Jour. Ind. Soc. Ag. Statistics
49 (Golden Jub. No.), 1996-97 : 215–236

Residual Maximum Likelihood Estimation of (Co) Variance Components in Multivariate Mixed Linear Models using Average Information

Just Jensen, Esa A. Mantysaari[1], Per Madsen and Robin Thompson[2]
Danish Institute of Animal Science, Tjele, Denmark

SUMMARY

An algorithm for the REML estimation of (co) variance components in general multivariate mixed linear models is described. The algorithm is based on the use of Average Information (AI) as second differentials of the likelihood function. The AI is obtained by averaging the information matrices based on observed and expected information. It is manipulated to a form that is much easier to calculate than either of the two. This involves the setting up of dummy variables as functions of residuals and calculating sums of squares and cross-products associated with these. Procedures that are based on second differentials can lead to estimates outside the parameter space. By contrast, the EM-algorithm always ensures that estimates are in the parameter space. An alternative formulation of the EM-algorithm allows the possibility of constructing algorithms that are intermediate between AI and EM and can ensure estimates within the parameter space without the problem of slow convergence of the EM algorithm.

The new algorithm was compared to derivative-free (DF) and EM algorithms by analysing two sets of field data under several models. The AI algorithm converged in much fewer rounds than the other algorithms and was in general able to locate a higher maximum of the likelihood function.

Key words : Multivariate mixed linear models, Residual maximum likelihood, Estimation by EM algorithm.

1. *Introduction*

Variance and co-variance components are of paramount importance in animal breeding as well as in many other areas of research (Searle *et al.* [28]). In many cases data are multivariate such that covariances among traits also must be considered. The most common method currently used for the estimation

1 Agricultural Research Centre, Finland
2 Roslin Institute, Edinburgh, Scotland
Present address : IACR - Rothamsted, United Kingdom

of variance and covariance components in animal breeding research is the REML method suggested for unbalanced data by Patterson and Thompson [25]. The REML method is computationally very intensive especially in large multivariate models with several random effects. Much efforts has, therefore, gone into the search for efficient algorithms and computational procedures. Procedures currently used for the estimation of (co) variance components are usually based on either derivative-free (DF) methods, as suggested by Smith and Graser [30] and Graser *et al.* [5] or they are using first derivatives as in the EM-algorithm (Dempster *et al.* [1]).

Algorithms and computer packages implementing DF methods for multivariate models used in animal breeding have been presented by Meyer [19] and Jensen and Madsen [11]. Derivative-free REML algorithms are, however, plagued by numerical problems, especially if the likelihood function contains many parameters to be estimated (Misztal [21]). Misztal also showed that as the number of traits increases the DF methods become less efficient than methods using first derivatives, i.e. procedures based on the EM-algorithm. The major part of the computations in one round of a DF 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 and Perez-Enciso [22] have shown that these elements can be computed in about three times the computer time needed to compute the determinant of the coefficient matrix. Since the EM-algorithm may need fewer rounds, the total computing time might well be less than the time needed in algorithms based on DF methods.

The computations involved in estimating (co) variance components by the REML method can often be immense, especially if the model contains many traits that are influenced by several random factors. In special cases it is possible to use transformations that simplify the analysis of multivariate models considerably, e.g. Meyer [18], Jensen and Mao [10], Lin and Smith [14], Juga and Thompson [13] and Van Vleck and Boldman [31]. Unfortunately all these algorithms are highly specialized and cannot be used in general multivariate linear models.

The earlier mentioned poor numerical properties of the DF methods in multivariate mixed models have spurred new interest in the development of algorithms utilizing first and may be second derivatives of the likelihood function. The matrix of second derivatives of the likelihood function is called the observed information matrix. By taking expectations, one obtains the Fisher information matrix, sometimes just called the information matrix. REML algorithms utilizing observed or expected information will lead to either the

Newton-Raphson or the Fisher-scoring algorithms, respectively (e.g. Searle *et al.* [28]). Both the observed and the expected information matrices involve terms that are difficult to compute. Using univariate models, Johnson and Thompson [12] noted that the average of observed and expected information is considerably easier to compute than either of the components. This leads to an algorithm somewhat between the Newton-Raphson and the Fisher scoring algorithms.

The purpose of this paper is to extend the method of Johnson and Thompson [12] to a general multiple trait model with several random effects and allowing different models for each trait. Another purpose is to use an alternative formulation of the EM-algorithm for restricted maximum likelihood to derive algorithms that are intermediate between EM and AI algorithms and can be used to enable parameter estimates to stay in the parameter space.

2. *Model*

In this section the general multivariate linear mixed model is defined.

Let :

$$\mathbf{y} = \mathbf{X}\beta + \sum_{i=1}^{r} \mathbf{Z}_i \mathbf{u}_i + \mathbf{e} \tag{1}$$

be the Multivariate mixed model, where $\mathbf{y}$ denote the vector of observations on t traits, β is a vector of fixed effects, $\mathbf{u}_i$, $i = 1, 2, ..., r$ are vectors of random effects for the i^{th} random factor and $\mathbf{e}$ is a vector of random residuals. The design matrices $\mathbf{X}$ and $\mathbf{Z}_i$, $i = 1, 2, ..., r$ are assumed known. Without loss of generality it is assumed that $\mathbf{X}$ has full column rank.

The design matrices $\mathbf{X}$ and $\mathbf{Z}_i$ are structured. Consider the situation where records are ordered by trait and the i'th random effect affects each trait only once; i.e., $p_i = t$, where p_i is the number of traits included in the i'th random effect.

$$\text{Then}\quad \mathbf{Z}_i = \begin{bmatrix} \mathbf{Z}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{Z}_{i2} & & \vdots \\ & & \ddots & \\ \mathbf{0} & & & \mathbf{Z}_{ip_i} \end{bmatrix} = [\,\mathbf{Z}^*_{i1} \,\vdots\, \mathbf{Z}^*_{i2} \,\vdots\, \cdots \,\vdots\, \mathbf{Z}^*_{ip_i}\,] \tag{2}$$

corresponding to a partition of $\mathbf{u}_i$ as $\mathbf{u}'_i = [\, \mathbf{u}'_{i1} \,\vdots\, \mathbf{u}'_{i2} \,\vdots \cdots \vdots\, \mathbf{u}'_{ip_i} \,]$, where subvectors $\mathbf{u}_{ij}$ are the effects of the i'th random effect on the j'th trait.

