

## ALGORITHMS FOR SIMULATION OF ANIMAL MODELS WITH MULTIPLE TRAITS AND WITH MATERNAL AND NON-ADDITIVE GENETIC EFFECTS\*

L.D. Van Vleck

### ABSTRACT

The Choleski decomposition  $L_v$  of the variance-covariance matrix  $V = L_v L_v'$  can be used for simulation of genetic values for a population of animals with known numerator and dominance relationship matrices,  $A$  and  $D$ . If the variances of additive and dominance genetic effects are  $\sigma_a^2$  and  $\sigma_d^2$  and  $v_a$  and  $v_d$  are vectors of order of the number of animals ( $N$ ) of standard random normal values, then  $a = L_A v_a$  and  $d = L_D v_d$  are the vectors of simulated additive and dominance genetic values for the  $N$  animals. The calculations to accumulate elements of  $a$  or  $d$  can be done one random normal value at a time. Simulation of the multiple trait analog can be done similarly by taking advantage of the direct product property of  $G_{tN}$ , the genetic covariance matrix for the  $t$  traits and  $N$  animals. With traits ordered within animal,  $L_{G_{tN}} = L_A \otimes L_G$  where  $L_G$  is the Choleski decomposition of  $G$ , the matrix of genetic covariances among the traits and  $\otimes$  is the direct product operator. The pattern of accumulating the genetic values is such that the accumulation can be done sequentially, one vector of order,  $t$ , of standard random normal values at a time.

### INTRODUCTION

Simulation of quantitative genetic models usually involves starting with simulation of genotypic values for a parent generation followed by transmission of those simulated effects to progeny which have terms added to simulate Mendelian sampling (e.g., Kennedy, 1986). Effects of selection on estimation of breeding values or estimates of variance components can be studied in this way (e.g., Sorensen and Kennedy, 1984a,b; Walter and Mao, 1985). Models including additive, maternal, cytoplasmic, and epistatic genetic effects can also be studied (e.g., Southwood *et al.*, 1989).

Another use of simulation is to investigate sampling variances for estimates of variance-covariance components for a particular data structure. For animal models, the data structure includes the calculated numerator relationship matrix,  $A$ , for the animals with records and their connecting ancestors. The purpose of this

note is to describe a simulation procedure for single and multiple trait situations when  $A$  is known without simulation of genotypes in the sequence of parents, progeny, grandprogeny, etc. Random mating without selection must be assumed.

### METHODS

Typically with methods used to estimate variance components such as REML that utilize numerator relationships, the inverse of the numerator relationship matrix,  $A_+^{-1}$ , is computed by the rules of Henderson (1976) or by the modified rules when inbreeding is present (Quaas, 1976). The  $A_+$  notation indicates base animals are included. These calculations will include base animals that are assumed unrelated and which may or may not have records. One way to obtain  $A_+$ , which may be efficient if the order is not too large or if sparse matrix or partitioned matrix methods can be utilized, is to invert  $A_+^{-1}$ . the numerator relationship matrix for animals with records,  $A$  will be a portion of  $A_+$ . Knowledge of  $A_+$  or  $A$  will allow calculation of the additive by additive covariance matrix ( $A:A$ ) as the Hadamard product of  $A$  with itself. The dominance relationship matrix,  $D$ , can be calculated from  $A_+$  if there is no inbreeding, or approximated from  $A_+$  if there is inbreeding. Similarly, coefficients for epistatic

\* Published as Paper No. 9828, Journal Ser., Nebraska Agric. Res. Div., Univ. of Nebraska, Lincoln 68583-0908.

covariance matrices such as A:D can be calculated from A and D. In further development, A and D will be assumed known.

The following result will be the basis for using A and D for single or multiple trait simulations.

A variance-covariance matrix, V, can be written as the product of its Choleski factors;  $L_V L_V' = V$  where  $L_V$  is the lower triangular Choleski decomposition of V. A sample vector, s, of variables from a population having variance, V, can be simulated from  $L_V$  and a vector, v, of standardized random uncorrelated normal variables (mean = 0, variance = 1) as  $s = L_V v$ . Means or effects of fixed factors can be added if desired but do not affect the variance-covariance matrix of s.

