ANIMAL MODEL ESTIMATION USING SIMULATED REML

D. J. Klassen and S. P. Smith
Animal Genetics and Breeding Unit
University of New England
Armidale NSW 2351, Australia

SUMMARY

A two step derivative free procedure to estimate the expected information matrix is proposed, termed Simulated REML. Simulation on the realized data structure is followed by multiple regression of the Fisher quadratics on the simulated parameters and the expected information. An application to a single trait animal model is presented along with computing suggestions for implementation. Preliminary results of a simulation study are also presented.

INTRODUCTION

Restricted maximum likelihood (REML) estimation of variance components (Patterson and Thompson, 1971) has become widely accepted in animal breeding because of the desirable statistical properties of the estimates and the increasing availability of high-speed computers. The non-linear nature of the REML equations generally render analytical solutions impossible and iterative procedures must be used.

For likelihood functions whose first and second derivatives are readily evaluated and which are strictly increasing over the interval between the initial guess and the true maximum, the quadratic convergence of the Newton-Raphson method makes it the method of choice. The gradient vector is used to find the direction of steepest descent and the curvature (or Hessian) matrix defines the step size in that direction. However, because of the difficulty in calculating first and second derivatives for highly unbalanced animal breeding data it is more common to use numerical or statistical approximations of one or both of these derivatives (i.e. quasi-Newton methods) in order to reduce calculations while retaining some of the power of the Newton-Raphson method (e.g. Smith and Graser, 1986; Simianer, 1988; Meilijson, 1989; Meyer, 1989). Meyer (1990) reviews the present status of these and other methods in variance component estimation.

Fisher (1925) suggested that the gradient and Hessian needed for Newton-Raphson could be efficiently estimated from their expectations. Fisher's method of scoring (FMS) is an algorithm based on this statistical modification and has been widely used for REML analyses in animal breeding (e.g. Thompson, 1973; Schaeffer *et al.*, 1978; Meyer, 1983,1985). Although the expected information matrix, $E[\mathbf{I}(\theta)]$, is often easier to compute than the observed information matrix (Meyer, 1989b), the calculations required are often prohibitive.

In this paper a general algorithm for use in estimating the expected Fisher information will be presented. The algorithm utilizes a simulation step followed by a regression step. Its major computational burden is the application of BLUP to a number of simulated data sets. The applicability of this approach to REML estimation for a given data set and model is then determined by the availability of suitable prediction software. The inverse of the estimated expected information matrix provides a useful ancillary statistic to assess the value of the parameter estimates.

ESTIMATING THE EXPECTED INFORMATION MATRIX

Let the linear model of analysis be

$$\mathbf{y}=\mathbf{Xb}+\mathbf{Zu}+\mathbf{e}$$

where **y**, **b**, **u** and **e** are vectors of observations, fixed and random effects and residuals, and **X** and **Z** are incidence matrices for fixed and random effects respectively. The observations are assumed to come from a multivariate normal distribution with mean **Xb** and variance **V**.

Simulation Step

Using the assumed model and the exact design matrix of the data, m data vectors are simulated, with the i^{th} data vector having the variance-covariance matrix ($\mathbf{V}_i$), chosen from a grid of

possible values. BLUP solutions for the random effects ($\hat{\mathbf{u}}_i$) and the residuals ($\hat{\mathbf{e}}_i$) are produced for each data vector using an *a priori* variance-covariance matrix ($\mathbf{V}_0$) for all BLUP calculations.

