# Approximation of Prediction Error Variance in Large-Scale Animal Models

I. MISZTAL and G. R. WIGGANS[1]
Department of Animal Sciences
University of Illinois
Urbana 61801

## ABSTRACT

Computation of prediction error variances for genetic evaluations estimated by mixed model methodology requires inversion of the coefficient matrix, which is not practical for large populations. Although methods have been developed to approximate prediction error variance for sire models, they are not suitable for animal models, because sizable effects of the relationship matrix are not considered. To approximate reciprocal of prediction error variance, an iterative algorithm was developed that combines contributions due to production records (if any) and due to relationships. Contribution due to production records is a weighted number of records; contribution due to relationships is sum of contributions from parents and offspring. Accuracy of the algorithm was investigated with a simulated data set for three generations of animals that included 1000 cows, 40 sires, 2315 records, and 100 herd-year-seasons. The model included herd-year-season and permanent environmental effects. Iteration involved reading the file with records once and reading the relationship file once per round (seven rounds were required in the simulation). Correlation between repeatability estimates obtained by the algorithm and by inversion was .996.

## INTRODUCTION

At present, the method of choice for genetic evaluation of dairy cattle is an animal model. The animal model allows for simultaneous sire and cow evaluation, prevents certain kinds of selection bias, and potentially improves accuracy of prediction. Although the theory required for genetic evaluation with an animal model was developed almost 40 yr ago (2), large-scale evaluation programs based on this model have begun only recently because of computational limitations.

The following developments have enhanced greatly computational feasibility of the animal model for large data sets. Henderson (3) discovered a rapid method to construct the inverse of a numerator relationship matrix at a fraction of the cost necessary for direct inversion. Quaas and Pollak (6) presented a reduced animal model in which size of the coefficient matrix is reduced by absorbing nonparents. Westell and Van Vleck (12) and Westell et al. (13) developed an efficient algorithm for incorporating groups in animal models. Schaeffer and Kennedy (8) presented a computational method for solving mixed model equations by iterating "on data" without creating the coefficient matrix; however, their method required keeping several sorted copies of data and relationship files. Misztal and Gianola (4) presented the theory behind iteration on data and showed that simplification was possible with Jacobi iteration. Misztal et al. (5) derived formulas for estimating accuracy of slowly converging animal model solutions and gave an algorithm for optimal selection of relaxation factors. Wiggans and Misztal (14) processed lactation data from over 100,000 first lactation Ayrshire cows with an animal model. They demonstrated that large-scale evaluation programs are possible at a reasonable cost using Jacobi iteration and selecting a nearly optimum relaxation factor. With the same data set, Wiggans et al. (15) used an animal model with all lactations and herd-sire interaction included. This recent research suggests that replacement of the Modified Contemporary Comparison on a national level is possible.

Reciprocals of diagonals of the absorbed coefficient matrix for sires have been found to provide good approximations of prediction error variance (PEV) in a sire model (7, 9, 10, 11). If relationships among sires were included in the model, reciprocals of diagonal elements ignoring relationships provided acceptable approximations

Accepted April 25, 1988.

[1]Animal Improvement Programs Laboratory, Agricultural Research Service, United States Department of Agriculture, Beltsville, MD 20705.

of PEV (11), although methods that partly considered the relationship matrix resulted in better approximations (7, 16). With an animal model, relationships cannot be ignored, because for many animals they provide the only source of information for evaluation and corresponding PEV.

The method for obtaining solutions for the animal model that will be implemented on a national scale (15) suggests a prospective algorithm for approximating PEV. This model includes management, herd-sire interaction, and permanent environmental effects. Although genetic group effects are not explicitly in the model, they are included through modification of the relationship matrix (12, 13). Complexity of the genetic grouping scheme and artificial nature of the herd-sire interaction might result in ignoring these effects when approximating PEV. Because of the large number of animals to be evaluated, evaluations are obtained by iterating on data, which does not create the coefficient matrix. Thus, for computing efficiency, approximation of PEV should be derived without using the coefficient matrix. In addition, cost of approximating PEV should not be excessive, preferably not exceeding cost of evaluation.

The goal of this study was to develop and evaluate an algorithm to approximate PEV in an animal model with the restriction that the algorithm should be suitable for large data sets.

