Genet. Sel. Evol. (1989) 21, 317-340

Original article

Restricted maximum likelihood to estimate variance components for animal models with several random effects using a derivative-free algorithm

K. Meyer

Edinburgh University, Institute of Animal Genetics, West Mains Road, Edinburgh EH9 3JN, Scotland, UK

(received 21 March 1988, accepted 11 January 1989)

Summary – A method is described for the simultaneous estimation of variance components due to several genetic and environmental effects from unbalanced data by restricted maximum likelihood (REML). Estimates are obtained by evaluating the likelihood explicitly and using standard, derivative-free optimization procedures to locate its maximum. The model of analysis considered is the so-called Animal Model which includes the additive genetic merit of animals as a random effect, and incorporates all information on relationships between animals. Furthermore, random effects in addition to animals' additive genetic effects, such as maternal genetic, dominance or permanent environmental effects are taken into account. Emphasis is placed entirely upon univariate analyses. Simulation is employed to investigate the efficacy of three different maximization techniques and the scope for approximation of sampling errors. Computations are illustrated with a numerical example.

variance components – restricted maximum likelihood – animal model – additional random effects – derivative – free approach

Résumé – Utilisation du maximum de vraisemblance restreint et d'un algorithme sans dérivation, pour estimer les composantes de variance d'un caractère, selon un modèle animal avec plusieurs effets aléatoires. *On décrit une méthode pour estimer simultanément les composantes de la variance d'un seul caractère, dues au milieu ou plusieurs effets génétiques. La méthode admet des données non équilibrées et se fonde sur le maximum de vraisemblance restreint (« REML »). Les composantes estimées sont obtenues par l'évaluation explicite de la fonction de vraisemblance, dont on recherche le maximum par des techniques générales d'optimisation, ne nécessitant pas le calcul des dérivées. Le modèle d'analyse est un « modèle animal », où l'on considère la valeur génétique individuelle des animaux comme un effet aléatoire, et tient compte de toute l'information généalogique disponible. Des effets aléatoires complémentaires (effets maternels génétiques, effets de dominance, effets de milieu permanent) sont aussi pris en compte. La simulation est utilisée pour évaluer l'efficacité de trois techniques de maximisation, et pour déterminer approximativement les distributions des estimateurs. Les calculs sont illustrés par un exemple numérique.*

composantes de la variance – maximum de vraisemblance restreint – modèle animal – effets aléatoires complémentaires – approche sans dérivation

INTRODUCTION

Over the last decade, restricted maximum likelihood (REML) has become the method of choice for estimating variance components in animal breeding and related disciplines trying to partition the phenotypic variation into genetic and other components. This has been facilitated not only by an increase in the general level of computational resources available, but by the development of numerous specialized algorithms, exploiting specific features of the data structure or model of analysis as well as utilizing a variety of numerical techniques.

So far, REML has found most practical use in the analysis of dairy cattle data under a "sire model". For this model, records of progeny are used only to obtain information on half of their sires' breeding value, while dams and relationships between females are ignored. Recently, interest has increased in more detailed models, in particular the conceptually simplest breeding value or "Animal Model" (AM) where each record is taken to provide information on the additive genetic merit of the animal measured. By including animals which do not have records themselves but are parents, this allows for all information on relationships to be taken into account.

A large proportion of REML applications have been restricted to models with one random factor (*e.g.* sires) apart from random residual errors, estimating two variance components only in a univariate analysis, or p $(p+1)$ for a multivariate analysis of p traits. While algorithms for more complicated models have been described, they are by and large computationally demanding. Often they involve inversion of a matrix of size equal to the total number of levels of all random effects fitted. This can be prohibitive for practically sized data sets. Thus REML has found comparatively little use so far for models fitting several random effects.

Maximum likelihood estimation involves, by definition, location of the maximum of the likelihood function for a given set of data, model of analysis and parameters to be estimated. Estimating variance components for unbalanced data generally requires iterative schemes. Standard textbooks on numerical analysis classify procedures to find the optimum (minimum or maximum) of a function according to the amount of information required from derivatives of the function. The so-called Newton methods utilize both first and second derivatives, *i.e.* geometrically speaking slope and curvature, and are thus quickest to converge. Methods relying on first derivatives only include steepest descent, conjugate gradient and Quasi-Newton procedures approximating second derivatives. Finally, there are derivative-free methods involving direct search strategies or numerical approximation of derivatives (see for example Gill *et al.*, 1981).

In the main, REML algorithms currently employed in animal breeding fall into the first two categories. Fisher's Method of Scoring is a special case of the Newton procedures, requiring expected values of second derivatives of the log likelihood function ($\mathcal{L}$) to be evaluated. As these are often difficult to obtain, Expection-Maximization (EM) type algorithms (Dempster *et al.*, 1977), exploiting first derivative information, are used more widely.

A derivative-free REML algorithm has been suggested by Graser *et al.* (1987) for univariate analyses to estimate the additive genetic and error variance under an animal model. Exploiting sparse matrix techniques, they showed that their

procedure was suitable for data from large selection experiments involving several thousand animals.

This paper describes the use of a derivative-free approach to estimate variance components by REML for AMs which include not only animals' additive genetic merit but also additional random effects, and thus cover a wide range of models suitable for the analysis of animal breeding data. Univariate analyses only are considered at present, extensions to multivariate situations will be discussed elsewhere.

CALCULATING THE LIKELIHOOD

The Model

Let:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{b} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{e} \quad [1]$$

denote the linear model of analysis with:
$\mathbf{Y}$ the vector of N observations,
$\boldsymbol{b}$ the vector of NF fixed effects (including any linear of higher order covariables)
$\mathbf{X}$ the NxNF incidence or design matrix for fixed effects with column rank NF*,
$\boldsymbol{u}$ the vector of all NR random effects fitted,
$\mathbf{Z}$ the NxNR incidence matrix for random effects, and
$\boldsymbol{e}$ the vector of N random residual errors.
Assume:

$$V(\boldsymbol{u}) = \mathbf{G},$$
$$V(\boldsymbol{e}) = \mathbf{R} \quad \text{and}$$
$$\text{Cov}\ (\boldsymbol{u}, \boldsymbol{e}') = \mathbf{0}$$

which gives:

$$V(\boldsymbol{y}) = \mathbf{ZGZ}' + \mathbf{R}.$$

The mixed model equations (MME) pertaining to [1] are then (Henderson, 1973):

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\boldsymbol{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\boldsymbol{y} \end{bmatrix} \quad [2]$$

or $\mathbf{C}\ \hat{\mathbf{F}} = \boldsymbol{r}$. If $\mathbf{C}$ is not of full rank, as it is often the case, estimates for $\boldsymbol{b}$ are not unique.

The Likelihood

REML operates on the likelihood of linear functions of the data vector with expectations zero, so-called error contrasts, or, equivalently, on the part of the likelihood (of the data vector) which is independent of fixed effects. This results in the loss in degres of freedom due to fitting of fixed effects being taken into account (Patterson & Thompson, 1971). For $\boldsymbol{Y} \sim \text{N}(\mathbf{X}\boldsymbol{b}, \boldsymbol{V})$, the log likelihood is (*e.g.* Harville 1977):

$$\log \mathcal{L} = -1/2[\text{const} + \log|\mathbf{V}| + \log|\mathbf{X}^{*\prime}\mathbf{V}^{-1}\mathbf{X}^{*}| + (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{b}})'\ \text{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{b}}) \quad [3]$$

where $\mathbf{X}^*$ (of order N×NF*) is a full rank submatrix of $\mathbf{X}$. Using matrix equalities given by Harville (1977) and Searle (1979), [3] can be rewritten as:

$$-2 \log \mathcal{L} = \text{const} + \log|\mathbf{R}| + \log|\mathbf{G}| + \log|\mathbf{C}^*| + \boldsymbol{y}'\mathbf{P}\boldsymbol{y} \quad [4]$$

where $\mathbf{C}^*$ is the coefficient matrix in [2] with $\mathbf{X}$ replaced by $\mathbf{X}^*$, and $\mathbf{P}$ is a matrix:

$$\begin{aligned} \mathbf{P} &= \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \\ &= \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{V}^{-1}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{V}^{-1} \end{aligned} \quad [5]$$

Calculation of the first two terms required in [4] depends on the specific structure of $\mathbf{R}$ and $\mathbf{G}$ in a given analysis. The latter two, however, can be determined in a general fashion, as suggested by Graser *et al.* (1987), by Gaussian Elimination (as described in most Numerical Analysis textbooks, or by Smith & Graser (1986)) applied to the mixed model array: the coefficient matrix in [2] augmented by the right hand side and a quadratic in the data vector.