Note that the variance-covariance matrix of s,

$$V(s) = E[ss'] = E[L_V v v' L_V'] = L_V E[v v'] L_V' \text{ and}$$

because  $E[v v'] = I$ ,

$$V(s) = L_V L_V' = V \text{ as required.}$$

Note that V must be positive definite. A check is to make sure all of the eigenvalues are positive.

Terms in  $L_V$  follow those derived in Van Vleck and Henderson (1961) in a simulation study of the sampling variances of estimates of genetic correlation but which were not recognized then as being elements of the Choleski decomposition of V.

## RESULTS

Several symbolic examples of the general result will be given, terminating with multiple traits and relationships among animals.

### *Independent parent-progeny pairs with environmental covariance and genetic-environmental covariance*

This example was for many years used in my class on simple methods of estimating genetic parameters which included simulation of genetic models.

Simulation is for

$$P_x = G_x + E_x$$

$$P_y = G_y + E_y \quad \text{with}$$

$$V(P_x) = \sigma_g^2 + \sigma_e^2 + 2\sigma_{ge}$$

$$V(P_y) = \sigma_g^2 + \sigma_e^2 + 2\sigma_{ge}$$

$$\text{COV}(P_x, P_y) = .5\sigma_g^2 + \sigma_{e_x e_y}.$$

Simulation of  $P_x$  and  $P_y$  is really simulation of a 4-variate sample vector,  $s = (G_x E_x G_y E_y)'$  with

$$V = \begin{pmatrix} \sigma_g^2 & \sigma_{ge} & .5\sigma_g^2 & 0 \\ \sigma_{ge} & \sigma_e^2 & 0 & \sigma_{e_x e_y} \\ .5\sigma_g^2 & 0 & \sigma_g^2 & \sigma_{ge} \\ 0 & \sigma_{e_x e_y} & \sigma_{ge} & \sigma_e^2 \end{pmatrix}$$

For numerical values of elements of V designated for the simulation, decompose V to  $L_V L_V'$ .

For each parent-progeny pair, generate a vector of four standard random normal variables,

$$v = (g_x e_x g_y e_y)'$$

Then  $s = L_V v$ .

### *Simulation of genetic values for multiple traits in unrelated animals*

As an example, assume simulation is for genetic values of three traits (A, B, C); i.e., for

$$s_g = (G_A, G_B, G_C)' \text{ with}$$

$$V(s_g) = G = \begin{pmatrix} \sigma_{G_A}^2 & \sigma_{G_A G_B} & \sigma_{G_A G_C} \\ \sigma_{G_A G_B} & \sigma_{G_B}^2 & \sigma_{G_B G_C} \\ \sigma_{G_A G_C} & \sigma_{G_B G_C} & \sigma_{G_C}^2 \end{pmatrix}$$

For numerical values of G, decompose G to  $L_G L_G'$ .

For each animal, generate a vector of three standard random normal variables,

$$g = (g_A g_B g_C)'$$

Then  $s_g = L_G g$ .

The corresponding environmental effects vector,  $s_e$ , with environmental covariance matrix, E, can be generated similarly from an independent set of three standard random normal variables,  $e = (e_A e_B e_C)'$  as  $s_e = L_E e$ .

### *Simulation for additive, dominance, and additive by dominance genetic effects of related animals*

Let A and D be the numerator (additive) and dominance relationship matrices among the animals. Let  $\sigma_a^2, \sigma_d^2, \sigma_{a:d}^2$  be the variances for additive, dominance and additive by dominance genetic values.

Vectors a, d, and a:d to represent the additive, dominance and additive by dominance genetic values of

the animals are to be simulated. A vector of environmental effects,  $e$ , could be simulated for independent environmental effects.

The first step is to decompose  $A$ ,  $D$  and  $A:D$  and multiply, respectively, by  $\sigma_a$ ,  $\sigma_d$  and  $\sigma_{a:d}$  to obtain  $L_A$ ,  $L_D$ ,  $L_{A:D}$ . This step illustrates that the method is limited to designs where the number of animals is such that the Choleski factors can be obtained in a reasonable amount of time. These steps, however, are done only once and not for each replicate. With REML or derivative-free REML procedures, each round of iteration for each replicate will require either the inverse of the coefficient matrix of at least the order of number of animals for additive effects alone, or the calculation of the logarithm of the determinant of the coefficient matrix with sparse methods. These analysis steps are usually, in fact, the real limiting factors for the simulation; not the calculation of  $A$  from  $A^{-1}$  or  $L_A$  from  $A$ . The next mathematical step is to generate vectors  $v_A$ ,  $v_D$  and  $v_{A:D}$  of standard random normal values of length the number of animals and then calculate:

$$a = L_A v_A, d = L_D v_D, \text{ and } a:d = L_{A:D} v_{A:D}.$$

Van Tassell (1989) used such a procedure for simulation of additive genetic values for a sire model with 20 sires and 50 herds with herd by sire interaction. He acknowledged L.R. Schaeffer "for suggesting the use of the Choleski decomposition to simulate the correlation structure for related sires".