For the random vectors in (1) we further assume :

$$\begin{aligned} E[\,\mathbf{u}_i\,] &= \mathbf{0} \\ E[\,\mathbf{e}\,] &= \mathbf{0} \\ \mathrm{Var}[\,\mathbf{u}_i\,] &= \mathbf{G}_i \\ \mathrm{Var}[\,\mathbf{e}\,] &= \mathbf{R} \\ \mathrm{Cov}[\,\mathbf{u}_i, \mathbf{u}'_j\,] &= \mathbf{0}, \text{ if } i \neq j \text{ and} \\ \mathrm{Cov}[\,\mathbf{u}_i, \mathbf{e}'\,] &= \mathbf{0}, \ \forall\, i \end{aligned} \tag{3}$$

Let

$$\mathbf{G} = \sum_{i=1}^{r}{}^{+}\, \mathbf{G}_i$$

$$\mathbf{Z} = [\, \mathbf{Z}_1 \,\vdots\, \mathbf{Z}_2 \,\vdots \cdots \vdots\, \mathbf{Z}_r \,]$$

and

$$\mathbf{u}' = [\, \mathbf{u}'_1 \,\vdots\, \mathbf{u}'_2 \,\vdots \cdots \vdots\, \mathbf{u}'_r \,]$$

then $\mathrm{Var}[\mathbf{y}] = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$. Generally, the (co)variance matrices are structured such that $\mathbf{G}_i = \mathbf{G}_{0_i} \otimes \mathbf{A}_i$, where $\mathbf{A}_i$ is a known matrix and $\mathbf{G}_{0_i}$ is a $p_i \times p_i$ matrix of variances and covariances among the traits in the i'th random effect. In many cases $\mathbf{A}_i$ is taken to be the identity matrix, or if $\mathbf{u}_i$ represents additive genetic effects then $\mathbf{A}_i$ is taken to be the numerator relationship matrix among the animals represented in $\mathbf{u}_i$. The dimension of $\mathbf{G}_{0_i}$ depends on the number of traits that are affected by the i'th random effect and on whether several correlated random effects affect the same trait as for example in models with direct and maternal additive genetic effects.

The residual (co)variance matrix $\mathbf{R}$, is a block diagonal matrix (for the moment assuming traits ordered within subject). The diagonal block corresponding to the i'th subject depends on which traits are measured. If all traits are measured the block is $\mathbf{R}_0$, a $t \times t$ matrix of residual (co)variances. If some traits are missing, the corresponding rows and columns in $\mathbf{R}_0$ must be deleted in order to form the diagonal block.

The parameters to be estimated are the N unique elements of the symmetric matrices $\mathbf{G}_{0_i}$, i = 1, 2, ..., r, and $\mathbf{R}_0$.

The collection of parameters are therefore :

$$\theta = [\,\mathbf{vech}\,(\mathbf{G}_{0_1})' \vdots\ \mathbf{vech}\,(\mathbf{G}_{0_2})' \vdots \cdots \vdots\ \mathbf{vech}\,(\mathbf{G}_{0_r})' \vdots\ \mathbf{vech}\,(\mathbf{R}_0)'\,]' \quad (4)$$

where **vech** is the operator putting unique elements of the argument in vector form (Searle [22]).

Individual elements in θ will generally be referred to as θ_j, for the j'th element or as e.g. $\theta_{i(j,k)}$ for a specific element corresponding to the j'k'th element in $\mathbf{G}_{0_i}$, and $\theta_{R(j,k)}$ for the j'k'th element in $\mathbf{R}_0$.

The mixed model equations corresponding to (1) are (Henderson [7]) :

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad (5)$$

Some useful relationship related to (5), discussed by Harville [6] and Searle [26] are given below :

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \quad (6)$$

which is a projection matrix mapping observations into weighted residuals :

$$\mathbf{P}\,\mathbf{y} = \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta}) = \mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\hat{\beta}-\mathbf{Z}\hat{\mathbf{u}}) \quad (7)$$

Similarly the following quantities are used to simplify derivation :

$$\mathbf{Z}'\mathbf{P}\,\mathbf{y} = \mathbf{G}^{-1}\hat{\mathbf{u}} \quad (8)$$

$$\mathbf{Z}'\mathbf{P}\mathbf{Z} = \mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{C}^{uu}\mathbf{G}^{-1} \quad (9)$$

where C is the coefficient matrix in (5) and $\mathbf{C}^{uu}$ is the submatrix of $\mathbf{C}^{-1}$ corresponding to **u**.

Finally :

$$\mathbf{y}'\mathbf{P}\,\mathbf{y} = \mathbf{y}'\mathbf{R}^{-1}\mathbf{y} - \hat{\beta}'\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} - \hat{\mathbf{u}}'\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \quad (10)$$

3. Likelihood

Assuming multivariate normality the restricted log-likelihood, modified from Harville [6], can be written as :

$$L(\theta) = -2ln(l) = \text{const} + ln\ |\mathbf{V}| + ln\ |\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}'\mathbf{P}\mathbf{y} \tag{11}$$

Again following Harville, the first and second derivatives of (11) with respect to θ can be written as :

$$\frac{\partial L(\theta)}{\partial \theta_j} = \mathbf{tr}\left[\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{P}\right] - \mathbf{y}'\mathbf{P}\left[\frac{\partial \mathbf{V}}{\partial \theta_j}\right]\mathbf{P}\mathbf{y} \tag{12}$$

$$\frac{\partial^2 L(\theta)}{\partial \theta_j \partial \theta_j'} = -\mathbf{tr}\left[\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_j'}\mathbf{P}\right] + 2\mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_j}\mathbf{P}\frac{\partial \mathbf{V}}{\partial \theta_j'}\mathbf{P}\mathbf{y} \tag{13}$$

4. Computation of First Derivatives

In order to compute the first derivatives of the log-likelihood, the terms in (12) must be evaluated. Consider initially the derivatives with respect to a (co)variance parameter in the i'th random effect.

