PROBLEMS IN ESTIMATING COVARIANCE MATRICES IN MULTITRAIT ANIMAL MODELS

L. R. Schaeffer
Centre for Genetic Improvement of Livestock
University of Guelph
Guelph, Ontario, CANADA N1G 2W1

SUMMARY

A method of obtaining estimates of covariance matrices from animal models in multitrait situations is put forward for further study. The method is based on approximating the prediction error covariance matrices of estimates of random effects in the model. The simplified approximations allow the estimation of covariance matrices from very large data sets, with different models for each trait and with missing observations on some traits. Comparison of the approximations with actual values in small data sets has not been done, and refinements to the approximations may be possible. Estimation will still require large amounts of computing due to the sheer number of animals that need to be included in such analyses.

INTRODUCTION

There have been numerous advances in methods of estimation of components of variance from animal models. Restricted Maximum Likelihood(REML) is a prefered method which can be computed in a number of different ways (i.e. as iterative MIVQUE, an EM algorithm, or Derivative Free). Even with these alternatives, however, there are severe computational limitations with very large data sets or with models whose structure cannot be fit into existing software. Often data are sampled so that the number of records to be analyzed is not too large, or the models for several traits are assumed to be the same with no missing observations on any trait so that canonical transformations can be applied. While these efforts are commendable, the practical problem remains how to compute estimates of covariance matrices for multitrait animal models when the models for each trait may not be the same, when missing observations may occur for one or more traits, and when there are millions of animals.

REML has been the method of choice because of its apparent ability to account for various types of selection bias (e.g. selection of mates, selection of animals over years, and selection of animals on one trait to be observed for later traits), and because the estimated covariance matrices are positive (semi)definite (if p.d. priors are used). Even if REML is used, however, the estimated covariance matrix may yield different values for one trait depending on how many and which other traits are included in the analysis(Lin and Lee, 1986). For example, the heritability of a trait may increase in going from a single trait analysis to a multitrait analysis. This problem may be a consequence of inappropriate models for one or more traits. Explanations are needed, perhaps through simulation as by Southwood et al. (1989), but this problem is ignored in this paper.

The objective of this paper is to present a method of estimating

covariance matrices from multitrait animal models for very large data sets. The method is based on the philosphy of the REML method, but is not REML. The properties of the method are unknown at this time, except that it is computationally feasible for very large data sets and multitrait animal models.

METHODS

A REML EM-type formula for estimating the additive genetic variance from a single trait animal model can be written as

$$\hat{\sigma}^2_a = (\hat{a}'A^{-1}\hat{a} + tr(A^{-1}C))/q$$

where $\hat{a}$ = solutions from mixed model equations for animal additive genetic effects,

A^{-1} = the inverse of the additive genetic relationship matrix,

C = the inverse elements of the mixed model coefficient matrix corresponding to animal additive genetic effects,

q = the number of animal additive genetic effects.

Notice that $\hat{a}'A^{-1}\hat{a} = \hat{a}'LDL'\hat{a}$ following Henderson(1976), because of the structure of L, then $L'\hat{a} = \hat{m}$, a vector of Mendelian sampling effects for each animal equal to $\hat{m}_i = \hat{a}_i - 0.5(\hat{a}_{sire} + \hat{a}_{dam})$ for the i-th animal, D is a diagonal matrix with either 2, 4/3, or 1 on the diagonals in the non-inbred situations (otherwise see Quaas 1976). Therefore, $\hat{a}'A^{-1}\hat{a} = \hat{m}'D\hat{m}$. REML can account for the selection of mates because the Mendelian sampling variance should not be affected by this selection. In a similar manner $tr(A^{-1}C)$ is the weighted sum of prediction error variances for Mendelian sampling effects.

Conceptually, the REML estimate of the additive genetic covariance matrix for multiple traits can be represented as

$$\hat{G} = (\text{data contribution} + \text{prediction error variance}) * q^{-1}.$$

Let the data contribution be the sum of $\hat{m}_i\hat{m}_i' d_i$ over q animals where $\hat{m}_i$ is the vector of additive genetic Mendelian sampling effects for animal i for t traits, and d_i is the appropriate element of D for animal i. The prediction error variance contribution must be approximated because C is not obtainable. Misztal and Wiggans (1988) and Meyer (1989) have given procedures for approximating diagonals of C that are based on the idea of absorbing various factors into the diagonal block for each animal and inverting the resulting diagonal block. Alternatively, let H_i represent the sum of information used to estimate the i-th animal's Mendelian sampling effect.