Inspection of the structure of a Choleski factor shows that the standard random normal values (random normal values) can be generated and used one at a time. For example, suppose the number of animals is three and that the appropriate Choleski factor ( $L_A$  or  $L_D$  or  $L_{A:D}$ ) is

$$L = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix}$$

and the vector of random normal values is  $(v_1 \ v_2 \ v_3)'$ .

$$\begin{aligned} \text{Then } s_1 &= l_{11} v_1 \\ s_2 &= l_{21} v_1 + l_{22} v_2 \\ s_3 &= l_{31} v_1 + l_{32} v_2 + l_{33} v_3. \end{aligned}$$

This structure shows that  $s$  can be accumulated sequentially from the separate  $v_i$  without saving the complete vector  $v$ . An environmental effect can be added when the accumulation of  $s_i$  is complete to obtain the phenotypic record for animal  $i$  which then could be written to a data file with appropriate pedigree and fixed effect information used by the programs to estimate variance components.

Here is a sketch of a Fortran program to calculate  $p$ , a vector of phenotypic records to be simulated for  $N$

animals, where  $L_A$ ,  $L_D$ ,  $L_{A:D}$  are the previously calculated Choleski factors for  $A$ ,  $D$  and  $A:D$ .

```
DO 1 I = 1, N
```

```
1   p(I) = 0
```

```
DO 2 I = 1, N
```

(generate 3 random normal values,  $v_A$ ,  $v_D$  and  $v_{A:D}$ , and multiply by the respective standard deviations  $\sigma_a$ ,  $\sigma_d$ , and  $\sigma_{a:d}$ :  $a = v_A * \sigma_a$ ,  $d = v_D * \sigma_d$ ,  $a:d = v_{A:D} * \sigma_{a:d}$ )

```
DO 3 J = I, N
```

```
p(J) = p(J) + LA(J,I)*a
```

```
+ LD(J,I)*d + LAD(J,I)*ad
```

```
3 CONTINUE
```

(generate a random normal value,  $v_E$ , times the environmental standard deviation:

```
e = vE*sigma_e)
```

```
p(I) = p(I) + e
```

```
2 CONTINUE
```

If an animal,  $I$ , does not have a record, e.g., a male or base animal needed for relationships, then the environmental effect need not be generated and the record is not used in the analysis program. Only the steps between statements labelled 3 and 2 are omitted.

Fixed effects can be added at the  $p(I) = p(I) + e$  step, according to the pattern from the original data file used as a basis for the simulation.

### *Simulation of genetic values for multiple traits on related animals*

This development will be described such that each animal has a genetic value for each trait. If a record is not to be generated for a trait, then the environmental effect need not be generated and the complete record will not enter the analysis. The loop for computation of genetic values for all traits for that animal, however, is needed.

To illustrate this case, consider only additive effects for related animals and environmental covariances among traits on an animal but not from animal to animal.

The crucial step is to take advantage of the direct product property for multiple traits ( $t$ ) of related animals. Instead of the Choleski factor being of order  $tN$ , only

Choleski factors of order  $t$  for the genetic variance-covariance matrix,  $G$ , and of order  $N$  for  $A$  are needed.

Let additive genetic values be ordered within animals and let  $G_{tN}$  be the genetic variance-covariance matrix.

Then

$G_{tN} = A \otimes G$  where  $\otimes$  is the direct product operator.

Fortunately, the direct product property works for the Choleski factorization, i.e.,

$$LG_{tN} = L_A \otimes L_G.$$

Note that  $LG_{tN} LG_{tN}' = G_{tN}$  and that  $(L_A \otimes L_G) (L_A \otimes L_G)' = L_A L_A' \otimes L_G L_G' = A \otimes G$ .