## METHODS

Consider the linear model:

$$y_{ijk} = h_i + a_j + p_j + e_{ijk} \qquad [1]$$

with $E(y_{ijk}) = h_i$ and

$$\text{Var}\begin{bmatrix}\mathbf{a}\\ \mathbf{p}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_a^2 & 0 & 0\\ 0 & \mathbf{I}\sigma_p^2 & 0\\ 0 & 0 & \mathbf{I}\sigma_e^2\end{bmatrix}$$

where $y_{ijk}$ is value of record k of animal j in herd-year-season i, $h_i$ is a fixed herd-year-season effect, $a_j$ is a random genetic effect of animal j, $p_j$ is a random permanent environmental effect, $e_{ijk}$ is random residual effect, **A** is a numerator relationship matrix among animals, and $\sigma_a^2$, $\sigma_p^2$, and $\sigma_e^2$ are variances of **a**, **p**, and **e** effects, respectively. The mixed model equations are:

$$\begin{bmatrix}\mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'Z}\\ \mathbf{Z'X} & \mathbf{Z'Z} + \mathbf{A}^{-1}\alpha & \mathbf{Z'Z}\\ \mathbf{Z'X} & \mathbf{Z'Z} & \mathbf{Z'Z} + \mathbf{I}\tau\end{bmatrix}\begin{bmatrix}\hat{\mathbf{h}}\\ \mathbf{a}\\ \mathbf{p}\end{bmatrix} = \begin{bmatrix}\mathbf{X'y}\\ \mathbf{Z'y}\\ \mathbf{Z'y}\end{bmatrix} \qquad [2]$$

where **X** and **Z** are design matrices for **h** and **a** effects, respectively, $\alpha = \sigma_e^2/\sigma_a^2$, and $\tau = \sigma_e^2/\sigma_p^2$.

The variance-covariance matrix of $\hat{\mathbf{a}} - \mathbf{a}$ is:

$$\text{Var}(\hat{\mathbf{a}} - \mathbf{a}) = \mathbf{C}^a\sigma_e^2 \qquad [3]$$

where $\mathbf{C}^a$ is the block submatrix due to animal genetic effects in the inverse of the coefficient matrix described in [2]. If $c_j$ is diagonal element j of $\mathbf{C}^a$, then PEV of $\hat{a}_j$ is:

$$\text{PEV}(\hat{a}_j) = c_j\sigma_e^2 \qquad [4]$$

With repeatability defined as $r^2_{\hat{a}_j a_j}$, repeatability of animal j can be written as:

$$r^2_{\hat{a}_j a_j} = 1 - \text{PEV}(\hat{a}_j)/\sigma_a^2 = 1 - \alpha c_j \qquad [5]$$

The element $c_j$ can be represented as:

$$c_j = 1/(\alpha + b_j) \qquad [6]$$

where $b_j$ is nonnegative and defined to be the total information for animal j.

Two basic assumptions were made. First, we assume the existence of such matrix **D** that diagonal elements of the matrix:

$$[\mathbf{D} + \mathbf{A}^{-1}\alpha]^{-1} \qquad [7]$$

are the same as in $\mathbf{C}^a$ where **D** is a diagonal matrix with elements $\mathbf{d} = \{d_j\}$. For example, **d** can be approximated by absorbing **h** and **p** equations into animal equations and discarding all off-diagonal elements of absorption. For [1], this leads to:

$$d_j \approx \sum_k \{[1 - 1/n_{jk}][1 - m_j/(m_j + \tau)]\} \qquad [8]$$

where $n_{jk}$ is number of animals for herd-year-season in which cow j had lactation k and $m_j$ is number of parities for cow j. In accord with common usage in sire models of the phrase "effective number of daughters," each $d_j$ can be called "effective number of records." The formula [8] does not account for unequal distribution of sires in herds.

The second assumption is that $b_j$ is composed of contributions from records ($f_j$) and from relationships ($\Sigma\, g_{jl}$):

$$b_j = f_j + \sum_l g_{jl} \qquad [9]$$

where index l varies over all progeny-animal-parent combinations that include animal j. Structure of these contributions will be determined.

Although no evidence exists for the validity of these two assumptions, reasonable approximations are expected to result.