Recall that $\mathbf{G}_i = \mathbf{G}_{0_i} \otimes \mathbf{A}_i$

Rewrite $\mathbf{G}_{0_i}$ as: $$\mathbf{G}_{0_i} = \sum_{j \geq k} \mathbf{D}_{i\{j,k\}}\ \theta_{i\{j,k\}} \tag{14}$$

where $\mathbf{D}_{i\{j,k\}}$ is a symmetric $p_i \times p_i$ indicator matrix containing ones in positions corresponding to the i, j'th parameter in $\mathbf{G}_{0_i}$ and zero's elsewhere, and $\theta_{i\{j,k\}}$ is the corresponding element in θ. Thus $\dfrac{\partial \mathbf{G}_{0_i}}{\partial \theta_{i\{j,k\}}} = \mathbf{D}_{i\{j,k\}}$ and accordingly when differentiating the likelihood in (12) :

$$\frac{\partial \mathbf{V}}{\partial \theta_{i\{j,k\}}} = \frac{\partial}{\partial \theta_{i\{j,k\}}}\left(\sum_{l=1}^{r} \mathbf{Z}_l \mathbf{G}_l \mathbf{Z}'_l + \mathbf{R}\right)$$

$$= \frac{\partial}{\partial \theta_{i\{j,k\}}} \mathbf{Z}_i \mathbf{G}_i \mathbf{Z}'_i$$

$$= \frac{\partial}{\partial \theta_{i\{j,k\}}} \mathbf{Z}_i \left(\left(\sum_{l>m} \mathbf{D}_{i\{l,m\}}\ \theta_{i\{l,m\}}\right) \otimes \mathbf{A}_i\right)\mathbf{Z}'_i$$

$$= \frac{\partial}{\partial \theta_{i\{j,k\}}} \mathbf{Z}_i \left(\left(\mathbf{D}_{i\{j,k\}} \, \theta_{i\{j,k\}} \right) \otimes \mathbf{A}_i \right) \mathbf{Z}'_i$$

$$= \mathbf{Z}_i \left(\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i \right) \mathbf{Z}'_i$$

Further define :

$$\hat{\mathbf{U}}_i = [\, \hat{\mathbf{u}}_{i_1} \vdots \hat{\mathbf{u}}_{i_2} \vdots \cdots \vdots \hat{\mathbf{u}}_{i_{pi}} \,] \, \mathbf{G}_{0_i}^{-1} \tag{15}$$

and let $\hat{\mathbf{u}}_{wij}$ be the j'th column in $\hat{\mathbf{U}}_i$, the set of weighted solutions corresponding to the i'th set of random effects.

For the trace part in (12) we obtain :

$$\begin{aligned}
\mathrm{tr}\left[\frac{\partial \mathbf{V}}{\partial \theta_{i\{j,k\}}} \mathbf{P} \right] &= \mathrm{tr}\ [\, \mathbf{Z}_i (\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i) \, \mathbf{Z}'_i \, \mathbf{P} \,] \\
&= \mathbf{tr}\ [\, \mathbf{Z}'_i \, \mathbf{P} \, \mathbf{Z}_i \, (\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i) \,] \\
&= \mathbf{tr}\ [\, (\mathbf{G}_i^{-1} - \mathbf{G}_i^{-1} \, \mathbf{C}^{u_i u_i} \mathbf{G}_i^{-1}) \, (\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i) \,] \\
&= \mathbf{tr}\, [\, ((\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1}) \\
&\qquad - (\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1}) \, \mathbf{C}^{u_i u_i} (\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1})) (\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i) \,] \\
&= \mathbf{tr}\, [\, \mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i) (\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1}) \\
&\qquad - (\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i) (\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1}) \, \mathbf{C}^{u_i u_i} \, (\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1})] \\
&= \mathbf{tr}\, [\, \mathbf{D}_{i\{j,k\}} \mathbf{G}_{0_i}^{-1} \otimes \mathbf{I}_{qi} \,] \\
&\qquad - \mathbf{tr}\, [\, (\mathbf{D}_{i\{j,k\}} \mathbf{G}_{0_i}^{-1} \otimes \mathbf{I}_{qi}) \, \mathbf{C}^{u_i u_i} (\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1}) \,] \\
&= q_i \, \mathbf{tr}\, [\, \mathbf{D}_{i\{j,k\}} \mathbf{G}_{0_i}^{-1}] - \mathbf{tr}\, [\, (\mathbf{G}_{0_i}^{-1} \mathbf{D}_{i\{j,k\}} \mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1}) \, \mathbf{C}^{u_i u_i}]
\end{aligned} \tag{16}$$

Similarly for the quadratic in (12), utilizing (8) :

$$\begin{aligned}
\mathbf{y}' \mathbf{P} \left[\frac{\partial \mathbf{V}}{\partial \theta_{i\{j,k\}}} \right] \mathbf{P} \ \mathbf{y} &= \mathbf{tr}\, [\, \mathbf{y}' \mathbf{P} [\mathbf{Z}_i (\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i) \mathbf{Z}'_i] \, \mathbf{P} \, \mathbf{y} \,] \\
&= \mathbf{tr}\, [\, \hat{\mathbf{u}}'_i \, \mathbf{G}_i^{-1} (\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i) \ \mathbf{G}_i^{-1} \hat{\mathbf{u}}_i \,] \\
&= \mathbf{tr}\, [\, \hat{\mathbf{u}}'_i \, (\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1}) (\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i) (\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1}) \, \hat{\mathbf{u}}_i \,]
\end{aligned}$$

$$= \mathbf{tr}\,[\,\hat{\mathbf{u}}'_i\,(\,\mathbf{G}_{0_i}^{-1}\,\mathbf{D}_{i\,\{j,\,k\}} \otimes \mathbf{I}_{q_i}\,)\,(\mathbf{G}_{0_i}^{-1} \otimes\ \mathbf{A}_i^{-1}\,)\,\hat{\mathbf{u}}_i\,]$$

$$= \mathbf{tr}\,[\,\hat{\mathbf{u}}'_i\,(\,\mathbf{G}_{0_i}^{-1}\,\mathbf{D}_{i\,\{j,\,k\}}\,\mathbf{G}_{0_i}^{-1}\ \otimes \mathbf{A}_i^{-1}\,)\,\hat{\mathbf{u}}_i\,] \tag{17}$$

Now consider all first derivatives with respect to $\mathbf{G}_{0_i}$ simultaneously. This can be written as a matrix $\dfrac{\partial\,\mathbf{L}\,(\theta)}{\partial\,\mathbf{G}_{0_i}}$ with diagonal elements $\dfrac{\partial\,\mathbf{L}\,(\theta)}{\partial\,\mathbf{G}_{0_{i\,(j,\,j)}}}$ and off-diagonal elements $\dfrac{1}{2}\,\dfrac{\partial\,\mathbf{L}\,(\theta)}{\partial\,\mathbf{G}_{0_{i(j\,k)}}}$.

By inspection of (16) and (17) it can be seen that the first differentials can be written in terms of cross products of solutions for individual traits and pertinent parts of the inverse of the coefficient matrix in (5).