Let R_i^{-1} be the inverse of the appropriate residual covariance matrix for the i-th animal (with zero rows and columns corresponding to the missing traits, if any). This matrix could be adjusted for the number of animals in the same management group as

$$R_{i*}^{-1} = R_i^{-1} - R_i^{-1} W_k^{-1} R_i^{-1}$$

where $W_k = \Sigma_i R_i^{-1}$ for animals in the k-th management group.
Then form

$$H_i = R_{i*}^{-1} + \Sigma_p 0.25\, R_{p*}^{-1}$$

and $$C_i = (H_i + G^{-1})^{-1}$$

where R_{p*}^{-1} are the inverses of the residual covariance matrices for each progeny weighted by 0.25, and G^{-1} is the inverse of the additive genetic covariance matrix of prior values. Note that there is no contribution of the parents to H_i, but one might include 0.25 times the parents' H_i. The prediction error covariance matrix(PE) for the i-th animal is C_i. Therefore, if the i-th animal does not have a record for any trait nor any progeny, then its PE is G. If the i-th animal has an infinite number of progeny for each trait, then PE is 0. Otherwise PE is somewhere between 0 and G.

The residual covariance matrix was estimated by computing a vector of residuals, $\hat{e}_i$, for each animal, and the prediction error covariance matrix for this estimate is $PE = [\, R_{i*}^{-1} + R^{-1}]^{-1}$. Let

$T = \Sigma_i \hat{e}_i \hat{e}_i'$ and let Q be the accumulation of PE for each animal, then

$$\hat{R} = (T + Q)/N$$

where N is the number of animals with records.

DISCUSSION

The methods were applied to 385,171 Canadian Hereford beef cattle records for calving ease, birthweight, weaning and postweaning gains. An animal model with effects for management group, age of dam by sex of calf effects, animal direct genetic effects, maternal genetic effects, and permanent environmental maternal effects was used. All possible combinations of missing traits existed. The model was assumed to be the same for each trait, but usually maternal effects are ignored for postweaning gain. Computationally, a reduced animal model was used to obtain parent animal solutions. During the backsolution phase for non-parent animals the sums of squares and cross products of solutions for Mendelian sampling effects (direct and maternal), and sums of prediction error variances were computed. Only one iteration was performed. With the addition of new data another iteration would be completed with the new priors. Initially, several iterations would be necessary. The

permanent environmental maternal covariance matrix was estimated separately because it was assumed to be uncorrelated to the direct and maternal genetic effects.

This procedure has the advantage that by accumulating sums of squares and prediction error variances within years of birth, then estimates of the genetic and residual parameters could be obtained for each year. The solutions for animals would be computed over all years so that biases from selection of animals over time would be accounted for properly, rather than sub-dividing the data by year of birth and estimating parameters within each subset independently. The same idea could be used to estimate parameters within management groups or herds so that heterogeneity of variances could be examined. Similarly, different estimates could be obtained if only parent animals were used versus non-parent animals versus all animals. Probably the best alternative would be to use only animals having records.

The method is based on the REML EM-type algorithm and may be the same if the approximations were the same as the exact inverse elements from the mixed model coefficient matrix. Therefore, further work is needed to determine the appropriateness of the approximations or to improve the agreement of the approximations to the inverse elements without increasing the computing complexity. The estimated covariance matrices of random effects will be positive definite if the prior values are p.d.

Another serious problem that animal breeders will be forced to solve in the near future is the problem of accounting for various types of selection in the estimation of parameters. Henderson(1980) presented a framework for MIVQUE to provide estimates unbiased by selection. The first step, however, will be to define the type of selection that has taken place in a mathematical sense (i.e. in the L matrix framework of Henderson, 1975). To date, many animal breeders have talked about selection bias in their publications without explicitly defining the type of selection. The implication being that any type of selection will or will not be accounted for by their procedures. We must become more precise in defining different types of selection.

REFERENCES

HENDERSON,C.R. 1975. Biometrics 31:423-448.
HENDERSON,C.R. 1976. Biometrics 32:69-84.
HENDERSON,C.R. 1980. J.Dairy Sci. 63:(Suppl. 1):110.
LIN,C.Y. and LEE, A.J. 1986. J. Dairy Sci. 69:2696-2703.
MEYER, K. 1989. Livest. Prod. Sci. 21:87-100.
MISZTAL, I. and WIGGANS,G.R. 1988. J. Dairy Sci. 71:(Suppl. 2):27-32.
QUAAS, R.L. 1976. Biometrics 32:949-953.
SOUTHWOOD, O.I., KENNEDY, B.W., MEYER, K., GIBSON, J.P. 1989. J. Dairy Sci. 72:3006-3012.