Calculation of $y'Py$ and log | C|*

The mixed model array for [1] is:

$$\mathbf{M} = \begin{bmatrix} \boldsymbol{y}'\mathbf{R}^{-1}\boldsymbol{y} & \boldsymbol{y}'\mathbf{R}^{-1}\mathbf{X} & \boldsymbol{y}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{X}'\mathbf{R}^{-1}\boldsymbol{y} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\boldsymbol{y} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \quad [6]$$

"Absorbing" rows and columns pertaining to random effects into the rest to the matrix then gives:

$$\begin{bmatrix} \boldsymbol{y}'\mathbf{V}^{-1}\boldsymbol{y} & \boldsymbol{y}'\mathbf{V}^{-1}\mathbf{X} \\ \mathbf{X}'\mathbf{V}^{-1}\boldsymbol{y} & \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \end{bmatrix} \quad [7]$$

and eliminating rows and columns for fixed effects correspondingly, yields $\boldsymbol{y}'\mathbf{P}\boldsymbol{y}$, the weighted sum of squared residuals required to evaluate log $\mathcal{L}$. Absorption is most easily carried out by Gaussian elimination: repeated absorption of one row and column at a time. This will also allow log $|\mathbf{C}^*|$ to be determined simultaneously.

Subdivide $\mathbf{M}$ of size K×K (K=NF+NR+1) with elements m_{ij} and column vectors $\boldsymbol{m}_i$ into rows 1 to K−1, and row K:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{\text{K-1}} & \boldsymbol{m}_{\text{K}} \\ \mathbf{m}' & m_{\text{KK}} \end{bmatrix}$$

Partitioned matrix results then give

$$\log|\mathbf{M}| = \log m_{\text{KK}} + \log|\mathbf{M}^*_{\text{K-1}}|$$

with $\mathbf{M}_{\text{K-1}}{}^* = \mathbf{M}_{\text{K-1}} - \boldsymbol{m}_{\text{K}}\mathbf{m}'_{\text{K}}/m_{\text{KK}} = \{m_{ij} - m_{\text{iK}}\, m_{jk}/m_{\text{KK}}\} = \{m_{ij}{}^*\}$ the matrix resulting when "absorbing" row and column K, or "pivoting" on m'_{KK}. Repeated use of this result shows that the required determinant is then simply the sum of the log of pivots log $m_{ii}{}^*$, $i = 2,\ldots,$ K) arising when absorbing all rows and columns of $\mathbf{M}$ into the first row, as required to evaluate $\boldsymbol{y}'\mathbf{P}\boldsymbol{y}$. If $\mathbf{X}$ is not of full rank, $\mathbf{M}$ has to be set up replacing $\mathbf{X}$ by $\mathbf{X}^*$ or, equivalently, absorptions have to be carried out skipping the NF-NF* rows with zero pivots.

Univariate analyses

Results presented so far hold for any model of form [1]. Consider now a univariate analysis with identically and independently distributed errors, *i.e.*

$$\mathbf{R} = \sigma_E^2 \mathbf{I} \qquad [8]$$

For given values of the other variance components, the error variance can be estimated directly in this case, from the residual sums of squares as (see Harville, 1977; or Graser *et al.*, 1987)

$$\sigma_E^2 = \boldsymbol{y}'\mathbf{P}\boldsymbol{y}/(\text{N-NF}^*) \qquad [9]$$

Let the other parameters to be estimated, *i.e.* (co)variances of the random effects fitted, be denoted by σ_i with $i = 1, ..., p-1$, and p *the total number of components with* $\sigma_p = \sigma_E^2$.

As discussed by Harville & Callanan (1988), a function of REML estimates of a set of parameters is also the REML estimate of this function. Hence, instead of maximizing log $\mathcal{L}$ with respect to the p components σ_i , we can reparameterize to σ_E^2 and $p-1$ functions f_i (σ_i, σ_E^2) of the other components and the error variance. An obvious choice is to express the σ_i as a proportion $(\lambda_i = \sigma_i/\sigma_E^2)$ of the latter, so that having found REML estimates of σ_E^2 and the λ_i, we can estimate $\widehat{\sigma}_i = \widehat{\lambda}_i\, \widehat{\sigma}_E^2$.

Furthermore, for fixed values of λ_i, log $\mathcal{L}$ attains its maximum with respect to σ_E^2 at the REML estimate of σ_E^2. This allows estimation to be conducted in two steps: Firstly, a "concentrated" likelihood is maximized with respect to the λ_i only which yields REML estimates $\widehat{\lambda}_i$. Secondly, $\widehat{\sigma}_E^2$ is obtained (from [9]) for the $\widehat{\lambda}_i$ (Harville & Callanan, 1988). The advantage of this approach is that it reduces the dimension of the numerical search for the maximum of log $\mathcal{L}$ by one. As the number of iterates and likelihoods to be evaluated to find the maximum of log $\mathcal{L}$ usually increases substantially with the number of parameters to be estimated, this can lead to a considerable saving in computational resources required.

Log $|\mathbf{R}|$ *and log* $|\mathbf{G}|$

From [8] it follows immediately that:

$$\log|\mathbf{R}| = \text{N} \log \sigma_E^2 \qquad [10]$$

Log $|\mathbf{G}|$ depends on the random effects fitted. For the simplest model with animals as the only random effect, as considered by Graser *et al.* (1987):

$$\boldsymbol{u} = \mathbf{a},$$
$$\mathbf{G} = \sigma_A^2\, \mathbf{A}, \text{ and}$$
$$\log|\mathbf{G}| = \text{NA} \log \sigma_A^2 + \log|\mathbf{A}| \qquad [11]$$

where σ_A^2 is the additive genetic variance, $\mathbf{A}$ the numerator relationship matrix between animals, $\mathbf{a}$ the vector of (direct) genetic effects for animals, and NA denotes the number of animals. Since log $|\mathbf{A}|$ does not depend on the parameters to be estimated, it is a constant and does not need to be calculated in order to maximize log $\mathcal{L}$. The inverse of $\mathbf{A}$ is required in [6] (for $\mathbf{G}^{-1}$) though, but this can be set up efficiently from a list of pedigree information, following rules described, for instance, by Quaas (1976).

Often, animals in the same environmental subclass are subject to a so-called common environment effect, for example a pen or litter effect in pig or mouse data.

Let $\mathbf{c}$ of length NC denote a vector of such effects to be included in the model of analysis, with

$V(\mathbf{c}) = \sigma_C^2\ \mathbf{I}$, and
Cov $(\mathbf{a}, \mathbf{c}') = \mathbf{0}$.

This gives:

$$\begin{aligned} \boldsymbol{u}' &= (\boldsymbol{a}'\ \boldsymbol{c}'), \\ \mathbf{G} &= \text{Diag}\ \{\sigma_A^2\ \mathbf{A};\ \sigma_C^2\ \mathbf{I}\}, \text{ and} \\ \log|\mathbf{G}| &= \mathbf{NA} \log \sigma_A^2 + \text{NC} \log \sigma_C^2 + \log|\mathbf{A}| \end{aligned} \quad [12]$$

In other cases, the model of analysis may involve two random effects for each animal. Let $\boldsymbol{m}$, of length NA, denote the second animal effect and assume each element has variance σ_M^2. If there are repeated records per animal for a trait, $\boldsymbol{m}$ represents the permanent effects due to animals, excluding additive genetic effects. These are usually assumed to be uncorrelated with any other effects in the model, so that

$$\begin{aligned} \boldsymbol{u}' &= (\boldsymbol{a}'\ \boldsymbol{m}'), \\ \mathbf{G} &= \text{Diag}\ \{\sigma_A^2\ \mathbf{A};\ \sigma_M^2\mathbf{I}\}, \text{ and} \\ \log|\mathbf{G}| &= \text{NA}\ (\log \sigma_A^2 + \log \sigma_M^2\ ) + \log|\mathbf{A}| \end{aligned} \quad [13]$$

If $\boldsymbol{m}$ had variance $\sigma_M^2\ \mathbf{D}$, [13] would be augmented by $\log|\mathbf{D}|$. As with $\log|\mathbf{A}|$, this term is constant and does not need to be evaluated. Note though that $\mathbf{G}^{-1}$ and consequently $\mathbf{D}^{-1}$ is required in [6]. A typical example for this kind of structure is a model where $\boldsymbol{m}$ stands for dominance effects, σ_M^2 for the respective variance and $\mathbf{D}$ for the dominance covariance matrix among animals.