#### Contribution Due to Records

Suppose that animals are unrelated and subsequently all $g_{jl}$ are 0. Then, [7] becomes:

$$[\mathbf{D} + \mathbf{I}\alpha]^{-1} \qquad [10]$$

with diagonal elements $1/(d_j + \alpha)$. From [6] and [9], $f_j = d_j$. Thus, record contribution equals effective number of records. If all effects are absorbed into animal effects (a) and animals are assumed still unrelated, diagonals of [10] should be identical with diagonals of:

$$[\mathbf{Z'MZ} + \mathbf{I}\alpha]^{-1} \qquad [11]$$

where **M** is an absorption matrix. Approximations to diagonal elements of [11] have been studied by VanRaden and Freeman (10), and any of their methods can be applied to find approximate **D**.

#### Contributions Due to Relationships

Suppose that $\mathbf{b}^* = (b_a, b_s, b_d)$ is the information on an animal, its sire, and its dam, respectively, with relationships taken into account. Suppose, that $\mathbf{q} = (q_a, q_s, q_d)$ is the information on the same animals if their relationships are not considered. The difference $\mathbf{r} = \mathbf{b}^* - \mathbf{q}$ is contribution due to a single 3 × 3 relationship structure. If $\mathbf{b}^*$ is known, **q** can be found analytically with an elementary 3 × 3 block of Henderson's method (3) for calculating a rapid inverse of **A**. Then, assuming no inbreeding:

$$\begin{bmatrix} 1.5\alpha + q_s & .5\alpha & -\alpha \\ .5\alpha & 1.5\alpha + q_d & -\alpha \\ -\alpha & -\alpha & 2\alpha + q_a \end{bmatrix}^{-1} = \begin{bmatrix} 1/(\alpha + b_s) & * & * \\ * & 1/(\alpha + b_d) & * \\ * & * & 1/(\alpha + b_a) \end{bmatrix} \qquad [12]$$

where structure of elements denoted by * is not known. Solutions for q from the nonlinear system of equations in [12] are difficult to obtain in a closed form but can be calculated iteratively (17).

*Algorithm 1 (A1).* Iteration steps with round numbers denoted by brackets are:

1. $\mathbf{r}^{[0]} = \mathbf{0}$.
2. $\mathbf{q}^{[1]} = \mathbf{b}^*$.
3. $i = 1$.
4. $\mathbf{b}^{*[i]} = (b_s^{[i]}, b_d^{[i]}, b_a^{[i]})$ through inversion as in [12a]:

$$\begin{bmatrix} 1.5\alpha + q_s^{[i]} & .5\alpha & -\alpha \\ .5\alpha & 1.5\alpha + q_d^{[i]} & -\alpha \\ -\alpha & -\alpha & 2\alpha + q_a^{[i]} \end{bmatrix}^{-1} = \begin{bmatrix} 1/(\alpha + b_s^{[i]}) & * & * \\ * & 1/(\alpha + b_d^{[i]}) & * \\ * & * & 1/(\alpha + b_a^{[i]}) \end{bmatrix} \qquad [12a]$$

5. $\mathbf{s}^{[i]} = \mathbf{b}^* - \mathbf{b}^{*[i]}$.
6. If $\mathbf{s}^{[i]}$ is not sufficiently small, perform:
   a) $\mathbf{q}^{[i+1]} = \mathbf{b}^* + \mathbf{s}^{[i]}$.
   b) $\mathbf{r}^{[i]} = \mathbf{b}^* - \mathbf{q}^{[i+1]}$.
   c) If any element in $\mathbf{q}^{[i+1]}$ is negative, set it to 0.
   d) $i = i + 1$.
   e) Repeat from 4).

The extra step 6c) insures that the A1 will work correctly if any parent is not identified ($b_s = 0$ or $b_d = 0$) or if only approximated b is known (early stages of iteration in approximating PEV). This algorithm, especially 4), is computationally intensive. A method to obtain diagonal inverse elements without calculating the complete inverse and utilizing the structure of matrices in [12] is in the Appendix.

#### Iterative Algorithm for Predicted Error Variance

Contributions due to records and due to relationships can be combined with an iterative algorithm.