Thus :

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

for

$$t_{i\,\{j,\,k\}} = \mathbf{tr}\,[\,\mathbf{A}_i^{-1}\,\mathbf{C}_{ii}^{u(j)\,u(k)}\,]\ \text{and}$$

$$s_{i\,\{j,\,k\}} = \hat{\mathbf{u}}'_{i_{(j)}}\ \mathbf{A}_i^{-1}\,\hat{\mathbf{u}}_{i_{(k)}} \tag{18}$$

where $\hat{\mathbf{u}}_{i_{(j)}}$ is the solution vector for the j'th trait in the i'th random factor. In understanding (18) it is useful to note that

$$\mathbf{G}_{0_i}^{-1}\ \mathbf{S}_i\,\mathbf{G}_{0_i}^{-1} = \hat{\mathbf{U}}'_i\ \mathbf{A}_i^{-1}\ \hat{\mathbf{U}}_i$$

where $\hat{\mathbf{U}}_i$ is defined in (15).

In evaluating first derivatives with respect to residual (co)variances we need to compute

$$\frac{\partial\,\mathbf{L}\,(\theta)}{\partial\,\mathbf{R}_{0_{\{i,\,j\}}}} = \mathrm{tr}\,[\,\mathbf{R}_{ij}\,\mathbf{P}\,] - \mathbf{y}'\,\mathbf{P}\,\mathbf{R}_{ij}\,\mathbf{P}\,\mathbf{y} \tag{19}$$

where $\mathbf{R}_{ij}$ is an indicator matrix defined in :

$$\mathbf{R} = \sum_{j \geq k} \mathbf{R}_{jk}\,\theta_{R\,\{j,\,k\}} \tag{20}$$

Similar to the definition of $\mathbf{D}_{i\,\{j,\,k\}}$ we obtain $\dfrac{\partial\,\mathbf{V}}{\partial\,\theta_{R\,\{j,\,k\}}} = \mathbf{R}_{jk}$.

By expanding **P** in terms of the coefficient matrix **C** we obtain for the trace part of (19) :

$$\mathrm{tr}\,[\,\mathbf{R}_{ij}\,\mathbf{P}\,] = \mathrm{tr}\,[\mathbf{R}_{ij}\,\mathbf{R}^{-1}] - \mathrm{tr}\,[\,\mathbf{C}^{-1}\,\mathbf{W}'\,\mathbf{R}^{-1}\,\mathbf{R}_{ij}\,\mathbf{R}^{-1}\,\mathbf{W}\,] \tag{21}$$

where $\mathbf{W} = [\mathbf{X} \,\vdots\, \mathbf{Z}\,]$

Similarly for the second term in (19) using (7) :

$$\mathbf{y}'\,\mathbf{P}\,\mathbf{R}_{ij}\,\mathbf{P}\,\mathbf{y} = \hat{\mathbf{e}}'\,\mathbf{R}^{-1}\,\mathbf{R}_{ij}\,\mathbf{R}^{-1}\,\hat{\mathbf{e}} \tag{22}$$

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 $\mathbf{W}'\,\mathbf{R}^{-1}\,\mathbf{W}$, 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 (21) and (22) can be computed in one pass through the data. A sufficient sparse inverse is one where the elements corresponding to non-zeros in the original matrix only are computed.

5. *Computation of Second Derivatives*

The matrix obtained by evaluating expression (13) for all j and j′ is the observed information matrix. Taking expectations one obtains the Fisher information matrix, with typical element :

$$\mathbf{E}\left[\frac{\partial^2\,\mathbf{L}\,(\theta)}{\partial\,\theta_j\,\partial\,\theta_j'}\right] = \mathbf{tr}\left[\frac{\partial \mathbf{V}}{\partial\,\theta_j}\,\mathbf{P}\,\frac{\partial\,\mathbf{V}}{\partial\,\theta_j'}\,\mathbf{P}\right] \tag{23}$$

Computation of either (13) or (23) might in many practical applications be prohibitively tedious. However, asymptotically they are identical, and this suggests taking the average of (13) and (23) as an expression of information (Johnson and Thompson [12]).

We therefore, define an average information matrix $\mathbf{I}_{\mathbf{A}}\,(\theta)$ with typical element :

$$\mathbf{I}_{\mathbf{A}}\,(\theta)_{j,j'} = \tfrac{1}{2}\left(\frac{\partial^2\,\mathbf{L}\,(\theta)}{\partial\,\theta_j\,\partial\,\theta_{j'}} + \mathbf{E}\left[\frac{\partial^2\,\mathbf{L}\,(\theta)}{\partial\,\theta_j\,\partial\,\theta_{j'}}\right]\right) = \mathbf{y}'\,\mathbf{P}\,\frac{\partial\,\mathbf{V}}{\partial\,\theta_j}\,\mathbf{P}\,\frac{\partial\,\mathbf{V}}{\partial\,\theta_{j'}}\,\mathbf{P}\,\mathbf{y} \tag{24}$$

To simplify (24) define a matrix **F** whose j'th column $\mathbf{f}_j$ consists of the vector $\dfrac{\partial\,\mathbf{V}}{\partial\,\theta_j}\,\mathbf{P}\,\mathbf{y}$. The number of columns in **F** thus equals the number of elements in θ to be estimated.

Then, $\mathbf{I}_A(\theta) = \mathbf{F}'\mathbf{P}\mathbf{F}$

$$= \mathbf{F}'\mathbf{R}^{-1}\mathbf{F} - \mathbf{T}'\mathbf{W}'\mathbf{R}^{-1}\mathbf{F} \tag{25}$$

using (10) and where $\mathbf{T}$ is a matrix whose j'th column is the solution to (5) with $\mathbf{f}_j$ used in place of $\mathbf{y}$, i.e. $\mathbf{T} = \mathbf{C}^{-1}\mathbf{W}'\mathbf{R}^{-1}\mathbf{F}$.

Therefore, once $\mathbf{F}$ is known the average information can be computed easily by solving systems like (5) once for each parameter in θ using efficient techniques for solving large and sparse linear systems, such that the solutions can be found without computing the full inverse of $\mathbf{C}$.