For other traits, for example measures of reproductive performance, we distinguish between a direct and a maternal (or paternal) additive-genetic component, allowing for a covariance between the two. In that situation, there may not be a record supplying information on $\boldsymbol{m}$ for each animal, but information is acquired indirectly *via* links arising from the genetic covariance and relationships. With σ_{AM} denoting the covariance between $\boldsymbol{a}$ and $\boldsymbol{m}$ and r_{AM} the corresponding correlation,

$$\mathbf{G} = \begin{bmatrix} \sigma_A^2\mathbf{A} & \sigma_{AM}\mathbf{A} \\ \sigma_{AM}\mathbf{A} & \sigma_M^2\mathbf{A} \end{bmatrix}$$

and partitioned matrix results give

$$\log|\mathbf{G}| = \text{NA}[\log \sigma_A^2 + \log \sigma_M^2 + \log\ (1 - r_{AM}^2)] + 2\log|\mathbf{A}| \quad [14]$$

For all models discussed so far, computational requirements to determine the part of $\log|\mathbf{G}|$, which depends on the parameters to be estimated, are trivial. This results from random effects being either uncorrelated, so that $\mathbf{G}$ is blockdiagonal ([12] and [13]), or $\mathbf{G}$ being the direct product of a matrix of parameters and a matrix describing correlations amongst levels of random effects as in [14]. Extensions to other models are straightforward, as long as $\mathbf{G}$ can be partitioned into blocks of such structure. For example, fitting permanent environmental effects ($\mathbf{c}$) as well as direct and maternal additive genetic effects, [14] would be augmented simply by (NC $\log \sigma_C^2$), provided $\mathbf{c}$ was uncorrelated to $\mathbf{a}$ and $\boldsymbol{m}$. Table I summarizes $\log \mathcal{L}$ for 10 models which may arise in the analysis of animal breeding data, with up to 3 random effects and involving up to 5 (co) variance components. Otherwise, $\mathbf{G}$ (or a submatrix thereof) needs to be set up explicitly and its determinant be obtained

using techniques as described above for log $|\mathbf{C}^*|$. For instance, if $\mathbf{G}$ contained a block of form

$$\begin{bmatrix} \sigma^2_A \mathbf{A} & \sigma_{AC}\mathbf{B} \\ \sigma_{AC}\mathbf{B}' & \sigma^2_C \mathbf{A} \end{bmatrix}$$

the contribution to log $|\mathbf{G}|$ would be

$$\text{NA} \log \sigma^2_A + \text{NC} \log \sigma^2_C + \log |\mathbf{A}| + \log |\mathbf{D} - r^2_{AC}\, \mathbf{B}'\mathbf{A}^{-1}\mathbf{B}| \qquad [15]$$

Table I. Summary of REML log likelihoods for individual Animal Models with up to 3 random effects and 5 (co)variance components; see section 2 for notation.

$-2 \log \mathcal{L} = \text{const} + \mathbf{y}'\mathbf{P}\mathbf{y} / \sigma^2_E + \log |\mathbf{C}^*| + \text{model-specific terms}$

Model No.	*Effects fitted*	*Covariances* v (m)	Cov (a,m′)	*Model-specific terms:*
1	**a**	-	-	NA log σ^2_A + (N-NF*NA) log σ^2_E + *log* $\|\mathbf{A}\|$
2	**a, c**	-	-	NA log σ^2_A + NC log σ^2_C + (N-NF*NA-NC) log σ^2_E + *log* $\|\mathbf{A}\|$
3	**a, m**	σ^2_M **A**	**0**	NA (log σ^2_A + log σ^2_M) + (N-NF*-2NA) log σ^2_E + 2 *log* $\|\mathbf{A}\|$
4	**a, m**	σ^2_M **A**	σ_{AM} **A**	NA [log σ^2_A + log σ^2_M + log (1-r^2_{AM})] + (N-NF*-2NA) log σ^2_E + 2 *log* $\|\mathbf{A}\|$
5	**a, m**	σ^2_M **I**	**0**	NA (log σ^2_A + log σ^2_M) + (N-NF*2NA) log σ^2_E + *log* $\|\mathbf{A}\|$
6	**a, m**	σ^2_M **D**	**0**	NA (log σ^2_A + log σ^2_M) + (N-NF*-2NA) log σ^2_E + *log* $\|\mathbf{A}\|$ + *log* $\|\mathbf{D}\|$
7	**a, m, c**	σ^2_M **A**	**0**	NA (log σ^2_A + log σ^2_M) + NC log σ^2_C + (N-NF*-2NA-NC) log σ^2_E + 2 *log* $\|\mathbf{A}\|$
8	**a, m, c**	σ^2_M **A**	σ_{AM} **A**	NA [log σ^2_A + log σ^2_M + log (1-r^2_{AM})] + NC log σ^2_C + (N-NF* -2NA-NC) log σ^2_E + 2 *log* $\|\mathbf{A}\|$
9	**a, m, c**	σ^2_M **I**	**0**	NA(log σ^2_A + log σ^2_M) + NC log σ^2_C + (N-NF*-2NA-NC) log σ^2_E + *log* $\|\mathbf{A}\|$
10	**a, m, c**	σ^2_n **D**	**0**	NA (log σ^2_A + log σ^2_M) + NC log σ^2_C + (N-NF*-2NA-NC) log σ^2_E + *log* $\|\mathbf{A}\|$ + *log* $\|\mathbf{D}\|$

Assume V(**a**) = σ^2_A **A**, Cov(**a**, **c**′) = Cov(**m**, **c**′) = 0 and V(**c**) = σ^2_C **I** for all models.
Terms are assumed to be the result of Gaussian Eliminations performed for **M** with σ^{-2}_E factored out.
Terms in light italic are constant and not required to maximize the likelihood.

Computational Considerations

Typically, the augmented coefficient matrix $\mathbf{M}$ is very large but also very sparse. Hence use of sparse matrix techniques, storing the non-zero elements of $\mathbf{M}$ only, is advantageous and allows matrices of order of thousands to be handled. Since $\mathbf{M}$ is symmetric, only the lower (or upper) triangle is required. One form of sparse matrix storage, described in standard text books such as Knuth (1973), is a so-called "linked list" . Such linked lists, one list for each row of $\mathbf{M}$ in conjunction with a vector pointing to the first element in each row, are well suited, and allow the Gaussian Elimination steps required to evaluate $\boldsymbol{y}'\mathbf{P}\boldsymbol{y}$ and $\log|\mathbf{C}^*|$ to be carried out efficiently.

In setting up $\mathbf{M}$, the order of equations can be of vital importance as it affects the "fill-in" during the absorption process, *i.e.* the number of additional non-zero off-diagonal elements arising. For computational efficiency this should be kept as small as possible. There is extensive literature concerned with numerical operations on sparse matrices. Tewarson (1973), for example, discusses techniques for the choice of pivot in each Gaussian Elimination step which yields the least local fill-in, and also considers the scope of *a priori* column permutations. A number of strategies for re-ordering matrices exists, often utilizing graph theory; see for instance Duff *et al.* (1986). Such general techniques, making little or no assumptions about the matrix structure can be computationally expensive. This may be prohibitive for situations where the direct solution of a large sparse system of equations is required a few times only, but may be worthwhile for our application where numerous likelihood evaluations are to be performed. Future research should consider this topic.

In the meantime, critical inspection of the data and relationship structure with their implications for the pattern of off-diagonal elements in the mixed model array, and judicious ordering of effects may achieve a large proportion of the potential benefits from general reordering algorithms. A standard strategy in attempting to minimize fill-in is to process rows with the fewest off-diagonals first. Graser *et al.* (1987) therefore suggested selection of pivots corresponding to the youngest animals first. For the models with several random effects for each animal, these should be assigned to successive rows. In other cases, it may be possible to exploit additional features of the data structure. For data from a multi-generation selection experiment with selection within families, for example, grouping of animals according to female "founders" appears preferable to a grouping according to generation. On the other hand, if animals are nested within contemporary (fixed) groups, it may be advantageous to order equations so that animals directly follow their group effects.

For $\mathbf{R}$ of form [8], σ_E^{-2} is usually factored from [6]. In this case, calculations to determine $\boldsymbol{y}'\mathbf{P}\boldsymbol{y}$ and $\log|\mathbf{C}^*|$, as described above, do not yield the terms required in [4] directly, but $(\boldsymbol{y}'\mathbf{P}\boldsymbol{y}\,\sigma_E^2)$ and $(\log|\mathbf{C}^*| + (\mathrm{NF}^*+\mathrm{NR})\log\sigma_E^2)$, which has to be born in mind when assembling the likelihood.

MAXIMIZING THE LIKELIHOOD

Choice of a strategy to locate the maximum of the likelihood function, or equivalently the minimum of $-2\log\mathcal{L}$, is determined by several considerations. Firstly, each function evaluation, *i.e.* likelihood calculation, is computationally very much

more demanding than any calculations required by the optimization procedure as such. Hence, a method which requires a small number of function evaluations in each iterate is desirable. Secondly, the procedure should be robust, *i.e.* cope with starting values for parameters considerably different from the eventual estimates, and should be little affected by problems of numerical accuracy, yielding sufficiently precise estimates of the minimum even for very flat functions. Thirdly, constraints on the parameter space should be accommodated and, preferably, not require extra function values or reduce the speed of convergence.

The suitability of three different approaches was examined using simulated data for models 1, 2, 4 and 8 as specified in Table I. Records were sampled according to the model of analysis for one or several generations (up to four), each comprising a given number of full-sib families (ranging from 25 to 800) of variable size (2 to 10), with dams nested within sires and each sire mated to a specified number of dams (1 to 5). Error variances were estimated directly, while all other components were expressed as a proportion of the phenotypic variance, *i.e.*, θ_A, θ_M, θ_C and θ_{AM} for σ^2_A, θ^2_M, θ^2_C and σ_{AM}, respectively. Obviously, θ_A is the heritability and θ_C what is commonly referred to as "c^2 effect" . As described above, this reduced the dimension of search to 1, 2, 3 and 4 for Models 1, 2, 4 and 8, respectively. This parameterization rather than expressing components as a proportion of the error variance (λ_i) was chosen since it allowed checks for parameter estimates out of bounds more readily and, for the limited cases examined, as it appeared to be more robust against bad starting values.