As with the single trait case with relationships, examination of the structure of  $L_A \otimes L_G$  leads to efficient simulation.

For example, let  $N = 3$  and  $t = 2$  so that the Choleski factors of  $A$  and  $G$  are:

$$L_A = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \text{ and}$$

$$L_G = \begin{pmatrix} g_{11} & 0 \\ g_{21} & g_{22} \end{pmatrix}.$$

Then  $L_A \otimes L_G =$

$$\begin{pmatrix} a_{11}L_G & 0 & 0 \\ a_{21}L_G & a_{22}L_G & 0 \\ a_{31}L_G & a_{32}L_G & a_{33}L_G \end{pmatrix}.$$

Note that the partitioned blocks on and below the diagonal,  $a_{ij}L_G$ , are all lower triangular, e.g.,

$$a_{21}L_G = \begin{pmatrix} a_{21}g_{11} & 0 \\ a_{21}g_{21} & a_{21}g_{22} \end{pmatrix}.$$

With no shortcuts, a vector of length  $tN$  of random normal values could be generated and premultiplied by  $L_A \otimes L_G$ . Let  $v_i$  be the vector of  $t$  random normal values associated with the  $t$  traits of animal  $i$ . Also let  $s_i$  be the vector of  $t$  simulated genetic values for animal  $i$ :

$$\begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} a_{11}L_g & 0 & 0 \\ a_{21}L_G & a_{22}L_G & 0 \\ a_{31}L_g & a_{32}L_G & a_{33}L_G \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}$$

If the multiplication is carried out:

$$\begin{pmatrix} s_1 \\ s_2 \\ s_3 \end{pmatrix} = \begin{pmatrix} a_{11}LGv_1 \\ a_{21}LGv_1 + a_{22}LGv_2 \\ a_{31}LGv_1 + a_{32}LGv_2 + a_{33}LGv_3 \end{pmatrix}$$

The terms in  $L_A$  obviously become scalar multipliers of the vectors obtained from  $LGv_i$ . The pattern is such that  $s$  can be accumulated easily from the  $v_i$  without storing the whole of  $v$ . A vector of correlated environmental effects  $s_e$  can be added to  $s_i$  to obtain the phenotypic vector, where  $s_e = LEE_i$  and  $e_i$  is a vector of  $t$  random normal values generated for animal  $i$ . Let  $p(I,J)$  represent the phenotype of trait  $J$  for animal  $I$ ,  $LG$  be the Choleski decomposition of  $G$ ,  $LE$  be the Choleski decomposition of  $E$ , the environmental variance-covariance matrix, and  $L_A$  be the Choleski decomposition of  $A$ . Then a sketch of a Fortran program is:

C Initialize P matrix

DO 1 I = 1,N

DO 1 J = 1,t

1 p(I,J) = 0.

C Generate animals 1 to N

DO 2 I = 1,N

Generate  $v_i$ , a vector of random normal values

Generate  $e_i$ , a vector of random normal values

C Calculate  $LGv_i$  and  $LEE_i$

DO 3 K = 1,t

SUMG = 0.

SUME = 0.

DO 4 J = 1,K

SUMG = SUMG + LG(K,J)\* $v_i$ (J)

SUME = SUME + LE(K,J)\* $e_i$ (J)

4 CONTINUE

S(K) = SUMG

E(K) = SUME

3 CONTINUE

C ACCUMULATE PRODUCTS OF  $LG*v_i$  by APPROPRIATE ELEMENTS of  $L_A$

DO 5 J = 1,N

AJI = LA(J,I)

DO 6 K = 1,t

p(J,K) = p(J,K) + S(K)\*AJI

C IF LAST LA IN ROW (i.e., I=J) ADD E TERMS TO TRAITS

IF (I.EQ.J) p(J,K) = p(J,K) + E(K)

6 CONTINUE

C COULD ADD FIXED EFFECTS AND WRITE RECORD

C FOR ANIMAL I FOR ANALYSIS

C DELETE RECORDS TO BE MISSING

5 CONTINUE

2 CONTINUE

### *Simulation of genetic values for direct and maternal genetic effects with covariance on related animals*

Direct and maternal genetic values for a single trait can be simulated as for the previous multiple trait case, with the two traits being direct and maternal genetic effects in order within animal. The difference is that when parts of the record are put together, the maternal genetic value (and when the dam repeats as a mother, the maternal permanent environmental effect) comes from the dam rather than the individual. Direct and maternal genetic values are generated for all animals whether or not they are a dam or have a record. The genetic variance-covariance matrix is

$$G = \begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix} \text{ where}$$

$\sigma_a^2$  and  $\sigma_m^2$  are the direct and maternal genetic variances and  $\sigma_{am}$  is the genetic covariance between direct and maternal effects.