Computational efficiency in obtaining (25) depends on how easy it is to form $\mathbf{F}$. The following shows how to compute $\mathbf{f}_j$ corresponding to elements in $\mathbf{G}_{0_i}$ and $\mathbf{R}_0$.

$$\mathbf{f}(\theta_{i\{j,k\}}) = \left[\frac{\partial \mathbf{V}}{\partial \theta_{i\{j,k\}}}\right]\mathbf{P}\mathbf{y}$$

$$= \mathbf{Z}_i(\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i)\mathbf{Z}'_i\ \mathbf{P}\ \mathbf{y} \tag{26}$$

Now utilizing (8) we can write (26) as :

$$\mathbf{f}(\theta_{i\{j,k\}}) = \mathbf{Z}_i\,[\,(\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i)\,]\,\mathbf{G}_i^{-1}\,\hat{\mathbf{u}}_i$$

$$= \mathbf{Z}_i\,[\,(\mathbf{D}_{i\{j,k\}} \otimes \mathbf{A}_i)(\mathbf{G}_{0_i}^{-1} \otimes \mathbf{A}_i^{-1})\,]\,\hat{\mathbf{u}}_i$$

$$= \mathbf{Z}_i\,[\,(\mathbf{D}_{i\{j,k\}}\mathbf{G}_{0_i}^{-1}) \otimes \mathbf{I}\,]\,\hat{\mathbf{u}}_i \tag{27}$$

The vectors in (27) are now very easy to compute using the weighted solutions of the MME given in (15).

$$\mathbf{f}(\theta_{i\{j,k\}}) = \mathbf{Z}^*_{ij}\,\hat{\mathbf{u}}_{wk} + \mathbf{Z}^*_{ik}\,\hat{\mathbf{u}}_{wj} \tag{28}$$

where $\mathbf{Z}^*_{ij}$ were defined in (2), and $\hat{\mathbf{u}}_{wi}$ are the weighted solutions defined in (15).

If j = k the indicator matrix $\mathbf{D}_{i\{j,k\}}$ contains only a one on the j'th diagonal all other elements being zero, so (28) simplifies to :

$$\mathbf{f}(\theta_{i\{j,j\}}) = \mathbf{Z}^*_{ij}\,\hat{\mathbf{u}}_{wj} \tag{29}$$

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

$$\mathbf{f}(\theta_{R\{j,k\}}) = \left[\frac{\partial \mathbf{V}}{\partial \theta_{R\{j,k\}}}\right]\mathbf{P}\,\mathbf{y}$$
$$= \left[\frac{\partial}{\partial \theta_{R\{j,k\}}}(\mathbf{Z}\,\mathbf{G}\,\mathbf{Z}' + \mathbf{R})\right]\mathbf{P}\,\mathbf{y}$$
$$= \left[\frac{\partial}{\partial \theta_{R\{j,k\}}}\left(\sum_{l>m}\mathbf{R}_{lm}\,\theta_{R\{l,m\}}\right)\right]\mathbf{P}\,\mathbf{y}$$
$$= \mathbf{R}_{jk}\,\mathbf{P}\,\mathbf{y} \tag{30}$$

where $\mathbf{R}_{jk}$ was defined in (20).

Using (7); (30) can be written as :

$$\mathbf{f}(\theta_{R\{j,k\}}) = \mathbf{R}_{jk}\,\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\,\hat{\mathbf{u}}) \tag{31}$$

The expression above is again easy to compute. If $j = k$ the elements in $\mathbf{f}(\theta_{R\{j,k\}})$ are 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. For computational efficiency (31) should be computed in the same pass through data as while collecting (18) for the first derivatives.

Furthermore, since $\mathbf{f}_j$ is not needed explicitly but only in setting up systems like (5) with $\mathbf{f}_j$ in place of $\mathbf{y}$ the corresponding right hand sides can be computed directly without actually forming $\mathbf{f}_j$.

6. Update of (Co) Variance Parameters

When $\mathbf{I}_A$ and $\dfrac{\partial L(\theta)}{\partial \theta}$ have been computed with an estimate θ_n of θ, a new estimate of the variance components can be found using the Newton update Δ from :

$$\Delta = \mathbf{I}_A^{-1}\,\frac{\partial L(\theta)}{\partial \theta} \tag{32}$$

The estimate of θ to be used in the next iteration is $\hat{\theta}_{n+1} = \theta_n - \Delta$. The procedure must be iterated until $\| \Delta \| < \varepsilon$ where ε is a small positive number. At the convergence $\mathbf{I}_A^{-1}$ will contain asymptotic estimates of Var $[\hat{\theta}]$.

A problem with the Newton update is that the parameters are not guaranteed to stay within the parameter space. Therefore, after each update it must be checked that all (co)variance matrices estimated $\mathbf{G}_{0_i}$, $i = 1, 2 \ldots r$ and $\mathbf{R}_0$ are positive (semi)definite. If the Newton update leads to $\hat{\theta}$ outside parameter space, Johnson and Thompson [12] suggested to modify (32) using the method of Marquardt [17]. This method amounts to adding a constant to the diagonal elements of the information matrix before solving (32). Another, perhaps simpler, approach is to switch to the EM algorithm of Dempster *et al.* [1]. This can be done either on all or on some of the parameters to be estimated. It is also possible to combine the AI and the EM algorithm.

7. *Estimation by EM Algorithm*

A typical EM estimate (see e.g. Mäntysaari and Van Vleck [23]) for a parameter in the i'th random effect in **G** is calculated as :

$$\hat{\mathbf{G}}_{0i\{jk\}} = \frac{1}{q_i\,(2 - \delta_{jk})}\,[\hat{\mathbf{u}}_i\,(\mathbf{D}_{i\,\{j,k\}}\,\mathbf{A}_i^{-1})\,\hat{\mathbf{u}}_i + \mathrm{tr}\,(\mathbf{D}_{i\,\{j,k\}}\,\mathbf{A}_i^{-1}\;\mathbf{C}^{u_{i\{j\}}\,u_{i\{k\}}})] \qquad (33)$$

where $\delta_{jk} = 1$ if $j = k$ and zero otherwise.

Therefore the update to estimate $\mathbf{G}_{0_i}$ can be written as :

$$\hat{\Delta}_i = [q_i\,\mathbf{G}_{0_i} - (\mathbf{T}_i + \mathbf{S}_i)\,]\,/\,q_i \qquad (34)$$

This update can be computed from the first differentials in (18) by pre- and post-multiplying by the corresponding $\mathbf{G}_{0_i}$ and dividing the product by q_i the number of levels in the i'th random factor.

In a notation similar to (32) the update to estimate $\mathbf{G}_{0_i}$ can be written as :

$$\Delta_i = \mathbf{I}_{EM}^{-1}\ \text{vech}\left(\frac{\partial\,L\,(\theta)}{\partial\,\mathbf{G}_{0_i}}\right) \qquad (35)$$

Consideration of the multiplications involved informing Δ_i show that the elements of $\mathbf{I}_{EM}^{-1}$ corresponding to $\theta_{i_{\{jk\}}}$ and $\theta_{i_{\{rs\}}}$ are :

$$[\, G_{0_{i\{jr\}}} \, G_{0_{i\{ks\}}} + G_{0_{i\{js\}}} \, G_{0_{i\{kr\}}} \,] / q_i \tag{36}$$

for jk different from rs. Note also that $\theta_{i\{jk\}} = G_{0i\{jk\}}$.