*Algorithm 2 (A2).* Iteration steps are:

1. Calculate contributions due to records **d**.
2. $\mathbf{b}^{[0]} = \mathbf{d}$.
3. $i = 1$.
4. $\mathbf{b}^{[i]} = \mathbf{d}$.
5. For every relationship (animal-parents):
   a) Extract $\mathbf{b}^* = (b_a, b_s, b_d)$ from elements of $\mathbf{b}^{[i-1]}$ corresponding to animals in selected relationship.
   b) Calculate **r** using A1.
   c) Add **r** to respective elements of $\mathbf{b}^{[i]}$.

6. If $\mathbf{b}^{[i]}$ and $\mathbf{b}^{[i-1]}$ are not close enough, perform:
   a) i = i + 1.
   b) Repeat from 4).

In A2, first round approximation for **b** is **d**. Relationship contributions are calculated with last round approximation of **b**. These contributions added to **d** form the new round approximation. Iteration continues until two subsequent approximations are close enough.

### Numerical Examples

*Example 1.* Assume that total information for an animal, its sire, and its dam are $b_a = 5$, $b_s = 3$, and $b_d = 1$, respectively, and $\alpha = 2$. Then:

$$\mathbf{b}^{*[1]} = (5.6087, 3.4688, 1.5)$$
$$\mathbf{s}^{[1]} = (.6087, .4688, .5)$$
$$\mathbf{q}^{[2]} = (4.3913, 2.5312, .5)$$
$$\mathbf{r}^{[1]} = (1.2174, .9375, 1)$$

After an additional five rounds, contributions due to relationships between animals stabilize at $\mathbf{r}^{[7]} = (.4763, .4385, .4754)$.

*Example 2.* Consider the design:

| Animal | Sire | Dam | Herd-year-seasons of records |
|---|---|---|---|
| 1 | 4 | . . . | 1, 2 |
| 2 | 4 | . . . | 1 |
| 3 | 5 | 1 | 2 |
| 4 | . . . | . . . | . . . |
| 5 | . . . | . . . | . . . |
| 6 | 5 | 2 | . . . |

If $\alpha = 2$ and $\tau = 3$, contributions due to records using [8] are approximately:

$$\mathbf{d} \approx [.6 \quad .375 \quad .375 \quad 0 \quad 0 \quad 0]$$

Contributions due to each relationship during the first two rounds of A2 are in Table 1. Subsequent round values for b are in Table 2 for iteration converged to four significant digits. The PEV for these six animals also are in Table 2.

## SIMULATION

Performance of A2 was assessed with a simulated data set. In the simulation, the base population consisted of 300 cows and 20 sires. These animals were bred to produce two additional (overlapping) generations of cows and sires. Each cow had from one to three records. Total data consisted of 2315 records of 1000 cows (daughters of 40 sires), with each record in one of 100 herd-year-seasons. Some animals were moderately inbred.

Repeatabilities for animals were obtained through two methods: 1) inversion by a sparse matrix package, SPARSPAK (1) and 2) A2. Two models were used for comparisons: 1) the full model in [1] and 2) the full model in [1] without herd-year-season effect (a "reduced" model). For both models, variance ratios were assumed to be $\alpha = \tau = 2.6$. For the reduced model, record contributions derived as in [8] with $n_{ik} = \infty$ are exact. Choice of the reduced model enabled differentiation between loss of accuracy in A2 caused by inaccurate "cow contributions" obtained as in

TABLE 1. Contributions due to information with relationships considered and relationships only for animals in example 2 during first two rounds of iteration.

| | | | | Information contribution (b) | | | Relationship contribution (r) | | |
|---|---|---|---|---|---|---|---|---|---|
| Round | Animal | Sire | Dam | Animal | Sire | Dam | Animal | Sire | Dam |
| 1 | 1 | 4 | . . . | .6 | 0 | 0 | 0 | .1224 | .1224 |
| | 2 | 4 | . . . | .375 | 0 | 0 | 0 | .0822 | .0822 |
| | 3 | 5 | 1 | .375 | 0 | .6 | .1123 | .0602 | .0598 |
| | 6 | 5 | 2 | 0 | 0 | .375 | .0822 | 0 | 0 |
| 2 | 1 | 4 | . . . | .6299 | .1023 | 0 | 0 | .1274 | .1274 |
| | 2 | 4 | . . . | .375 | .1023 | 0 | .0052 | .0811 | .0812 |
| | 3 | 5 | 1 | .4312 | .0301 | .6299 | .1156 | .0711 | .0705 |
| | 6 | 5 | 2 | .0411 | .0301 | .375 | .0903 | 0 | 0 |

