ESTIMATION OF (CO)VARIANCE COMPONENTS BY REML IN MULTIVARIATE MIXED LINEAR MODELS USING AVERAGE OF OBSERVED AND EXPECTED INFORMATION

Per Madsen [1], Just Jensen [1] and Robin Thompson [2]
[1] National Institute of Animal Science, Research Centre Foulum, P.O. Box 39, DK-8830 Tjele Denmark
[2] Roslin Institute, Edinburgh Roslin, Midlothian, EH25 9PS, United Kingdom

SUMMARY

A REML algorithm based on first and second derivatives for the general multivariate case were derived. The algorithm, is named AI-REML, and it the average of observed and expected information as the information matrix. In test runs, the AI-REML algorithm gave almost identical parameter estimates as DF-REML, but with a considerable reduction in computer time.

INTRODUCTION

Derivative-free multivariate REML algorithms are computationally expensive, especially if the likelihood function contains many parameters to be estimated. For example Misztal (1993) showed that as the number of traits increases the derivative-free methods become less efficient than methods using first derivatives, i.e. procedures based on the EM-algorithm. One round of a derivative-free method involves computing the determinant of the coefficient matrix of the mixed model equations. In the EM-algorithm, elements in the sparse inverse of this matrix are needed. However, Misztal & Perez-Enciso (1993) have shown that these elements can be computed in about three times the computer time needed to compute the determinant of the coefficient matrix.

The poor numerical properties of the derivative-free methods in multivariate mixed models have spurred new interest in the development of algorithms utilizing first and maybe second derivatives of the likelihood function.

The matrix of second derivatives is called the observed information matrix. Expectations of this matrix is the Fisher information matrix. REML algorithms utilizing observed or expected information will lead to either the Newton-Raphson or the Fisher-scoring algorithm, respectively. Both the observed and the expected information matrices involve terms that are difficult to compute. Using univariate models, Johnson and Thompson (1994) noted that the average of observed and expected information is considerably easier to compute than either of the components due to cancellation of terms. This leads to an algorithm called AI-REML which is a compromise between the Newton-Raphson and the Fisher scoring algorithms.

The purpose of this paper is to extend the method of Johnson and Thompson (1994) to the general multiple trait model with several random effects and to compare it with existing procedures based on the derivative-free approach.

METHODS

Assuming multivariate normality the restricted log-likelihood can, following Harville (1977), be written as:

$$L(\theta) = -2\ln l = \text{const} + \ln |\mathbf{V}| + \ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\mathbf{P}\mathbf{y} \quad [1]$$

where: y denote the vector of observation, $\mathbf{V} = \text{Var}[\mathbf{y}]$, X is the design matrix corresponding to the fixed effects, and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$.

The first and second derivatives of [1] with respect to elements of θ can be written as (e.g. Harville,1977):

$$\frac{\partial L(\theta)}{\partial \theta_{(j)}} = tr\left[\frac{\partial V}{\partial \theta_{(j)}} P\right] - y' P \left[\frac{\partial V}{\partial \theta_{(j)}}\right] P y \qquad [2]$$

$$\frac{\partial^2 L(\theta)}{\partial \theta_{(j)} \partial \theta_{(j')}} = - tr\left[\frac{\partial V}{\partial \theta_{(j)}} P \frac{\partial V}{\partial \theta_{(j')}} P\right] + 2y' P \frac{\partial V}{\partial \theta_{(j)}} P \frac{\partial V}{\partial \theta_{(j')}} Py \qquad [3]$$

The matrix obtained by evaluating expression [3] for all j and j' is the observed information matrix. Taking expectations one obtains the Fisher information matrix, with typical element:

$$E\left[\frac{\partial^2 L(\theta)}{\partial \theta_{(j)} \partial \theta_{(j')}}\right] = (-) tr\left[\frac{\partial V}{\partial \theta_{(j)}} P \frac{\partial V}{\partial \theta_{(j')}} P\right] \qquad [4]$$

Computation of both [3] or [4] might in many practical applications be prohibitively tedious. However taking average of [3] and [4] as an expression of information, we obtain an average information matrix $I_A(\theta)$ with typical element:

$$I_A(\theta)_{(j,j')} = \tfrac{1}{2}\left(\frac{\partial^2 L(\theta)}{\partial \theta_{(j)} \partial \theta_{(j')}} + E\left[\frac{\partial^2 L(\theta)}{\partial \theta_{(j)} \partial \theta_{(j')}}\right]\right) = y' P \frac{\partial V}{\partial \theta_{(j)}} P \frac{\partial V}{\partial \theta_{(j')}} P y \qquad [5]$$