Quadratic approximation

For a model with animals as the only random effect, Graser *et al.* (1987) fitted a quadratic function in $r = \sigma^2_A/\sigma^2_E$ to the log likelihood, predicting the maximum of log $\mathcal{L}$ as the maximum of this function. For one parameter, this required function values for 3 different r values per approximation. Having calculated log $\mathcal{L}$ for 3 initial points, each iterate then involved one function evaluation, for r^* which maximized the quadratic function of the previous step. This value and those pertaining to the two r values either side closest to r^* were then utilized in determining the next quadratic approximation to log $\mathcal{L}$. As reported by Graser *et al.* (1987), simulations for this model showed rapid convergence. A bad initial guess for r generally did not affect the estimation procedure greatly, as long as the three points in the initial approximation spanned a sufficiently large range. Though the number of iterates and likelihood evaluations required tended to increase, the same maximum of log $\mathcal{L}$ as for "good" starting values was attained without increasing computational demands excessively.

This approach extends to the case of multiple parameters. For $\boldsymbol{t}$, with elements θi, denoting the vector of parameters with respect to which log $\mathcal{L}$ is to be maximized, and log $\mathcal{L}$ $(\boldsymbol{t})$ the corresponding log likelihood, the quadratic approximation is:

$$\log \mathcal{L}\ (\boldsymbol{t}) = q + \boldsymbol{q}'\boldsymbol{t} + \boldsymbol{t}'\boldsymbol{Q}\boldsymbol{t} \qquad [16]$$

The vector maximizing [16] is then, for $\mathbf{Q}$ positive definite,

$$\boldsymbol{t}^* = -0.5\ \mathbf{Q}^{-1}\boldsymbol{q} \qquad [17]$$

For p parameters, a total of $z = 1 + p(p+3)/2$ different values of $\boldsymbol{t}$ and $\log \mathcal{L}(\boldsymbol{t})$ are required in each iterate to set up and solve a system of z equations for the intercept q, the vector of linear coefficients $\boldsymbol{q}$ and the symmetric matrix of quadratic coefficients $\mathbf{Q}$. This number increases rapidly with the number of paramaters, *e.g.* $z = 6$, 10, 15 and 21 for $p = 2$, 3, 4 and 5, respectively.

For one parameter, choice of the point to be replaced in each iterate was straightforward. In the multi-dimensional case, however, it was less obvious. Two strategies were explored. After z initial points had been obtained, the first involved, as for $p = 1$, in the regular case one function evaluation per iterate, *i.e.* calculation of $\log \mathcal{L}(\boldsymbol{t}^*)$ for $\boldsymbol{t}^*$ from the last iterate. This new point was added to the set of z points which formed the basis to predict $\boldsymbol{t}^*$ in the previous step. The worst of the resulting set of $z + 1$ points was then eliminated, and a new vector $\boldsymbol{t}^*$ determined. If the quadratic approximation failed, *i.e.* if $\log \mathcal{L}(\boldsymbol{t}^*)$ was lower than all z function values in the set, $\boldsymbol{t}^*$ was replaced by $(\boldsymbol{t}^* + \boldsymbol{t}_m) / 2$, where $\boldsymbol{t}_m$ was the parameter vector with highest function value in the set. If necessary, this was repeated until the replacement was successful. Hence, each iterate increased the average likelihood of the z current points.

The second strategy comprised z function evaluations per iterate. Given a vector of starting values $\boldsymbol{t}_0$ ($\boldsymbol{t}^*$ from the previous iterate), np vectors $\boldsymbol{t}_i$ were derived by multiplying the i-th element of $\boldsymbol{t}_0$ by a factor reflecting a chosen step size, 1.10 for steps of 10% in this case. Following a scheme described by Nelder & Mead (1965), further parameter vectors were then determined as $(\boldsymbol{t}_i + \boldsymbol{t}_j) / 2$ for $i < j = 0, ..., p$. This yielded the required total of z grid points and subsequent estimate $\boldsymbol{t}^*$. For both strategies, all vectors $\boldsymbol{t}$ were checked for elements out of the parameter space, and if necessary these were set to their respective bounds.

The quadratic approximation performed well for Model 2, though, for the limited number of examples considered, it was not consistently better than the two alternative procedures studied, in terms of the number of likelihood evaluations required. For Models 4 and 8, however, where the data structure was such that only a small proportion of animals had direct information on the second genetic effect, problems of numerical accuracy occurred. Often the system of z equations to be solved was indeterminate or almost so. Typically this yielded non-positive definite estimates of $\mathbf{Q}$ and useless predictions of $\boldsymbol{t}^*$. For the second strategy, an alternative approach, slightly more robust, was tried. This consisted of estimating elements of $\boldsymbol{q}$ and $\mathbf{Q}$ by numerical differentiation, *i.e.* as forward-difference approximations to the first and second derivatives of $\log \mathcal{L}$, respectively.

On the whole, quadratic approximation of the likelihood function involving multiple parameters appeared to be unsuitable as a general search procedure. For a one-dimensional search, however, it performed consistently best among the 3 strategies examined.

Quasi-Newton

Procedures which do not require second derivatives of the function to be minimized, but approximate the Hessian matrix (= matrix of second derivatives) are referred to as Quasi-Newton methods. This approximation is usually performed iteratively, starting from an identity matrix, utilizing rank-two update techniques based on the

vectors of changes in gradients (= first derivatives) and estimates between iterates. While most Quasi-Newton procedures require first derivatives, some are derivative-free, approximating the vector of gradients using finite differences. These have been found to show quadratic convergence and are recommended as the derivative-free methods to be used for smoth functions with continuous derivatives. Further details are beyond the scope of this paper and can be found in standard textbooks, for instance Gill *et al.* (1981).

Statistical library subroutines to find the minimum of a function using a Quasi-Newton algorithm, namely NAG routine E04JBF and IMSL routine ZXMIN, have been applied to $-2 \log \mathcal{L}$. These routines require the user to supply a subroutine to evaluate the function to be minimized, passing the number and vector of parameters as arguments and returning the function value. In addition, starting values for the parameters, the maximum number of iterates or function calls allowed and some criteria of accuracy of evaluation required and rounding errors tolerated have to be specified.

E04JBF provided the facility to constrain parameters between fixed upper and lower bounds (the IMSL equivalent ZXMWD was not tested), *i.e.* 0 to 1 for θ_A, θ_M and θ_C, and -1 to 1 for θ_{AM}. However, to impose these constraints, the routine required function values, setting all parameters simultaneously to their upper or lower limits. Obviously, this violated other restrictions on the parameter space in the genetic context, *i.e.* that the sum of components is bounded correspondingly ($0 \leq \theta_A + \theta_M + \theta_C + \theta_{AM} \leq 1$), and that the absolute value of the genetic correlation has a maximum of unity ($\theta_{AM}^2 \leq \theta_A \theta_M$). Consequently, $\log \mathcal{L}$ could often not be determined for the required constellation since $\mathbf{M}$ became negative definite or $\mathbf{y}'\mathbf{P}\mathbf{y}$ assumed a negative value. Hence minimization was carried out unconstrained. Techniques to implement more complicated constraints exist, and further research should investigate their suitabililty for the kind of models which are of interest in animal breeding.

Unless a parameter vector was encountered for which $-2 \log \mathcal{L}$ could not be evaluated, the Quasi-Newton algorithms performed well for all models examined. Each iterate required p function evaluations to approximate the vector of first derivatives. The number of iterates performed depended on user-specified criteria of accuracy and maximum number of function evaluations allowed. If likelihood functions were wery flat, routines would stop before the minimum of $-2 \log \mathcal{L}$ was determined as accurately as desired, flagging problems of numerical accuracy.

Figure 1 illustrates the typical pattern of changes in likelihood and estimates observed for an analysis under Model 4 for a "good" and a "bad" initial guess of parameter values. The simulated data for this example comprised 2 generations with 100 full-sib families of size 4 to 8 and 25 half-sib families each. Records were sampled for population values of $\sigma_P^2 = 100$ (phenotypic variance), $\theta_A = 0.50, \theta_M = 0.20$ and $\sigma_{AM} = -0.05$. Starting values used were the population values (Set I) and $\sigma_A = 0.30, \sigma_M = 0.15$ and $\sigma_{AM} = 0.05$ (Set II), respectively. For Set I, ZXMIN required 93 likelihood evaluations, for a given significance level of 6 significant digits. For Set II, however, the routine used 204 function calls before it considered the maximum of $-2 \log \mathcal{L}$ to be found, although Figure 1 suggests that likelihood and estimates were essentially identical after 60 function evaluations.

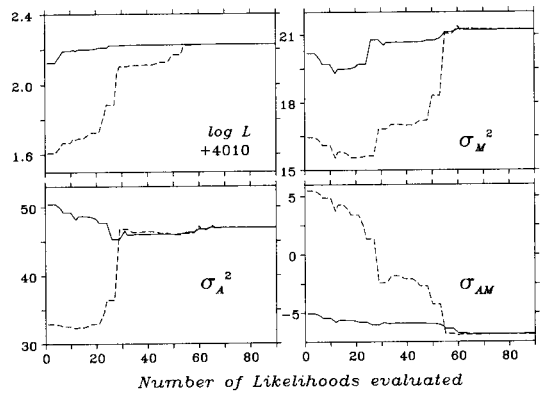

Fig. 1. Changes in likelihood and estimates of variance components for successive function evaluations, using a Quasi-Newton algorithm for "good" (solid line) and "bad" (dashed line) starting values; see text for notation and details of the data set.