### CONCLUSIONS

The direct product property of Choleski factors of the genetic variance-covariance and numerator relationship matrix allows for efficient simulation of records for single and multiple trait animal models using the relationship structure of an actual data set. Direct and maternal genetic effects as well as effects for animal models with non-additive genetic effects can also be simulated easily for relationships corresponding to a real data set.

### ACKNOWLEDGMENT

The origin of the idea of using the Choleski decomposition of the covariance matrix for simulation of records on multiple traits is unknown to me. I became aware of it from a Fortran Subroutine Dr. C.R. Henderson had written, probably in the 1960's. The general method was

taught for many years at Cornell University in my class on simple methods to estimate variances and covariances, e.g., parent-progeny records with genetic-environment covariance (Animal Science 422). Several animal breeders have learned the method, although many more probably have not seen the method described or the computing steps outlined.

### RESUMO

A decomposição do Choleski  $L_V$  da matriz de variância-covariância  $V = L_V L_V'$  pode ser usada para a simulação de valores genéticos para uma população de animais com matrizes de numerador e relacionamento de dominância conhecidos,  $A$  e  $D$ . Se as variâncias de efeitos genéticos aditivos e de dominância são  $\sigma_a^2$  e  $\sigma_d^2$ , e  $v_a$  e  $v_d$  são vetores da ordem do número de animais ( $N$ ) dos valores normais aleatórios padrão, então  $a = L_{AV} v_a$  e  $d = L_{DV} v_d$  são os vetores de valores genéticos aditivos e dominantes simulados para os  $N$  animais. Os cálculos para acumular elementos de  $a$  ou  $d$  podem ser feitos com um valor normal aleatório por vez. Simulação do análogo de característica múltipla pode ser feita de maneira semelhante, aproveitando da propriedade de produto direto do  $G_{tN}$ , a matriz de covariância genética para  $t$  características e  $N$  animais. Com características ordenadas para o animal,  $L_{G_{tN}} = L_A \otimes L_G$  onde  $L_G$  é a decomposição Choleski do  $G$ , a matriz de covariâncias genéticas entre características e  $\otimes$  é o operador de produto direto. O padrão de acúmulo de valores genéticos é tal que pode ser feito em seqüência, um vetor de ordem,  $t$ , de valores normais aleatórios padronizados por vez.

### REFERENCES

- Henderson, C.R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. *Biometrics* 32: 69-83.
- Kennedy, B.W. (1986). A further look at evidence for cytoplasmic inheritance of production traits in dairy cattle. *J. Dairy Sci.* 69: 3100-3105.
- Quaas, R.L. (1976). Computing the diagonal elements and inverse of a large numerator relationship matrix. *Biometrics* 32: 949-953.
- Sorensen, D.A. and Kennedy, B.W. (1984a). Estimation of response to selection using least squares and mixed model methodology. *J. Anim. Sci.* 58: 1097-1106.
- Sorensen, D.A. and Kennedy, B.W. (1984b). Estimation of genetic variances from unselected and selected populations. *J. Anim. Sci.* 59: 1213-1223.
- Southwood, O.I., Kennedy B.W., Meyer, K. and Gibson, J.P. (1989). Estimation of additive maternal and cytoplasmic variances in animal models. *J. Dairy Sci.* 72: 3006-3012.
- Van Tassell, C.P. (1989). Consideration of sire relationships when estimating variance components with herd by sire interaction. Master's Thesis, Iowa State University, Ames.
- Van Vleck, L.D. and Henderson, C.R. (1961). Empirical sampling estimates of genetic correlations. *Biometrics* 17: 359-371.
- Walter, J.P. and Mao, I.L. (1985). Multiple and single trait analyses for estimating genetic parameters in simulated populations under selection. *J. Dairy Sci.* 68: 91-98.

(Received November 11, 1993)