Regression Step

Fisher quadratics $\mathbf{q}_i$, (i.e. $\mathbf{y}_i'\mathbf{P}\partial\mathbf{V}_0/\partial\theta\mathbf{P}\mathbf{y}_i$, where $\mathbf{P}=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ and θ is the parameter vector) are calculated for the m data vectors by appropriately weighting sums of squares and cross-products of $\hat{\mathbf{u}}_i$ and $\hat{\mathbf{e}}_i$ by squares and cross-products of elements of the *a priori* θ_0. The quadratics are then equated to their expectations, which are weighted sums of the elements of $\mathbf{V}_i$. This can be written in matrix form as:

$$\mathbf{Q} = \mathbf{Bc} + \mathbf{e}$$

where $\mathbf{Q}$ = a vector of quadratics (nm x 1) from m data vectors with n parameters to be estimated

$\mathbf{B}$ = a design matrix (nm x $n(n+1)/2$) of the variances and covariances used to simulate the m data vectors

$\mathbf{c}$ = a vector of weights ($n(n+1)/2$ x 1) to be estimated (i.e. the elements of the information matrix, in a half-stored form)

$\mathbf{e}$ = a vector of residuals

The weighting vector $\mathbf{c}$ is then estimated by ordinary least squares and used to construct the estimator of $E[\mathbf{I}(\theta)]$. As m goes to infinity, this estimator approaches the true expected information matrix. Since this matrix is always positive-definite (Jennrich and Sampson, 1976), a non-negative definite estimator is an indication that the accuracy of the estimator is not sufficient. Least squares sampling errors can be used to assess this accuracy and, if necessary, more simulated data sets can be processed and a new estimator produced with higher accuracy.

APPLICATION TO REML ESTIMATION WITH AN ANIMAL MODEL

In a simple univariate animal model (Quaas and Pollak, 1980), assumed (co)variances are

$$\text{var}\begin{bmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{ZGZ}+\mathbf{R} & \mathbf{ZG} & \mathbf{R}\\ & \mathbf{G} & \mathbf{0}\\ \text{sym.} & & \mathbf{R}\end{bmatrix} = \begin{bmatrix}\mathbf{ZAZ}'\sigma^2_u+\mathbf{I}\sigma^2_e & \mathbf{ZA}\sigma^2_u & \mathbf{I}\sigma^2_e\\ & \mathbf{A}\sigma^2_u & \mathbf{0}\\ \text{sym.} & & \mathbf{I}\sigma^2_e\end{bmatrix}$$

where $\mathbf{y}$, $\mathbf{e}$, and $\mathbf{Z}$ are as previously defined
$\mathbf{u}$ is a vector of additive genetic effects
$\mathbf{A}$ is the numerator relationship matrix
$\mathbf{G}$ and $\mathbf{R}$ are the genetic and residual covariance matrices
σ^2_u and σ^2_e are the additive genetic and residual variances

FMS for the simple animal model can be written as

$$\begin{bmatrix}\text{tr}(\mathbf{PZAZ}'\mathbf{PZAZ}') & \text{tr}(\mathbf{P}^2\mathbf{ZAZ}')\\ \text{sym.} & \text{tr}(\mathbf{P}^2)\end{bmatrix}\begin{bmatrix}\hat{\sigma}^2_u\\ \hat{\sigma}^2_e\end{bmatrix} = \begin{bmatrix}\mathbf{y}'\mathbf{PZA}^{-1}\mathbf{ZPy}\\ \mathbf{y}'\mathbf{P}^2\mathbf{y}\end{bmatrix} = \begin{bmatrix}\hat{\mathbf{u}}'\mathbf{A}^{-1}\hat{\mathbf{u}}/\sigma^4_u\\ \hat{\mathbf{e}}'\hat{\mathbf{e}}/\sigma^4_e\end{bmatrix}$$

Computing hints

1) Pre-calculation of inbreeding coefficients (F) only once for every animal reduces the computing burden in the main program. Tier (1990) presents an algorithm to do this efficiently for even very large data sets. It is useful to store this information in terms of the diagonal elements of the $\mathbf{D}$ matrix used in Quaas' tabular method (Quaas *et al.* 1984), i.e. $d_{aa} = .5 - .25F_s - .25F_d$ for animal a with sire s and dam d.