Simplex

The Simplex method of Nelder & Mead (1965) is generally advocated as the derivative-free procedure to use if the multivariate function to be minimized is discontinuous, though, initially, it has been developed with the maximization of a likelihood function in mind. It relies on a comparison of function values without attempting to utilize any statistics related to derivatives of the function. Such optimization techniques are generally refered to as direct search procedures. While they often have been developed by heuristic approaches without proof of convergence, they have found to be effective in practice (Swann, 1972).

The Simplex or Polytope method, as some authors prefer to call it to avoid confusion with the Simplex technique used in Linear Programming, was initially suggested by Spendley *et al.* (1962). It operates on a set of parameter vectors and their pertaining function values, which form the vertices of a simplex in the parameter space, hence its name. As reviewed by Swann (1972), it is based on the conceptes of "evolutionary operations" , developed to optimize productivity of industrial plants, in which the parameter space is searched following some geometric configuration. The design which requires the least number of points and hence makes most efficient use of the function values calculated, is a regular simplex. This is defined simply as a set of mutually equidistant points, $n+1$ for a simplex of dimension n. For two dimensions, for example, the regular simplex is an equilateral triangle. A useful property of a regular is that a new simplex can be formed the existing simplex by addition of a single new point.

The search proceeds as follows. To begin, a simplex of specified size is set up, including the point representing an initial guess for the optimum, and corresponding function values are obtained. The aim in each iterate then is to replace the worst point, *i.e.* for a minimization problem the point with the highest function. The new point, defining the next simplex, is chosen so as to preserve the geometric shape in a direction away from the discarded point, but passing through the center of the remaining points. This cycle of rejection and regeneration of a vertex is

repeated until the simplex straddles the optimum. Reducing the size of the simplex, search then recommences with the new, smaller design, or terminates if the simplex becomes smaller than a specified limit (Swann 1972).

Subsequently, the Simplex method has been modified by Nelder & Mead (1965). Abandoning the regularity of design, their procedure allows the simplex to rescale itself automatically in each iterate, changing shape and size according to the landscape of the surface searched. This adaptability is achieved by a combination of so-called reflection, expansion and contraction steps.

Consider a function $F=f(\boldsymbol{t})$ to be minimized with respect to p independent variables, and let $\boldsymbol{t}_0$ denot the guess for the optimum to start with. The initial simplex then comprises $\boldsymbol{t}_0$ and p other points $\boldsymbol{t}_i$ $(i = 1, ..., p)$ obtained by modifying one co-ordinate of $\boldsymbol{t}_0$ at a time by a chosen step size. Each iterate begins by ordering and renumbering the points in the simplex according to the pertaining function values. The following three points are identified: $\boldsymbol{t}_0$ with function value F_0 is now the currently best point in the simplex ($F_0 < F_i, i = 1, ..., p$), $\boldsymbol{t}_p$ with function value F_p is the worst point ($F_p > F_i, i = 0, ..., p-1$) which is to be replaced in this iterate, and $\boldsymbol{t}_{p-1}$ is the next to worst point ($F_{p-1} < F_p$ and $F_{p-1} > F_i, i = 0, ..., p-2$). Next, the center of the points to remain in the simplex is obtained as:

$$t_m = (\sum_{i=0}^{p-1} t_i)/p$$

i.e. by averaging each coordinate over the set of points. A trial point is then generated by " reflecting" $\boldsymbol{t}_p$ towards the center:

$$\boldsymbol{t}_r = (1+\alpha)\ \boldsymbol{t}_m - \alpha\ \boldsymbol{t}_p$$

where the reflection coefficient α is a positive constant. The corresponding function value F_r is compared with F_0 to F_p to determine further steps in the iterate. Three possible outcomes are distinguished.

Firstly, $\boldsymbol{t}_r$ can be better than the worst point but not better than the best point ($F_0 < F_r < F_p$). In this case, $\boldsymbol{t}_r$ replaces $\boldsymbol{t}_p$ and a new iterate is started. Secondly, if $\boldsymbol{t}_r$ is the new best point ($F_r > F_0$), it is assumed that the direction of search has been a good one and search is "expanded" in this direction by examining a point

$$\boldsymbol{t}_e = \gamma\ \boldsymbol{t}_r + (1-\gamma)\ \boldsymbol{t}_m$$

with corresponding function value F_e. The expansion coefficient γ is positive constant. If F_e is less than F_r, the expansion has been successful and $\boldsymbol{t}_e$ replace $\boldsymbol{t}_p$. Otherwise $\boldsymbol{t}_e$ is discarded and $\boldsymbol{t}_r$ is substituted for $\boldsymbol{t}_p$ to complete the iterate. Thirdly, $\boldsymbol{t}_r$ may be worse than any of the remaining p points ($F_r > F_{p-1}$). This leads to "contraction" of the simplex, obtaining a new point $\boldsymbol{t}_c$ with pertaining function value F_c, as:

$$\boldsymbol{t}_c = \beta\ \boldsymbol{t}_r + (1\text{-}\beta\ )\ \boldsymbol{t}_m$$

if F_r is less than F_p, and as

$$\boldsymbol{t}_c = \beta\ \boldsymbol{t}_p + (1\text{-}\beta)\ \boldsymbol{t}_m$$

otherwise. The contraction coefficient β is a constant in the range from 0 to 1. If F_c is less than the smaller of F_r and F_p, the contraction has been successful and $\boldsymbol{t}_c$ replaces $\boldsymbol{t}_p$ to end the iterate. If not, the complete simplex is shrunk by moving each point halfway towards the best point, *i.e.*

$$t_i' = (t_0 + t_i) / 2$$

for $i = 1, ..., p$, before a new iterate is started. Suggested values for α, β and γ are 1.0, 0.5 and 2.0, respectively (Nelder & Mead 1965). Full computational details together with a flow diagram can be found in Nelder & Mead (1965), and a FORTRAN implementation for example in O'Neill (1971).

Since differences in function values are not utilized in establishing the direction of search, restrictions on the parameter space can be imposed easily by assigning a very large positive function value to parameter vectors out of bounds. As pointed out by Nelder & Mead (1965), this will be followed automatically by a contraction step which will eventually keep estimates within their bounds. The convergence criterion suggested by Nelder & Mead (1965) is the variance, or standard deviation, of function values in the simplex, rather than the more conventionally used change in estimates between iterates. The rationale for this was that in statistical problems such as maximum likelihood estimation, the curvature of the surface near the minimum reflects the amount of information available on the parameters. If the surface is flat, sampling errors are large and it is not worthwhile to determine the minimum with great accuracy.

For REML estimation of multiple variance components, the Simplex method proved robust and easy to use. For a given vector of starting values, $\boldsymbol{t}_0$, the initial simplex was made up of $\boldsymbol{t}_0$ and p vectors $\boldsymbol{t}_i$, obtained by multiplying the i-th element of $\boldsymbol{t}_0$ by 1.20. A factor this large, corresponding to a step size of 20% was chosen to ensure quick convergence even for bad starting values. This implied, however, that for starting values close to the estimates, the procedure would search unnecessarily in the wrong direction. A smaller step size may be sufficient and perhaps more efficient *i.e.* yield estimates with less likelihood evaluations. For cases where parameter vectors out of bounds were not a problem, the Quasi-Newton algorithm performed generally somewhat better than the Simplex method. The variance of function values in the Simplex had to be less than 10^{-8} or even 10^{-9} for the Simplex to find the minimum of $-2 \log \mathcal{L}$ with the same accuracy as ZXMIN (for an accuracy level of 5 or 6 significant digits). In terms of changes in parameter estimates, however, a limit of $10^{-}5$ to $10^{-}6$ generally appeared sufficient.

Figure 2 illustrates the convergence behaviour of the Simplex method for the same analyses as depicted in Figure 1. Oscillations in likelihood and estimates clearly reflect the "trial and error" mechanism of this approach. For both sets of starting values, 88 likelihood evaluations were required before the variance of function values in the Simplex dropped below the specified limit of 10^{-9}.

SAMPLING VARIANCES

The matrix of approximate, large sample covariances among variance component estimates is given by the inverse of the information matrix. This in turn can be approximated by the inverse of the Hessian matrix, *i.e.* the matrix of second derivatives of the log likelihood function with respect to the parameters to be estimated. A quadratic approximation of log $\mathcal{L}$ and Quasi-Newton procedures provide an approximate Hessian matrix ($\mathbf{Q}$) as a by-product. Using the Simplex method this has to be obtained explicitly. Nelder & Mead (1965) described an appropriate strategy, equivalent to the approximation by numerical differentiation

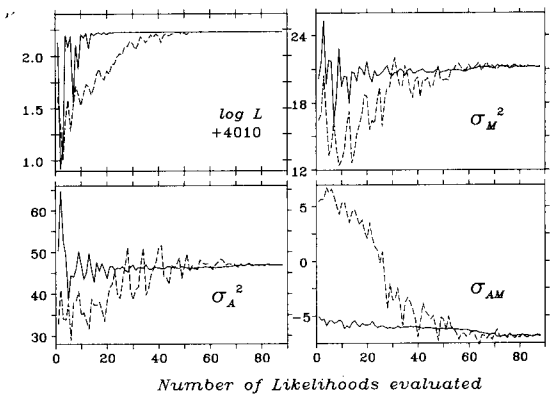

Fig. 2. Changes in likelihood and estimates of variance components for successive function evaluations, using the Simplex method for "good" (solid line) and "bad" (dashed line) starting values; see text for notation and details of the data set.

using forward differences (see *e.g.* Gill *et al.*, 1981). While the mechanics for approximating second derivatives are straightforward, it can be problematic in practice.

