*, 32: 181–187.

Tsuruta S, Misztal I, Strandén I. 2001. Use of the preconditioned conjugate gradient algorithm as a genetic solver for mixed-model equations in animal breeding applications. *Journal of Animal Science*, 79: 1166–1172.