Define a matrix F with dimension no of observations x no. of elements in θ, with the j'th column f_j equal to the vector $\frac{\partial V}{\partial \theta_{(j)}} P y$, let W be the total design matrix, C^{-1} the inverse of the coefficient matrix of the mixed model equations (MME), and R the residual (co)variance matrix, then:

$$I_A(\theta) = F' P F = F' R^{-1} F - (C^{-1} W' R^{-1} F)' W' R^{-1} F \qquad [6]$$

Therefore, once F is known the average information can be computed by solving the MME one time for each parameter in θ, with y replaced by the corresponding column of F.

For a column in F corresponding to an element in G_{0_i}, where G_{0_i} is a $p_i \times p_i$ (co)variance matrix among the traits in the i'th random effect, it can be shown that (Madsen et. al 1994):

$$f(\theta_{i(j,k)}) = \left[\frac{\partial V}{\partial \theta_{i(j,k)}}\right] P y = Z_i \left[\left(D_{jk} G_{0_i}^{-1}\right) \otimes I\right] \hat{u}_i \qquad [7]$$

where D_{jk} is a symmetric $p_i \times p_i$ indicator matrix containing ones in positions corresponding to the j, k'th parameter in G_{0_i} and zero's elsewhere, $\theta_{i(j,k)}$ is the corresponding element in θ, and $\hat{u}_i$ is the vector of all solutions for the i'th random effect. If j = k, D_{jk} contains only a one on the j'th diagonal all other elements being zero.

For the evaluation of columns in **F** corresponding to parameters in $\mathbf{R}_0$ we need:

$$\mathbf{f}\left(\theta_{R(j,k)}\right) = \left[\frac{\partial \mathbf{V}}{\partial \theta_{R(j,k)}}\right] \mathbf{P}\,\mathbf{y} = \mathbf{R}_{jk}\mathbf{P}\mathbf{y} = \mathbf{R}_{jk}\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}}) \qquad [8]$$

If j = k the elements in f ($\theta_{R(j,k)}$) is the weighted residuals for the j'th trait with all other elements zero. If $j \neq k$, the weighted residuals for both traits are used, with all other positions equal to zero. The effect of $\mathbf{R}_{jk}$ is to interchange the weighted residuals for each trait. In practice $\mathbf{f}_i$ are not computed, in stead elements of $\mathbf{W'R^{-1}F}$ are accumulated directly during one pass through data.

In order to compute first derivatives of the log-likelihood, the terms in [2] must be evaluated. Consider the derivatives with respect to a (co)variance parameter in the i'th random effect it can be shown that (Madsen et. al, 1994):

$$\frac{\partial L(\theta)}{\partial G_{0_i}} = q_i\, G_{0_i}^{-1} - G_{0_i}^{-1}\,[\mathbf{T} + \mathbf{S}]\, G_{0_i}^{-1} \qquad [9]$$

where q_i is the number of levels for the i'th random effect, **T** and **S** are matrices with elements

$t_{jk} = tr\left[\mathbf{A}_i^{-1}\, C_{ii}^{u_j u_k}\right]$, and $s_{jk} = \hat{u}_{i_{0j}}\, \mathbf{A}_i^{-1}\, \hat{u}'_{i_{0k}}$ respectively, and where $\mathbf{A}_i$ is the covariance structure for the i'th random effect, $C_{ii}^{u_j u_k}$ is a submatrix of C^{-1} and $\hat{u}_{i_{0j}}$ is the solution vector for the j' trait in the i'th random factor.

The first derivatives with respect to residual [co]variances can be computed as (Madsen et. al, 1994):

$$\frac{\partial L(\theta)}{\partial R_{0_{kj}}} = tr\,[R_{jk}\,P] - y'\,P\,R_{jk}\,P\,y = tr[R_{jk}\,R^{-1}] - tr[C^{-1}\,W\,R^{-1}\,R_{jk}\,R^{-1}W'] = \hat{e}'\,R^{-1}\,R_{jk}\,R^{-1}\,\hat{e} \qquad [10]$$

where $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\,\hat{\beta} - \mathbf{Z}\,\hat{\mathbf{u}}$.