2) When creating grids of possible variance-covariance matrices, it is advisable to do a Cholesky decomposition of a matrix made up of the elements of θ_0, and vary each of the elements in a regular way. By ensuring that diagonal elements are always positive, the problem of $\mathbf{V}_i$ being outside the parameter space is avoided.

3) For the purpose of simulation, fixed effects are given a value of zero. The only information required to simulate the i^{th} data set is the realized pedigree and $\mathbf{V}_i$. The pedigree is processed oldest animals first with the random genetic effect for animal *a* with sire *s* and dam *d* simulated as

$$\alpha_{ai} = .5(\alpha_{si}+\alpha_{di}) + \sqrt{d_{aa}}\sigma_{ui}\varepsilon_{1i}$$

where d_{aa} is as defined above
α_{ai}, α_{si} and α_{di} are the i^{th} simulated genetic values for animal *a* and its parents
σ_{ui} is the i^{th} grid value for the additive genetic standard deviation
ε_{1i} is a pseudo-randomly generated number ~N(0,1)

The i^{th} phenotypic value for animal *a* is then simulated as

$$T_{ai}=\alpha_{ai} + \sigma_{ei}\varepsilon_{2i}$$

where σ_{ei} is the i^{th} grid value for the residual variance
ε_{2i} is a pseudo-randomly generated number ~N(0,1)

4) Because of the repetitive nature of the algorithm (i.e. *m* data sets being produced and evaluated independently of one another), it is natural to consider vectorization as a means to greater efficiency. The data set need only be read once per round to simulate all *m* data sets and construct the coefficient matrix and *m* right-hand-sides (if required). When considering prediction software, strategies which allow vectorization of the problem to *m* sub-problems are preferable. The method proposed by Tier and Graser (1990) has been shown to be efficient in terms of memory requirements and speed, and vectorization is straightforward. Alternatively, an indirect approach which avoids setting up the mixed model equations altogether (Schaeffer and Kennedy, 1986) may be the method of choice for very large problems.

5) There is nothing inherent in FMS which constrains the estimates to the parameter space. In fact, if MIVQUE estimates were required, unconstrained FMS would be desirable. However, by definition, REML estimates lie in the parameter space (Thompson, 1962). Therefore, a method to impose constraints (where necessary) is required. If a round of FMS yields estimates outside the parameter space, the real data Fisher quadratics ($\mathbf{q}_0$) and the estimator of $E[\mathbf{I}(\theta)]$ can be used to construct a function in the form

$$f(\theta) = \text{constant} + \mathbf{q}_0\theta - .5\theta'\widehat{E[\mathbf{I}(\theta)]}\theta$$

The maximum of this function when no constraints are present is the same as FMS. The function can be maximized with parameter constraints using any direct search routine (e.g.. simplex method, conjugate direction method). Press *et al.* (1986) review available methods and give Fortran code. The efficiency of the search routine is of little importance since function evaluation is not demanding.

BEHAVIOUR OF METHOD

The results in this section are based on a simulation study presented in more detail in Klassen and Smith (1990). Four factors which were found to affect the efficiency of the simulation approach are discussed.

1) Regression theory indicates that chosing data points farther from the mean will increase the accuracy of the regression coefficient. However, the risk of non-linear situations being encountered increases as values become more and more extreme. To test the effect of choosing increasingly extreme $\mathbf{V}_i$, the small data set of 8 animals with records and pedigrees presented in Sorensen and Kennedy (1986) was analysed using the simple model presented earlier. One hundred repeats of one iteration of FMS were performed for each of several progressively more extreme grids. Initially the empirical standard deviation (ESD) consistently declined, indicating a benefit to using more extreme grids. However, numerical problems became evident when variance ratios became greater than 1000 to 1. Most of the benefit of choosing extreme values was realized by the time variance ratios approached 25 to 1.