As noted when examining the quadratic approximation of log $\mathcal{L}$ for multiple parameters, **Q** was often not positive definite, whether derived by least-squares as described by Smith & Graser (1986) or using finite differences, and consequently yielded inappropriate predictions of the parameters maximizing log $\mathcal{L}$. Similarly, the approximate Hessian produced by Quasi-Newton routines did not necessarily give meaningful estimates of sampling variances. Indeed, a large proportion of the literature on Quasi-Newton algorithms is concerned with procedures to deal with bad or non positive definite approximations to the matrix of second derivatives.

Simulation was employed to examine the sampling distribution of variance component estimates and their predicted variances for models 1, 2 and 4. Data were sampled for one to four generations consisting of 25 to 800 full-sib families of size 2 to 10 each. Dams were nested within sires with each sire mated to 1 to 5 dams. Variance component estimates were obtained using the Simplex method with population parameters as starting values and a convergence criterion of $V(-2 \log \mathcal{L}) < 10^{-8}$. Second derivatives were approximated as described by Nelder & Mead (1965) using a step size of 0.1% of the estimates.

Approximation of sampling variances worked well for model 1, *i.e.* estimating the additive genetic and error variance only. It was satisfactory for model 2 which included a additional environmental component due to full-sib families (litters), but it failed for analyses under model 4, including a maternal genetic effect as an additional random effect for each animal.

Table II summarizes empirical and predicted sampling variances and empirical correlations between estimates for a variety of design for model 2. Results clearly emphasize the need to ensure that the data provide sufficient information to estimate all effects fitted. Considering only one generation, variance components were derived solely from the covariances between full-and half-sibs, causing a

high negative sampling correlation between σ^2_A and σ^2_C or, equivalently, between heritability and c^2 effect. In terms of the likelihood surface, this implied a maximum along a flat ridge, *i.e.* an area where for a constant sum of the two parameters the value of the likelihood changed very little with changes in the parameter values. If each sire was mated to only one dam, *i.e.* there were no half-sib families, there was little scope to partition the within family variance into its genetic and environmental components. This yielded a likelihood surface of a shape which did not allow sampling variances to be approximated.

Including data for a second generation provided the covariance between parents and their offspring as an additional source of information, thus allowing a considerably better discrimination between σ^2_A and σ^2_C, as evidenced by the decreased (absolute value) sampling correlation between the two components. While for one generation sampling variances decreased with increasing number of dams per sire the reverse generally held for data spanning several generations.

These relationships are also illustrated in Figure 3 for analyses under Model 4. Considering one generation only (top row), no animal had records expressing both its direct and maternal genetic effects. Hence the sampling correlation between σ^2_A and σ^2_M was large and negative. This was accompanied by little variation in estimates of σ_{AM} depicting that this was determined by the genetic covariance structure among animals only. For data including 3 generations (bottom row) sufficient comparisons between and within generations were available to yield estimates of σ^2_A and σ^2_M virtually independent of each other, while estimates of σ_{AM} were highly variable and showed a strong negative association with σ^2_M.

NUMERICAL EXAMPLE

Table III contains simulated data for 282 animals in two generations. Each generation consists of 18 full-sib families of size 6 to 10, where each sire is mated to 3 dams. Parents of generation one did not have records, which introduced 24 base animals, yielding 306 animals in total. Records were sampled according to Model 8 (see Table I) for population values of $\sigma^2_A = 40, \sigma^2_M = 15, \sigma_{AM} = -5, \sigma^2_C = 10$ and $\sigma^2_E = 50$, a phenotypic mean of 200 and a fixed effect of 20 for generation 2.

Data were analysed under Models 1, 2, 3, 4, 7 and 8, as described in Table I. Analyses were carried out using the Simplex method to find the maximum of the likelihood. Two sets of starting values were chosen, namely the population values as a good (Set I: θ_A=0.40, $\theta_M = 0.15$, $\theta_{AM} = -0.05$ and $\theta_C = 0.10$) and $\theta_A = 0.10$, $\theta_M = 0.30$, $\theta_{AM} = 0.10$, $\theta_C = 0.20$ (= Set II) as a bad initial guess. For models other than 8, the appropriate subset of parameters was used. Estimates for two values for the minimum variance of function values in the Simplex are summarized in Table IV. Characteristics of the augmented coefficient matrix **M** and intermediate results for the first likelihood evaluated in each run are given to help with the validation of computer programs. Gaussian elimination steps are not dependent on the model of analysis, hence the example given by Graser *et al.* (1987) should suffice as an illustration. Numerical examples to be used in checking the building of the mixed model array for the various models can be found elsewhere in the literature. In particular, Henderson (1988) gives examples for a variety of animal models.

Table II. Example data.

Family No.	*Sire ID*[A]	*Dam ID*	*Progeny records*
Generation 1			
1	1	2	220, 212, 221, 207, 218, 201, 214, 229
2	1	3	214, 198, 194, 211, 212, 228, 210, 198
3	1	4	223, 223, 215, 226, 212, 231, 229
4	5	6	221, 210, 213, 223, 239, 222, 223
5	5	7	217, 216, 206, 204, 216, 212
6	5	8	218, 218, 220, 202, 209, 213, 201, 211, 225
7	9	10	213, 211, 221, 215, 228, 204
8	9	11	225, 224, 225, 216, 224, 220, 225
9	9	12	215, 222, 219, 224, 225, 212, 214, 219
10	13	14	239, 233, 241, 232, 226, 228, 211, 215
11	13	15	228, 211, 202, 226, 222, 203, 213, 219
12	13	16	230, 238, 217, 220, 222, 227, 227, 234
13	17	18	222, 221, 203, 210, 210, 229, 227, 221
14	17	19	237, 222, 232, 237, 223, 229, 231, 229, 228, 232
15	17	20	213, 232, 227, 231, 232, 230, 220
16	21	22	234, 223, 226, 221, 230, 214, 233, 220
17	21	23	223, 216, 224, 227, 226, 217, 218
18	21	24	224, 228, 223, 225, 230, 213, 224, 227
Generation 2			
19	107	81	245, 245, 253, 252, 249, 240, 238, 254, 262, 258
20	107	82	232, 252, 243, 245, 236, 254, 233, 228
21	107	112	240, 238, 235, 230, 252. 238, 228, 246
22	118	120	253, 242, 252, 250, 231, 239, 246, 233
23	118	113	235, 254, 233, 233, 262, 250, 248
24	118	146	243, 255, 237, 230, 240, 236, 227, 247, 238
25	140	128	241, 234, 243, 229, 239, 243, 243
26	140	131	269, 240, 233, 246, 252, 240, 256
27	140	105	238, 232, 244, 221, 243, 251, 236, 241
28	100	98	254, 249, 237, 238, 248, 247, 233
29	100	37	232, 228, 225, 229, 264, 228, 224
30	100	52	234, 243, 238, 238, 237, 235, 243, 256, 241
31	123	137	237, 238, 234, 241, 221, 237, 236, 250, 237
32	123	69	247, 239, 238, 239, 233, 234, 224, 248, 235
33	123	39	248, 232, 240, 236, 235, 253, 243
34	51	65	234, 233, 222, 233, 213, 228
35	51	31	209, 225, 223, 226, 224, 223, 221, 224, 219
36	51	138	229, 238, 230, 240, 246, 236, 230, 230, 230

[A] Animal number: Progeny are numbered consecutively as listed, *i.e.* records 220, 212, ..., 229 in family 1 pertain to animals 25, 26, ..., 32, and family 36 consists of animals 298 to 306.

Table III. Estimates of variance components and pertaining likelihoods for data in Table II, for 6 different models of analysis and 2 convergence criteria, using the Simplex procedure to determine the maximum of the likelihood; together with computational characteristics of the first likelihood evaluated in each run, for two sets of starting values.

	Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 7*	*Model 8*
Components of the likelihood for starting values I						
$\mathbf{y}''\mathbf{P}\mathbf{y}$	15,471.911	13,415.534	12,869.662	13,518.880	11,263.396	11,946.071
$\log \lvert\mathbf{C}^*\rvert$	435.531	468.504	922.420	986.058	851.871	933.370
σ^2_A	36.838	38.330	40.856	38.625	45.973	42.665
σ^2_M	–	–	15.321	14.485	17.239	15.999
σ_{AM}	–	–	–	−4.828	–	−5.333
σ^2_C	–	9.583	–	–	11.493	10.666
σ^2_E	55.257	47.913	45.963	48.282	40.226	42.665
$\log \lvert\mathbf{G}\rvert$	1103.597	1197.106	1970.420	1923.036	2130.539	2069.119
$\log \lvert\mathbf{R}\rvert^A$	1131.382	1091.165	1079.440	1093.329	1041.856	1058.450
	−104.312	−239.901	−1270.843	−1287.182	−1359.585	−1381.239
$\log \mathcal{L}^B$	−857.40813	−852.85432	−950.49887	−950.95584	−951.42259	−950.62486
Components of the likelihood for starting values II						
$\mathbf{y}'\mathbf{P}\mathbf{y}$	22,128.650	17,688.306	16,917.472	16,181.931	15,220.205	14,428.670
$\log \lvert\mathbf{C}^*\rvert$	900.164	907.906	1234.475	1261.095	1048.232	1007.667
σ^2_A	8.781	9.025	10.070	11.559	13.589	17.177
σ^2_M	–	–	30.210	34.676	40.768	51.531
σ_{AM}	–	–	–	11.559	–	7.177
σ^2_C	–	18.049	–	–	27.179	34.354
σ^2_E	79.031	63.173	60.420	57.793	54.358	51.531
$\log \lvert\mathbf{G}\rvert$	664.820	777.339	1749.622	1709.926	2051.952	2079.738
$\log \lvert\mathbf{R}\rvert$	1232.295	1169.135	1156.570	1144.035	1126.756	1111.696
	−113.616	−257.044	−1361.636	−1346.878	−1470.377	−1450.723
$\log \mathcal{L}$	−866.68398	854.10060	−951.23030	−952.07137	−954.90335	−958.31668
Convergence criterion: $\mathrm{Var}(-2 \log \mathcal{L}) < 10^{-6}$						
σ^{2C}_A	44.175	30.384	24.344	38.195	26.501	30.405
σ^2_M	–	–	19.741	33.402	7.919	13.879
σ_{AM}	–	–	–	−20.722	–	−6.449
σ^2_C	–	14.920	–	–	9.449	7.968
σ^2_E	50.825	50.672	54.027	46.939	52.473	50.513
$\log \mathcal{L}$	−857.23731	−852.50969	−950.55494	−950.32041	−950.08570	−950.06236
Number of Simplex iterates[D]						
Run I	4	10	11	30	29	22
Run II	8	15	14	33	39	82
Number of likelihoods evaluated[E]						
Run I	10	22	22	58	55	47
Run II	18	31	29	59	72	144

Table III continued.

Model 1	*Model 2*	*Model 3*	*Model 4*	*Model 7*	*Model 8*	
Convergence criterion: Var($-2 \log \mathcal{L}$) $< 10^{-9}$						
σ^{2C}_{A}	43.954	30.981	24.822	37.872	26.577	30.518
σ^{2}_{M}	–	–	19.817	32.751	7.689	13.420
σ_{AM}	–	–	–	19.874	–	−6.405
σ^{2}_{C}	–	14.878	–	–	9.667	8.368
σ^{2}_{E}	50.954	50.343	53.738	47.091	52.421	50.433
log $\mathcal{L}^{F}$	−857.23720	−852.50918	−950.55422	−950.31979	−950.08554	−950.06174
	−759.50344	−754.77543	−755.08671	−754.85228	−754.61803	−754.59424
Number of Simplex iterates						
Run I	7	16	17	37	37	36
Run II	11	20	20	41	45	94
Number of likelihoods evaluated						
Run I	16	33	34	72	71	73
Run II	24	41	40	73	83	166
Time/log $\mathcal{L}^{G}$	0.122	0.265	0.204	0.247	0.464	0.508

[A]Second line: adjusted for factoring of σ^2_E from mixed model array
[B]Omitting constants not depending on parameters to be estimated
[C]Estimates given are for Run I
[D]No. of Simplex iterates carried out to reach specified minimum variance of function values
[E]No. of likelihoods evaluated in these iterates
[F]First line omitting constants; second line including log $|\mathbf{A}|$
[G]CPU time (in seconds) required per likelihood evaluation

To demonstrate calculation of the likelihood, consider the analysis for model 8 with starting values equal to the population parameters (Run I). Absorbing all fixed and random effects into $\boldsymbol{y}'\boldsymbol{y}$ (= 51, 165.234) gives the residual sum of squares $\boldsymbol{y}'\mathbf{P}\boldsymbol{y} = 11,946.071$ and log $|\mathbf{C}^*| = 933.370$ (taking natural logarithms throughout). Fitting generations as the only fixed effects, their design matrix $\mathbf{X}$ has full column rank 2. Hence the estimate of the residual variance is 11, 946.071/280 = 42.665. For $\theta_A = 0.40$, $\theta_M = 0.15$, $\theta_{AM} = -0.05$ and $\theta_c = 0.10$ this gives estimates of 42.665, 15.999, −5.333 and 10.666, respectively for the corresponding (co)variances. Log $|\mathbf{R}|$ is 282 × log 42.665 = 1058.450, but the term required in calculating log $\mathcal{L}$ is (282−2−2×306 −36) × log 42.665 = −368 × 3.753 = −1381.239. Log $|\mathbf{G}|$ is made up of three parts, namely NC × log σ^2_C = 36 × 2.367 = 85.215, NA × log σ^2_A = 306 × 3.753 = 1148.531, and NA × log $(\sigma^2_M - \sigma^2_{AM}/\sigma^2_A)$ = 306× log 15.333 = 835.374, which gives a total of 2069.119. The first log likelihood evaluated for this run is then −0.5 × (11,946.071/42.665 + 933.370 − 1381.239 + 2069.119 = −950.625.

For model 1, data were also analyzed using an EM-algorithm with tridiagonalisation of the coefficient matrix as described by Smith & Graser (1986). The

Table IV. Empirical (E) and predicted (P) sampling variances, with empirical standard deviation (sd) of predicted values, and sampling correlations of variance component estimates, based on 100 or more replicates. Data were simulated under model 2 with population parameters of $\sigma_A^2 = 40$, $\sigma_C^2 = 15$ and $\sigma_E^2 = 45$.

				Var (σ_A^2)			*Var* (σ_C^2)			*Var* (σ_E^2)			*Sampling correlation*			
FS^A	D^B	n^C	N^D	*E*	*P*	*sd*	*E*	*P*	*sd*	*E*	*P*	*sd*	*A,C*	*A,E*	*C,E*	h^2,c^2
One generation																
100	1	3-5	400	170	-	-	5	-	-	97	-	-	0.52	−0.88	−0.31	0.14
		6-10	800	135	-	-	2	-	-	50	-	-	0.40	−0.87	−0.15	−0.48
	2	6-10	800	411	584	315	100	117	62	117	159	78	−0.84	−0.94	0.82	−0.93
	5	3-5	400	599	629	407	87	100	39	201	186	100	−0.75	−0.92	0.61	−0.84
		6-10	800	435	513	350	57	70	32	116	141	87	−0.70	−0.94	0.63	−0.85
200	2	3-5	800	350	410	148	84	93	36	95	118	36	−0.81	−0.94	0.71	−0.90
	5	3-5	800	316	304	175	53	50	22	100	91	42	−0.74	−0.91	0.60	−0.84
400	5	2	800	224	209	65	45	47	9	79	74	16	−0.72	−0.86	0.44	−0.79
		3-5	1600	133	142	42	25	25	4	40	43	10	−0.75	−0.90	0.62	−0.86
800	2	2	1600	155	188	28	42	47	8	50	58	7	−0.78	−0.88	0.56	−0.85
	5	2	1600	87	101	19	24	24	2	34	36	5	−0.66	−0.85	0.38	−0.76
		3-5	3200	78	72	17	11	12	1	22	22	4	−0.76	−0.93	0.64	−0.86
Two generations																
25	1	3-5	200	441	437	208	99	93	62	163	154	56	−0.42	−0.79	0.21	−0.63
	5	3-5	200	477	526	335	77	76	42	183	173	80	−0.23	−0.86	0.13	−0.53
50	1	3-5	400	190	216	75	51	45	18	70	76	19	−0.28	−0.80	0.06	−0.55
	5	3-5	400	244	244	128	34	39	12	99	84	32	−0.33	−0.83	0.18	−0.55
100	5	3-5	800	180	199	81	21	23	6	75	62	20	−0.37	−0.91	0.30	−0.60
		1-7	800	193	202	79	25	24	7	62	62	19	−0.40	−0.90	0.26	−0.65
Three generations																
25	1	3-5	300	281	272	111	31	49	15	31	49	15	−0.08	−0.80	−0.11	−0.41
	5	3-5	300	320	364	156	55	50	20	121	115	39	−0.25	−0.79	0.04	−0.52
33	1	3-5	400	235	191	67	38	38	14	75	68	16	−0.15	−0.83	0.06	−0.50
	5	3-5	400	229	259	134	37	36	13	75	85	35	−0.29	−0.84	0.19	−0.58
Four generations																
25	1	3-5	400	217	197	74	28	35	12	68	66	17	−0.16	−0.83	0.01	−0.46
	5	3-5	400	216	247	103	34	33	11	83	79	23	−0.20	−0.78	−0.03	−0.49

[A] No. of full-sib families per generation

[B] No. of dams mated to each sire

[C] No. of progeny per full-sib family

[D] Average no. of records per data set

[E] A, C: σ_A^2, σ_C^2, A,E: σ_A^2, σ_E^2, C,E: σ_C^2, σ_E^2 and h^2, c^2: heritability, c-squared effect