TABLE 2. Contributions due to total information with relationships considered by iteration round and resulting prediction error variance (PEV) for animals in example 2.

| Iteration round | Animal | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | .6598 | .375 | .4873 | .2046 | .0602 | .0822 |
| 2 | .6705 | .3802 | .4906 | .2086 | .0711 | .0903 |
| ⋮ | | | | | | |
| 15 | .7021 | .4073 | .5007 | .2180 | .0830 | .1103 |
| PEV | $.3701\sigma_e^2$ | $.4154\sigma_e^2$ | $.3999\sigma_e^2$ | $.4509\sigma_e^2$ | $.4801\sigma_e^2$ | $.4739\sigma_e^2$ |

[8] or by possible inaccurate "relationship contributions."

## RESULTS AND DISCUSSION

For convergence, 7 rounds of iteration were required for the full model and 10 rounds for the reduced model. Comparison of repeatabilities obtained by the two methods for the two models are in Table 3. For the full model (in which record contributions are approximated), correlation between repeatabilities for the two methods was .996 and SD of difference (inversion repeatability − A2 repeatability) is .007. For the reduced model (in which record contributions are calculated accurately), SD of difference and maximum absolute differences were less than half those for the full model. Mean of difference decreased even more from .008 to below .001. This indicates that A2 accounted relatively well for the numerator relationship matrix, and most of the inaccuracy of A2 for the full model was caused by approximations in contributions due to records.

Repeatabilities for many animals in the reduced model were the same from both methods except for rounding errors. This observation led to the hypothesis that differences for remaining animals were caused by ignoring inbreeding. However, these differences diminished only slightly if inbreeding was considered.

TABLE 3. Comparison of repeatabilities obtained by inversion (RI) and by algorithm 2 (R2) for full model and model without herd-year-season effects.

| Measure | Full model | Model without herd-year-season effect |
|---|---|---|
| Mean (RI) | .421 | .437 |
| SD (RI) | .074 | .077 |
| Minimum (RI) | .089 | .091 |
| Maximum (RI) | .763 | .796 |
| Mean (R2) | .429 | .437 |
| SD (R2) | .078 | .077 |
| Maximum (R2) | .785 | .791 |
| Correlation (RI, R2) | .996* | .999* |
| Mean (RI —R2) | .008 | .000 |
| SD (RI —R2) | .007 | .003 |
| Maximum \|RI —R2\| | .063 | .038 |

*$P<.001$.

## CONCLUSIONS

The algorithms provide an approximate PEV for all animals. Contributions due to relationships are derived similarly for all animals without making distinction for sex or presence of record. More complex models can be accommodated by deriving appropriate procedures to approximate contributions of records to PEV.

Computational demands were modest; storage required amounts to three variables per animal, and iteration involved reading the record file once and the relationship file several times. New questions are raised from this study. Can the record contribution d be computed exactly for various sets of fixed and random effects, or is this impossible for a more general class of models? Modification of animal model equations as a result of incorporating groups in the relationship matrix is very simple. Is it possible to modify A2 so that group effects are accounted for in PEV? Finally, A2 should be evaluated with larger field data sets, in which amount of unbalancedness can be greater than in the simulated data of this study.

## ACKNOWLEDGMENTS

The helpful comments of C. R. Henderson, P. M. VanRaden, and J. I. Weller in addition to those of two anonymous referees are acknowledged gratefully. This research was supported by US-Israel Binational Agricultural Research and Development Project Number US-805-84; the Holstein Association, Brattleboro, VT; and the Cornell Theory Center, Ithaca, NY.