This inverse information matrix gives the (co)variance matrix of $\mathbf{G}_{0_i}$ if the q_i random effects were directly observed. The elements of $\mathbf{I}_{EM_i}$ are functions of the elements $G_{i\{jk\}}^{-1}$ of $\mathbf{G}_{0_i}$ and it can be verified that they are :

$$q_i \; [\, G_{0_{i\{jr\}}}^{-1} \, G_{0_{i\{ks\}}}^{-1} + G_{0_{i\{js\}}}^{-1} \, G_{0_{i\{kr\}}} \,] / (1 + \delta_{jk} \, \delta_{rs}) \tag{37}$$

The EM updates for $\mathbf{R}_0$ can be derived in similar way but it is necessary to take the possible different missing data patterns into account (Mäntysaari, Jensen and Thompson, unpublished). For example the update to form $\mathbf{R}_0$ from $\mathbf{R}_0$ is $\mathbf{R}_0 \dfrac{\partial L}{\partial \mathbf{R}_0} \mathbf{R}_0$, which is in the same form as the update to form $\hat{\mathbf{G}}_0$ from $\mathbf{G}_0$ in (34). Therefore in vector form Δ_R can be written as :

$$\Delta_R = \mathbf{I}_{EMR}^{-1} \; \text{vech} \left(\frac{\partial L(\theta)}{\partial \mathbf{R}_0} \right) \tag{39}$$

The matrices $\mathbf{I}_{EMR}^{-1}$ and $\mathbf{I}_{EMR}$ are given by use of (36) and (37) with $\mathbf{R}_0$ replacing $\mathbf{G}_{0_i}$. By combining the updates for all sets of (co)variance parameters we obtain:

$$\mathbf{I}_{EM} = \left(\sum^{+} \mathbf{I}_{EM_i} \right) \otimes \mathbf{I}_{EMR} \tag{40}$$

The advantage of the alternative formulation of the EM-REML algorithm is that it allows the combination of AI and EM information. If the AI algorithm in a certain round yields parameters outside parameter space a switch to the EM-algorithm will yield estimates inside the parameter space with increased likelihood. Unfortunately the increase in likelihood can be very small and a better alternative might be to use a combined information matrix :

$$\mathbf{I}_{AEM} = (1 - b_{EM}) \, \mathbf{I}_A + b_{EM} \, \mathbf{I}_{EM} \tag{41}$$

where $b_{EM} \in [0, 1]$ and must be chosen such that estimates are within parameter space.

8. *Summary of Computational Steps*

The following summarizes the necessary steps in the proposed algorithm:

1. Form the structure of the multiple trait mixed model equations in sparse form, e.g. as in Duff *et al.* [2].
2. To facilitate effective factorization and to minimize fill in, reorder the MT-MME using a minimum degree algorithm or such e.g. George and Liu [3] and carry out symbolic factorization of the coefficient matrix.
3. Given the current value of θ, form the numerical part of MT-MME by one pass through the data, carry out numerical factorization, solve the system, and compute the sparse inverse of the coefficient matrix.
4. In one pass through the data, compute residuals, first derivatives of $L(\theta)$ and form $\mathbf{W}'\mathbf{R}^{-1}\mathbf{f}_i(\theta_i)$ for $i = 1, 2, .., \mathbf{N}$.
5. Solve MT-MME one time for each parameter in θ using $\mathbf{W}\mathbf{R}^{-1}\mathbf{f}_i(\theta_i)$ formed in step 4 as right hand sides, and compute $\mathbf{I}_A$.
6. Update (co)variance components using (32).
7. Check that the new parameters are within parameter space. If yes check for the convergence. If no, discard AI update and compute a combined update with a sufficient weight on EM to ensure that estimates stay within parameter space.

Steps 3-7 must be repeated until convergence. Each iteration requires two passes through the data, computation of the sparse inverse of the coefficient matrix and solving the mixed model equations 1+N times. Efficient use of sparse matrix technology, e.g. as in FSPAK (Misztal and Perez-Enciso [22]) is therefore of crucial importance for large models.

9. *Examples*

The new algorithm was tested on two different sets of field data. The first data set consisted of records from the Danish beef-performance test stations for dual purpose cattle. A total of 5489 Holstein bulls had records on the traits weight at 1.5, 6 and 11 months of age. Tracing pedigree information increased the total number of animals in the analysis to 15241. The model for each trait included fixed effects of station-year-season and effects of age, proportion of Holstein-Friesian genes and heterotsygosity as covariables. The only random effects in the model were animal and error.

The second data set consisted of records of weights of Texel sheep at birth and at 2 months of age. A total of 7863 animals had records, and after tracing pedigree information the data set contained a total of 9460 animals.

For each trait the model included 5 cross classified fixed effects with a total of 421 levels, and as random effects the model included : permanent environmental effect of dam, litter within dam, additive direct effect, additive maternal effect and random error. The model assumed a non-zero additive genetic covariance between direct and maternal additive genetic effects.

Results for one, two, and for cattle example, three-trait analyses are reported. Some characteristics of each example model are shown in Table 1. For the sheep data the number of parameters to be estimated increases dramatically as more traits are included in the analysis. This is because the model includes four random effects and both direct and maternal additive genetic effects so that the dimension of the additive genetic covariance matrix to be estimated becomes twice the number of traits included in the analysis.

Table 1. Characteristics of example models analyzed

Item	Cattle data			Sheep data	
No. of traits	1	2	3	1	2
NZ[1] in MME[2]	72914	263942	564634	260729	976255
Rank of MME	15537	31074	46581	27705	55410
Parameters to estimate	2	6	12	6	19
Average NZ per equation	4.69	8.49	12.12	9.41	17.62
Pct. filled cells	0.054	0.051	0.050	0.064	0.062

[1] Non-zeros
[2] Mixed model equations

For comparison some of the models were also analysed using a DF and an Em algorithm. All the computations were performed using the DMU-package of Jensen and Madsen [11]. The starting values for parameters to be estimated were the same in all the analyses. Due to very long computing times for EM the two trait analyses on sheep data were run using the DF and AI algorithms only.

10. Results and Discussion

The algorithm presented here is an extension of basic ideas presented by Johnson and Thompson [12] in order to analyse general multiple trait models.