Note that $\mathbf{W'}\,\mathbf{R}^{-1}\,\mathbf{R}_{ij}\,\mathbf{R}^{-1}\,\mathbf{W}$ has exactly the same structure as $\mathbf{C} = \mathbf{W'R^{-1}W}$, but with $\mathbf{R}^{-1}$ replaced by $\mathbf{R}^{-1}\,\mathbf{R}_{ij}\,\mathbf{R}^{-1}$. Therefore, if a sparse inverse of **C** is available [9] and [10] can be computed together with [8] needed in the second derivatives in one pass through the data.

Once first derivatives and the average information matrix is computed new estimates of θ can be computed in a way similar to the update in the Fisher scoring algorithm.

EXAMPLE

The AI-REML algorithm was compared with the DF-REML procedure of Jensen and Madsen (1992). Data from the Danish performance testing scheme for potential AI bulls were used. Data consists of weight at 1½, 6 and 11 months of age, and daily gain and feed intake in different ages periods. A total of 3870 animals with records and 1081 sires and maternal grand sires were included in the analyses.

The comparisons were made for univariate and bivariate analyses. In addition multivariate analyses up to 5 traits were conducted by AI-REML. In all cases, the same model were used for all traits. Per trait, the model included 97 station x year x quarter classes, 21 dam year of birth classes, 4 regressions and 4951 animal effects. Parameters estimated by the two procedures for a bivariate analyses of weight at 1½ and 6 months of age are shown i Table 1.

Table 1. Variance and covariance components for weight at 1½ and 6 months of age, estimated by DF-REML and AI-REML.

	Estimated parameters*)					
	$\sigma^2_{a_1}$	$\sigma_{a_1a_2}$	$\sigma^2_{a_2}$	$\sigma^2_{e_1}$	$\sigma_{e_1e_2}$	$\sigma^2_{e_2}$
DF-REML	13.04	29.86	100.83	41.06	57.62	222.64
AI-REML	13.00	29.79	100.67	41.08	57.66	222.75

*) Likelihoods are in Table 2.

Table 2. Comparison of DF-REML and AI-REML for analyzing different number of traits.

	DF-REML[1)]			AI-REML		
No. of traits	L=-2ln l	Likelihood evaluations	CPU-time (sec)[2)]	L=-2ln l	Likelihood evaluations	CPU-time (sec)[2)]
1	20737.796	39	80.3	20737.796	5	60.1
2	46045.431	472	1837.6	46045.430	9	341.2
3				74370.553	10	1024.0
4				111481.805	13	2785.1
5				148905.075	13	5285.9

1) For DF-REML: 3, 4 and 5 trait analyses not run
2) On a CRAY C92A

DISCUSSION

In test runs, the AI-REML and DF-REML algorithms gave almost identical parameter estimates. The AI-REML algorithm converged in 13 or less rounds for all the analyses conducted, while DF-REML used 472 rounds for the bivariate analyses (3, 4 and 5 traits DF-REML analyses were not run due to constrains on CPU time). Even though that the AI-REML algorithm require more computations per round of iteration, the savings in computer time were considerable. In the bivariate analyses, the AI-REML algorithm used less than 20% of the computer time used by the DF-REML algorithm.

After further testing and optimization of the program, it will be included in the DMU-package of Jensen and Madsen (1992). One possible problem is that iterates might not be in the acceptable parameter space. Possible solutions including using a Choleski decomposition (Lindstrom and Bates, 1988) are under investigation. This might avoid numerical problems if the components are near zero.

REFERENCES

HARVILLE, D.A. (1977) J. of the American Stat., Assoc. (JASA) 72:320-340.
JENSEN, J. and MADSEN, P. (1992) Mimeo, National Inst. of Anim. Sci., DK
JOHNSON, D. and THOMPSEN, R. (1994) J. Dairy Sci. (submitted)
LINDSTROM, M.J. and BATES, D.M. (1988) J. Amer. Stat. Ass. 1-14-1022
MADSEN, P., JENSEN ,J. and THOMPSON, R. (1994) In preparation.
MISZTAL, I. (1993) Personal communication.
MISZTAL, I. and PEREZ-ENCISO, M. (1993) J. Dairy Sci., 76:1479-1483.