## REFERENCES

1 George, A., and J. W. Liu. 1981. Computer solution of large sparse positive definite systems. Prentice-Hall, Inc., Englewood Cliffs, NJ.
2 Henderson, C. R. 1949. Estimation of changes in herd environment. J. Dairy Sci. 32:706.
3 Henderson, C. R. 1976. A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69.
4 Misztal, I., and D. Gianola. 1987. Indirect solution of mixed model equations. J. Dairy Sci. 70:716.
5 Misztal, I., D. Gianola, and L. R. Schaeffer. 1987. Convergence rates in animal model solutions. J. Dairy Sci. 70:2577.
6 Quaas, R. L., and E. J. Pollak. 1980. Mixed model methodology for farm and ranch beef cattle testing programs. J. Anim. Sci. 51:1277.
7 Robinson, G. K., and L. P. Jones. 1987. Approximations for prediction error variances. J. Dairy Sci. 70:1623.
8 Schaeffer, L. R., and B. W. Kennedy. 1986. Computing strategies for solving mixed model equations. J. Dairy Sci. 69:575.
9 Ufford, G. R., C. R. Henderson, and L. D. Van Vleck. 1979. An approximate procedure for determining prediction error variances of sire evaluations. J. Dairy Sci. 62:621.
10 VanRaden, P. M., and A. E. Freeman. 1985. Rapid method to obtain bounds on accuracies and prediction error variances in mixed models. J. Dairy Sci. 68:2123.
11 Weller, J. I., H. D. Norman, and G. R. Wiggans. 1985. Estimation of variance of prediction error for best linear unbiased prediction models with relationships included. J. Dairy Sci. 68:930.
12 Westell, R., and L. D. Van Vleck. 1987. Simultaneous genetic evaluation of sires and cows for a large population of dairy cattle. J. Dairy Sci. 70:1006.
13 Westell, R. $A_{ij}$, R. L. Quaas, and L. D. Van Vleck. 1988. Genetic groups in an animal model. J. Dairy Sci. 71:1310.
14 Wiggans, G. R., and I. Misztal. 1987. Supercomputer for animal model evaluation of Ayrshire milk yield. J. Dairy Sci. 70:1906.
15 Wiggans, G. R., I. Misztal, and L. D. Van Vleck. 1987. Animal model evaluation of Ayrshire milk yield with all lactations, herd-sire interaction, and groups based on unknown parents. J. Dairy Sci. 71:1319.
16 Wilmink, J. B. M., and J. Dommerholt. 1985. Approximate reliability of best linear unbiased prediction in models with and without relationships. J. Dairy Sci. 68:946.

## REFERENCE ADDED IN PROOF

17 Dahlquist, D., and Åke Björk. 1974. Page 249 *in* Numerical methods. Prentice-Hall, Englewood Cliffs, NJ.

## APPENDIX

To obtain diagonal elements of an inverse of a 3 x 3 matrix, consider the symmetric matrix:

$$\begin{bmatrix} a & b & c \\ b & d & e \\ c & e & f \end{bmatrix}^{-1} = \begin{bmatrix} m & * & * \\ * & l & * \\ * & * & k \end{bmatrix}$$

where elements denoted by * are not known. Solutions to m, l, and k are:

$$\begin{aligned} m &= 1/[a - (b^2f - 2ebc + c^2d)/(df - e^2)] \\ l &= 1/[d - (b^2f - 2ebc + e^2a)/(af - c^2)] \\ k &= 1/[f - (c^2d - 2ebc + e^2a)/(ad - b^2)] \end{aligned} \quad [13]$$

Translating [13] to notation in [12]:

$$\begin{aligned} b_s &= .5\alpha + q_s - [.25\alpha^2(2\alpha + q_a) \\ &\quad - \alpha^3 + \alpha^2(1.5\alpha + q_d)]/ \\ &\quad [(1.5\alpha + q_d)(2\alpha + q_a) - \alpha^2] \\ b_d &= .5\alpha + q_d - [.25\alpha^2(2\alpha + q_a) \\ &\quad - \alpha^3 + \alpha^2(1.5\alpha + q_s)]/ \\ &\quad [(1.5\alpha + q_s)(2\alpha + q_a) - \alpha^2] \\ b_a &= \alpha + q_a - [\alpha^2(1.5\alpha + q_d) \\ &\quad - \alpha^3 + \alpha^2(1.5\alpha + q_s)]/ \\ &\quad [(1.5\alpha + q_s)(1.5\alpha + q_d) - .25\alpha^2] \end{aligned}$$