2) It was found that data sets with more information (i.e. more observations and close relationships) required less simulations to achieve the same accuracy. The ESD was approximately

inversely proportional to the square root of the estimated expected information. When comparing data sets of similar structure, the amount of information in a data set was approximately proportional to the number of observations. For example, a data set of 80 observations required four hundred simulations to reach approximately the same accuracy acheived with an 8000 observation data set of similar structure and only four simulations.

3) The effect of deeper pedigrees (more generations) was to increase the number of prediction iterations required for convergence of the BLUP solutions. In some cases this considerably increased the computing requirements of the procedure. Attempting to use unconverged BLUP solutions introduced biases in the estimates.

4) The minimum number of simulations required to allow estimation of $E[\mathbf{I}(\theta)]$ is $(n+1)/2$ where n is the number of parameters to estimate. For several examples of pig field data containing 8000 to 10,000 records, estimates for two and three variance component models required approximately n^2 simulations per FMS iteration to approach convergence. This is expected to approach the theoretical minimum number of simulations with very large data sets. The estimate of the large-sample sampling covariance matrix can be used to establish suitable convergence criteria. Extra simulations can be added on the last couple of iterations to refine the estimates if required.

CONCLUSIONS

The efficiency of Simulated REML estimation has been shown to be affected by choice of grid, size of data set, depth of pedigree and number of parameters estimated. Further improvements to the method will result with experience. A side product of FMS is the inverse of the estimator of $E[\mathbf{I}(\theta)]$ which yields an estimate of the asymptotic large-sample sampling covariance matrix. This can be useful when assessing the value of data to estimate the parameters of interest. Basing convergence criteria on the expected information ensures that iteration only continues when there is enough information available on the parameters to justify the effort (Meyer, 1988).

Extension of Simulated REML to more complicated models (e.g. multi-trait) appears to be straightforward, although it is not known how the method reacts when sampling covariances are high. Application to very large data sets has not yet been attempted. Further research in this area will allow comparisons with other estimation methods, for computing requirements and overall suitability. This will indicate whether situations exist where the simulation approach would be the estimation method of choice.

REFERENCES

Fisher, R.A. 1925. Cam.Phil.Soc.Proc. **22**:700-725.
Jennrich, R.T. and Sampson, P.F. 1976. Technometrics **18**:11-17.
Klassen, D.J. and Smith, S.P. 1990. (in preparation)
Meilijson, I. 1989. J.Roy.Statist.Soc.B **51**:127-138.
Meyer, K. 1983. J.Dairy.Sci. **66**:1988-1997.
Meyer, K. 1985. Biometrics **41**:153-166.
Meyer, K. 1989a. Genet.Sel.Evol. **21**:317-340.
Meyer, K. 1989b. In Festschrift in Honour of A. Robertson (in press)
Meyer, K. 1990. (This proceedings)
Patterson, L.D. and Thompson, R. 1971. Biometrika **58**:545-559.
Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. 1987. "Numerical Recipes". Cambridge University Press, Cambridge.
Quaas, R.L., Anderson, R.D., and Gilmour, A.R. 1974. "BLUP School Handbook". Animal Genetics and Breeding Unit, Armidale.
Schaeffer, L.R. and Kennedy, B.W. 1986. In Proc. 3rd World Congr.Genet.Appl.Livest.Prod. Vol. **XII**, 382-393.
Schaeffer, L.R., Wilton, J.W. and Thompson, R. 1978. Biometrics **34**:199-208.
Simianer, H. 1988. J.Anim.Breed.Genet. **104**:334-339.
Smith, S.P. and Graser, H.-U. 1986. J.Dairy.Sci. **69**:1156-1165.
Sorenson, D.A. and Kennedy, B.W. 1986. J.Anim.Sci. **63**:245-258.
Thompson, R. 1973. Biometrics **29**:527-550.
Thompson, W.A. 1962. Ann.Math.Stat. **33**:273-289.
Tier, B. 1990. Genet.Sel.Evol. (submitted)
Tier, B. and Graser, H. -U. 1990. J.Anim.Breed.Genet. (submitted)