In contrast of DF and EM algorithms the AI algorithm usually converges in very few rounds. This is clearly seen in Table 2 where the number of rounds, computing time and the log-likelihood at convergence are presented for each

Table 2. Number of rounds/evaluations, computing time, and log-likelihood (L) at convergence for DF, EM and AI algorithms run on example models

Model and measure	Algorithm		
	DF	EM	AI (EM)
Cattle, 1 trait			
Rounds	41	77	4 (0)**
Time (s)	142	1237	84
–2 In L	4.87005	4.87012	4.87006
Cattle, 2 traits			
Rounds	351	1000*	6 (0)
Time (s)	4704	108353	710
–2 In L	89.78331	89.78341	89.78306
Cattle, 3 traits			
Rounds	1435	1000*	6(0)
Time(s)	54896	319127	2068
–2 In L	16.84969	15.15830	15.15638
Sheep, 1 trait			
Rounds	471	620	5(1)
Time(s)	1267	47596	418
–2 In L	40.67845	40.67893	40.67846
Sheep, 2 traits			
Rounds	5813	—	6(1)
Time(s)	110222	—	3570
–2 In L	94.51863	—	44.53450

* Maximum number of iterations reached.
** No of evaluations with weight on EM.

example run. Although each round of iteration in AI scheme can take more time than rounds for DF or EM, the amount of computer time needed is much less for the AI algorithm than for the DF and EM. The superiority of AI algorithm is obvious when the number of parameters to be estimated is large as the number of iterations seems not to be affected by the number of dimensions in the maximization. The same was seen before in Madsen *et al.* [16] where the same data as here was analyzed considering simultaneously up to 5 traits. The number of iterations remained constantly under 15, although the number of parameters estimated was 30 (Madsen *et al.* [16]). Moreover, the AI algorithm was generally able to locate a higher ln L, than the EM and DF algorithms. In cases with many parameters to estimate; (3 traits in cattle data or 2 traits with sheep model) the DF algorithm was not able to satisfactorily locate the maximum of the likelihood function. This is most likely to be due to the poor numerical properties of algorithms based upon derivative free methods as discussed by Misztal [21].

A problem in comparing the number of rounds with different algorithms is the stopping criteria used. In DF the Simplex algorithm of Nelder and Mead [24] was used, and it was required that the variance of the log-likelihood values in the polytope was less than 10^{-8}. In EM and AI several alternatives were tested. The norm of the update vector $\| \Delta \| < \varepsilon_1$ has a disadvantage since it can be very small in EM when the solutions are still far away from the maximum. If the algorithm converges to a point inside the parameter space, the vector of first derivatives (the gradients) should approach zero. An alternative stopping criteria would therefore be $\| \frac{\partial L(\theta)}{\partial \theta} \| < \varepsilon_2$, where ε_2 is a small positive number.

Usually the parameters in the likelihood are estimated with varying precision. A parameter estimated with a low precision corresponds to a dimension in the likelihood where the surface is relatively flat. Thus a third stopping criteria could therefore be $\left\| \frac{\mathbf{diag.}\ (\mathbf{I}_{AI}^{-1})}{\sqrt{\mathbf{N}}} \frac{\partial L(\theta)}{\partial \theta} \right\| < \varepsilon_3$ where N is the number of parameters to be estimated. The last stopping criteria has the advantage that dimensions corresponding to parameters estimated with a low accuracy will get more weight. The disadvantage of the third convergence criteria is that if estimates are at the boundary of the parameter space the vector of first derivatives is not necessarily zero. In our implementation we have therefore chosen to stop whenever criteria 1 or criteria 3 were fulfilled but with $\varepsilon_1 << \varepsilon_3$. In the examples here EM and AI algorithms had the same convergence criterium. The main emphasis was on criteria 3, i.e., the norm of weighted gradient, which was fulfilled in all AI runs ($\varepsilon_3 = 0.05$). However, in all cases but one EM failed to fulfill the criteria, and the maximum limit of 1000 iterations were reached instead. Slow convergence near the maximum seems to be a characteristic of EM algorithm in cases where the likelihood is complicated and the information from the data is limited.

The update in (32) do not guarantee estimates within the parameter space, and if estimates are inadmissible some action must be taken. One possibility is to discard the AI update and replace it with an EMn update, since that ensures that estimates stay within parameter space and an EM update will always increase the likelihood. However, our experience revealed that using an update from the combined information matrix in (41) was a better alternative. This leaves the problem of choosing the relative weights on the two algorithms. A practical approach that we used was to set b_{EM} to a small number and then increase it until estimates from the combined algorithms are within parameter space. Experience, where b_{EM} initially was set to 1/200 and increased by 1/200 until admissible parameters were obtained suggested that a very small weight in EM in many practical cases is sufficient. Obviously our choice of b_{EM} was very arbitrary and better ways of choosing b_{EM} could be derived. In the dairy cattle example the AI algorithm performed without problems but with the much more complicated model for sheep data, AI steps pointing out of parameters space were encountered during the first round of iteration, but admissible estimates were recovered with a single 1/200 step towards EM information (Table 2).

The probability of getting intermediate estimates outside parameter space tends to increase with the number of parameters to be estimated. A poor choice of starting values for the (co)variance parameters also can create problems in the first rounds of iteration. In such cases our approach was to gradually mix AI and EM information matrices. More efficient solutions to this problem might exist. One possibility is to compute updates to a transformation of the parameters in the likelihood function as was suggested by Lindstrom and Bates [15]. Updating, e.g. on a Cholesky decomposition of all the (co)variance matrices will ensure that all estimates of the covariance matrices themselves will remain within parameter space.

Meyer and Smith [20] investigated several alternative schemes for maximizing the restricted likelihood, including Newton-Raphson and Fisher scoring. They used exact second differentials and considered a number of techniques to ensure that the likelihood increases in each iteration. Also in their study second derivative methods typically converged in much smaller number of iterations than derivative free methods. They were using a method of backward differentiation by Smith [29] that requires 6N times as much computation as when evaluating only the likelihood. Our approach of average information requires much less work in each iteration, typically less than required for calculating the first derivative. The computation of exact second differentials also have large memory requirements.

It might be thought that using second differentials opposed to average information should speed convergence. However, for these there are two different possibilities, corresponding to observed and expected information, i.e. Newton-Raphson or Fisher scoring, respectively. There are no consensus on which is better in all circumstances. Jennrich and Sampson [8] and Jennrich and Schlucter [9] suggest that Newton-Raphson is less robust against poor starting values than Fisher scoring and they advocate switching between algorithms. Meyer and Smith [20] failed to show any consistent advantages of either of the two methods but found, however, mixed algorithm starting with Fisher scoring and later on switching to Newton-Raphson to be more robust than either of two alone. Gilmour *et al.* [4] saw in several situations that AI and Newton-Raphson algorithms converged with the same small number of iterations.