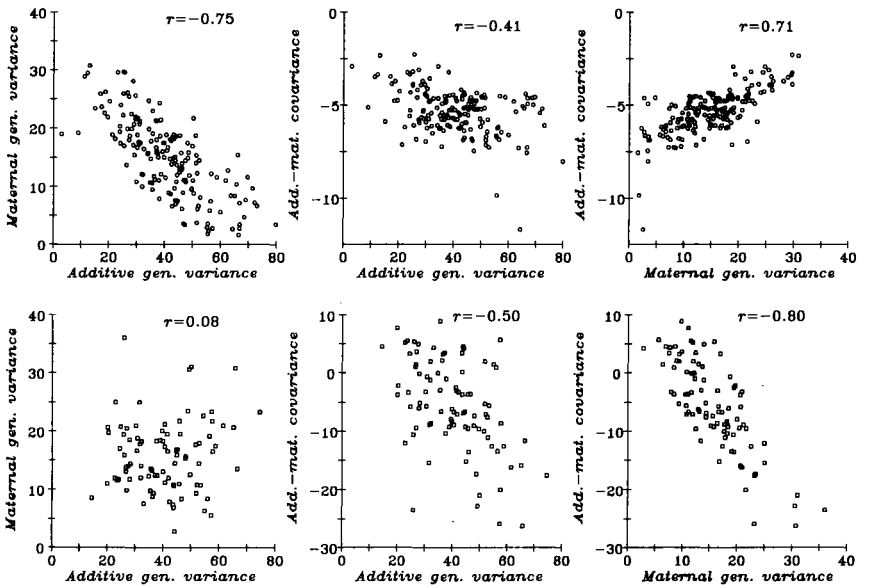

Fig. 3. Distribution of variance component estimates and corresponding sampling correlations (r) for data simulated under model 4 with population parameters of σ^2_A=40, σ^2_M= 15, σ_{AM}= −5 and σ^2_E=50 for a) one generation of 400 full-sib families of size 2 with each sire mated to 2 dams (circles (top row), 171 replicates), and b) 3 generations of 100 full-sib families of size 6 to 10 with each sire mated to 4 dams (squares (bottom row), 90 replicates).

derivative-free method required 0.6 seconds CPU time for set-up operations and 2.0 seconds to obtain estimates, using the Simplex procedure with population parameters as starting values (Run I) and performing 16 likelihood evaluations. The quadratic approximation for this case was somewhat faster, needing 1.5 seconds and 12 likelihood evaluations. The EM-algorithm required 36.0 seconds for various set-up steps, 31.0 seconds alone to tridiagonalize the coefficient matrix of order 306. Estimation was then fast, needing 0.7 s only to carry out 120 iterates. Part of the differences in computational requirements could be attributed to differences in computing techniques: programs for the EM-algorithm operated on the full coefficient matrix (though halfstored) rather than the derivative-free method which dealt with its non-zero elements only. Smith & Graser's (1986) procedure could also have been utilised to perform an analysis under model 2. This would have required at least 5 tridiagonalizations of the coefficient matrix, however, *i.e.* more than 180 seconds CPU time, in contrast to 9.5 seconds used by the derivative-free algorithm (Run I).

CONCLUSIONS

Direct maximization of the likelihood provides an attractive alternative to REML algorithms relying on information from derivatives. Though it has been discussed

here only with reference to Animal Models, it is extremely flexible and can be adapted to a wide range of models of interest in the analysis of animal breeding data. Graser *et al.* (1987) describe the application to a Reduced Animal Model which reduces the number of Gaussian Elimination steps required by not setting up equations for animals which do not have progeny but absorbing these directly into equations for their parents. Though not considered here, this equivalent model can be used to reduce computations for a number of animal models containing additional random effects though it is generally not feasible for non-additive genetic models; see Henderson (1988) for a discussion and examples. For AMs, a set of computer programs has been written accommodates all 10 models of Table I (Meyer 1988). Using sparse matrix techniques, models involving several thousand random effects levels can be handled computationally. Limitations on models and size and structure of data sets are imposed by the fact that each analysis requires numerous likelihood evaluations.

For univariate analyses, a reduction in the dimension of search is possible by estimating the variance of residual errors directly. The Simplex method is recommended as a robust and easy-to-use optimization procedure when the likelihood is to be maximized with respect to several parameters. For one parameter, the quadratic approximation used by Graser *et al.* (1987) appeared best. Extensions to multivariate analyses are straightforward though computationally considerably more demanding and will be considered in a subsequent paper.

Simulation showed the Simplex method to perform well. Means of estimates over replicates agreed with the population values for which data were sampled, indicating that the global rather than a local maximum of the likelihood function had been located. Limited work investigated the effect of different starting values on parameter estimates: while the number of likelihood evaluations required to obtain estimates varied with the initial guess, estimates for a particular data set did not depend on it. This suggests that local maxima are not a problem in derivative-free REML estimation. Höschele (1988) recently reported that REML is less likely to converge to local maxima than Bayesian procedures of variance component estimation. In general, however, this remains an as yet unsolved problem. Restarting the search at the predicted maximum or repeating it with different starting values will provide a reasonable degree of confidence in the assumption that the global maximum has been found, but we cannot be sure. Especially for complicated designs, the number of maxima of the likelihood surface over the parameter space is not known.

For models fitting animals' additive genetic merit as the only genetic effect, approximation of second derivatives of the likelihood function provided appropriate estimates of sampling covariances between estimates. For models containing other genetic effects, both non-additive genetic and maternal genetic, large negative sampling correlations between estimates were observed for a number of designs. This resulted in a shape of the likelihood surface which in general did not allow second derivatives to be approximated by numerical differentiation.

ACKNOWLEDGEMENTS

This study was supported by the Agricultural and Food Research Council. I am grateful to W.G. Hill for helpful comments and to B. Tier for explaining "linked list absorption".

REFERENCES

Dempster A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data *via* the EM algorithm. *J. R. Soc. Stat.* B. 39 1-38

Duff I.S., Erisman A.M. and Reid J.K. (1986) *Direct Methods for Sparse Matrices.* Monographs on Numerical Analysis, Clarendon Press, Oxford

Gill P.E., Murray, W. and Wright M.H. (1981) *Practical Optimization.* Academic Press, New York

Graser H.-U., Smith S.P. and Tier B. (1987) A derivative-free approach for estimating variance components in animal models by Restricted Maximum Likelihood. *J. Anim. Sci.* 64, 1362-1370

Harville D.A. (1977) Maximum Likelihood approaches to variance component estimation and to related problems. *J. Amer. Stat. Ass.* 72, 320-338

Harville D.A. & Callanan T.P. (1988) Computational aspects of likelihood-based inference for variance components. *In: Proc. Int. Symposium on Advances in Statistical Methods for Genetic Improvement of Livestock* (D. Gianola and K. Hammond eds) Springer Verlag, Heidelberg, FRG (in press)

Henderson C.R. (1973) Sire evaluation and genetic trends. *In: Proc. of Anim. Breed. Genet. Symp. in Honor of Dr. J.L. Lush*, ASAS/ADSA, Champaign, Illinois, 10-41

Henderson C.R. (1988) Theoretical basis and computational methods for a number of different animal models. *J. Dairy Sci.* 71 Suppl. 2, 1-16

Höschele I. (1988) Local maxima in likelihood and Bayesian estimation of variance components. *J. Dairy Sci.* 71 Suppl. 1, 262 (Abstr.)

IMSL library reference manual (1982). Version 9.2. Houston, Texas

Knuth D.E. (1973) *The Art of Computer Programming Vol. 1. Fundamental Algorithms*, 2nd edition. Addison- Wesley Pub. Co., Reading, Massachusetts

Meyer K. (1988) DFREML a set of programs to estimate variance components under an Individual Animal Model. *J. Dairy Sci.* 71 Suppl. 2, 33-34 (Abstr.)

NAG reference manual. Mark 11. Numerical Algorithms Group

Nelder J.A. & Mead R. (1965) A simplex method for function minimization. *Comput. J.* 7, 147-151

O'Neill R. (1971) Algorithm AS47: Function minimization using a Simplex procedure. *Appl. Stat.* 20, 338-345

Patterson H.D. & Thompson R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika* 58, 545-554

Quaas R.L. (1976) Computing the diagonal elements of a large numerator relationship matrix. *Biometrics* 32, 949-953

Searle S.R. (1979) *Notes on variance component estimation: A detailed account of Maximum Likelihood and kindred methodology.* Paper BU-673M, Biometrics Unit, cornell University, Ithaca, NY

Smith S.P. & Graser H.-U (1986) Estimating variance components in a class of mixed models by Restricted Maximum Likelihood. *J. Dairy Sci.* 69, 1156-1165

Spendley W., Hext G.R. & Himsworth F.R. (1962) Sequential application of simplex designs in optimization and evolutionary operation. *Technometrics* 4, 441-461

Swann W.H. (1972) Direct search methods. *In: Numerical Methods for Unconstrained Optimization* (W. Murray ed.) Academic Press, Inc., London and New York, p. 13-28

Tewarson R.P. (1973) *Sparse Matrices.* Academic Press, Inc., London and New York