Meyer and Smith [20] considered also reparameterization of (co)variance matrices into Cholesky scale. As was mentioned earlier we expect that their suggestions could improve the robustness of the AI algorithm as well. The transformations require only little extra computations once the first and second differentials are available. However, the transformation may change the shape of the likelihood so that the maximum is more difficult to locate. Thus more experience is needed before a general recommendation can be made. In its simplicity the combined EM and AI algorithm may be of interest because by suitable choice of b_{EM} it will always lead to a point in the parameter space with increased likelihood. It could also be thought of as a Marquardt type algorithm by making the adjustments in the canonical parameter space.

ACKNOWLEDGEMENTS

The Danish Institute of Animal Science, Foulum, Denmark is thanked for financial support during the visits by E. Mäntysaari and R. Thompson. The National Research Council in Denmark is thanked for providing access to supercomputing facilities and finally the Agricultural Research Centre of Finland is thanked for support for J. Jensen and R. Thompson during several visits.

REFERENCES

[1] Dempster, A.P., Laird, N.M. and Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *J.Roy. Statist. Soc.*, **B39**, 1-38.

[2] Duff, I.S., Erisman, A.M. and Reid, J.K., 1989. *Direct Method for Sparse Matrices.* Clarendon Press, Oxford, England.

[3] George, A. and Liu, J.W., 1981. *Computer Solution of Large Sparse Positive Definitive Systems.* Prentice-Hall Inc., Englewood Cliffs, New Jersey.

[4] Gilmour, A.R., Thompson, R. and Cullis, B.R., 1995. Average information : An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, **51**, 1440-1450.

[5] Graser, H.U., Smith, S.P. and Tier, B., 1987. A derivative-free approach for estimating variance components in animal models by REML. *J. Anim. Sci.*, **64**, 1362-1370.

[6] Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.*, **72**, 320-340.

[7] Henderson, C.R., 1973. Sire evaluation and genetic trends. Proc. of the animal breeding symposium in honour of Dr. Jay L. Lush, Champaign, Illinois, *American Soc. of Anim. Sci.*, 10-41.

[8] Jennrich, R.I. and Sampson, P.F., 1976. Newton-Raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics*, **18**, 11-17.

[9] Jennrich, R.I. and Schluchter, M.D., 1986. Unbalanced repeated measures models with structural covariance matrices. *Biometrics*, **42**, 805-820.

[10] Jensen, J. and Mao, I.L., 1988. Transformation algorithms in analysis of single-trait and of multitrait models with equal design matrices and one random factor per trait : A review. *J. Anim. Sci.*, **66**, 2750-2761.

[11] Jensen, J. and Madsen, P., 1994. DMU : A package for the analysis of multivariate mixed models. Proc. of 5th World Congress on Genetics Applied to Livestock Production, **22**, 45-46.

[12] Johnson, D.L. and Thompson, R., 1995. Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. *J. Dairy Sci.*, **78**, 449-456.

[13] Juga, J. and Thompson, R., 1992. A derivative-free algorithm to estimate bivariate (co)variance components using canonical transformations and estimated rotations. *Acta Agric. Scand., Sect. A. Animal Sci.*, **42**, 191-197.

[14] Lin, C.Y. and Smith, S.P., 1990. Transformation of multitrait to unitrait mixed model analysis of data with multiple random effects. *J. Dairy Sci.*, **73**, 2494-2502.

[15] Lindstrom, M.J. and Bates, D.M., 1988. Newton-Raphson and EM algorithms for linear mixed-effect models for repeated-measures data. *J. Amer. Statist. Assoc.*, **83**, 1014-1022.

[16] Madsen, P., Jensen, J. and Thompson, R., 1994. Estimation of (co)variance components by REML in multivariate mixed linear models using average of observed and expected information. Proc. of World Congress on Genetics Applied to Livestock Production, Guelph, Ontario, Canada, **22**, 19-22.

[17] Marquardt, D., 1963. An algorithm for least-squares estimation of non-linear parameter. *J. Appl. Math.*, **11**, 431.

[18] Meyer, K., 1985. Maximum likelihood estimation of variance components for a multivariate model with equal design matrices. *Biometrics*, **41**, 153-165.

[19] Meyer, K., 1991. Estimating variances and covariances for multivariate animal models by restricted maximum likelihood. *Genet. Sel. Evol.*, **23**, 67-83.

[20] Meyer, K. and Smith, S.P., 1996. Restricted maximum likelihood estimation for animal models using derivatives of the likelihood. *Genet. Sel. Evol.*, **28**, 23-49.

[21] Misztal, I., 1994. Comparison of computing properties of derivative and derivative-free algorithms in variance-component estimation by REML. *J. Anim. Breed. Genet.*, **111**, 346-355.

[22] Misztal, I. and Perez-Enciso, M., 1993. Sparse matrix inversion for restricted maximum likelihood estimation of variance components by expectation maximization. *J. Dairy Sci.*, **76**, 1479-1483.

[23] Mäntysaari, E.A. and Van Vleck, L.D., 1989. Restricted maximum likelihood estimates of variance components from multitrait sire models with large number of fixed effects. *J. Anim. Breed. Gent.*, **106**, 409-422.

[24] Nelder, J.A. and Mead, R., 1965. A simplex method for function minimization. *Computer J.*, **7**, 308-313.

[25] Patterson, H.D. and Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545-554.

[26] Searle, S.R., 1979. *Notes on variance component estimation*. Cornell Univ., Biometrics Unit, New York.

[27] Searle, S.R., 1982. *Matrix Algebra Useful for Statistics*. John Wiley and Sons, New York.

[28] Searle, S.R., Casella, G. and McCulloch, C.E., 1992. *Variance Components*. John Wiley and Sons., New York.

[29] Smith, S.P., 1995. Differentiation of the Cholesky algorithm. *J. Computational and Graphical Statistics*, **4**, 134-147.

[30] Smith, S.P. and Graser, H.U., 1986. Estimating variance components by restricted maximum likelihood. *J. Dairy Sci.*, **69**, 1156-1165.

[31] Van Vleck, L.D. and Boldman, K.G., 1993. Sequential transformation for multiple traits for estimation of (co)variance components with a derivative-free algorithm for restricted maximum likelihood. *J. Anim. Sci.*, **71**, 836-844.