Pedigree Analysis for Quantitative Traits:
I. Variance components without matrix inversion.

E.A.Thompson (1) and R.G. Shaw (2)

(1) Department of Statistics GN22, University of Washington, Seattle, WA 98195
(2) Dept. Botany and Plant Sciences, University of California, Riverside, CA 92521

SUMMARY

Recent developments in the animal breeding literature facilitate estimation of the variance components in quantitative genetic models. However, computation remains intensive, and many of the procedures are restricted to specialised designs and models, unsuited to data arising from studies of natural populations. We develop algorithms that allow maximum likelihood estimation of variance components for data on arbitrary pedigree structures. The proposed methods can be implemented on small computers, since no intensive matrix computations or manipulations are involved. Although parts of our procedures have been previously presented, we unify these into an overall scheme whose intuitive (and theoretical) justification clarifies the approach. An example of data on a natural population of *Salvia lyrata* is analysed.

KEYWORDS: Complex pedigrees, EM-algorithm, Mixed model, Variance components.

ACKNOWLEDGEMENTS

Both authors are greatly indebted to Joe Felsenstein for his encouragement and for many helpful discussions during the course of this work. The pedigree diagrams (figures 1 and 2) are produced with the software of Geyer (1988). The work of one author (R.G.S.) was supported in part by NIH grant F32 GM09664-02, and by a travel grant from the Academic Senate of the University of California at Riverside. Support for production of this report was from USDA grant 88-37151-3958.
We consider the mixed linear model:
\[ y = \mu + Tz + e, \]  
(1)
where \( \mu \) is the vector of means, \( T \) is a design matrix, \( z \) is a vector of random effects, and \( e \) is the vector of environmental effects. The estimation of the variance of random effects \( (z) \) is a long-standing and difficult problem. Methods for analysing balanced data, based on analysis of variance, are relatively straightforward, yet even in this case there is uncertainty and controversy (Searle, 1971: Pp. 401-403; 406-408). In the complementary case of unbalanced data, numerous methods, many of them derived from the analysis of variance of balanced data, are available. Searle (1971, pp. 421-515) discusses complexities and deficiencies of these at length. Maximum likelihood is a theoretically sound approach that provides unambiguous resolutions to many of the difficulties encountered in other methods. Shaw (1987) has recently discussed maximum likelihood methods for the genetic analysis of quantitative traits in natural populations. However, the approach has until recently been avoided in this area, because it does not generally provide analytical solutions, and it therefore requires computationally intensive iterative methods, routinely involving large matrices.

Beginning with the work of Hartley and Rao (1967), recent theoretical developments have considerably reduced the computational burden of likelihood estimation of variance components. Harville (1977), and more recently Meyer (1987a), contrast two main classes of algorithms: those employing second derivatives of the likelihood function (Fisher's method of scoring; Anderson 1973) and those using only the likelihood equation (EM algorithms; Dempster, Laird and Rubin, 1977). The latter, although often requiring many iterations to converge, are nonetheless advantageous, because each round of iteration normally uses less computation and memory than is required by a scoring algorithm, because the likelihood necessarily increases at each iteration, and because the EM-estimates of variance components are necessarily non-negative. Thompson and Meyer (1986) have suggested a reparameterisation of the EM algorithm that speeds convergence but destroys the non-negativity property of the iterates. On another front, Hemmerle and colleagues (Hemmerle and Hartley, 1973; Hemmerle and Lorenz, 1976; Goodnight and Hemmerle, 1979) have made successive improvements to methods for obtaining the required inverse of the matrix
\[ V = var(y) = \sigma^2TG'T + \tau^2I, \]  
(2)
with the notation, \( var(e) = \tau^2I \) and \( var(z) = \sigma^2G \). Dempster et al. (1984) have provided a recent discussion of several aspects of computational techniques.

As a result of these efforts, maximum likelihood (and the related method, restricted maximum likelihood) has been applied to estimation in the mixed linear model in quantitative genetics increasingly often in recent years (for example, Rothschild and Henderson, 1979; Tong, Kennedy, and Moxley, 1979; Meyer, 1983). These analyses have been feasible, only because the computing programs explicitly take advantage of the relatively simple structure often found in animal breeding data. For example, Meyer
employing an EM-algorithm, considers alternative computing strategies for
hierarchical data (dams nested within sires) depending on whether there are few or many
levels of fixed effects. Gianola (1986) has employed EM to estimate variance
components when the residual variance differs among environments. There has recently
been a literature explosion of papers proposing alternative procedures for solving the
iterative equations with more efficient computational schemes; for example, Schaeffer
and Kennedy (1986), Graser, Smith and Tier (1987) and Henderson (1986a,b). All of
these papers assume a genetic model with only additive genetic effects and/or special
design structures. Smith and Graser (1986) have proposed a method avoiding matrix
inversion, but require other matrix procedures and a grid search. For problems involving
multiple fixed and random effects and non-standard experimental designs (pedigrees, in
the genetic case), the computational problems remain daunting.

Many of the elements of the approach presented below have been implemented in
the context of animal breeding genetics, but the overall scheme here is novel in its
application to extended pedigrees and to the models of natural population genetics. In
contrast to Smith and Graser (1986) who state that "computational expedience and not
intuitive appeal is the justification for [their] algorithm", we feel it useful to present an
algorithm whose equations may be slow to converge but do have an intuitive
interpretation in the context of the genetic model. Additionally, the equations developed
below involve no explicit matrix inversion or determinant evaluation, either in the
estimation of variance components or in the evaluation of the likelihoods. The computer
storage required is minimal, and the equations are readily implemented.

2. Simple model for a quantitative trait; the EM equations.

We consider first a model which is considered by many other papers in the animal
breeding genetics literature. We shall assume that the trait is observed for \( k \) members of
a pedigree, giving rise to a \( k \)-variate Normal observation \( y \) with mean \( \mu I \) and variance-
covariance matrix, \( V = \sigma^2 G + \tau^2 I \), where now \( G \) is the matrix of genetic
relationships between the observed members of the pedigree, \( I \) is the \( k \times k \) identity matrix and \( I \) is a
\( k \)-vector of ones. In terms of underlying breeding values of the individuals this model may
also be expressed

\[
y_i = \mu + x_i, \quad x_i = z_i + e_i, \quad 1 \leq i \leq k
\]

where the \( e_i \) are independently and identically distributed \( N(0, \tau^2) \) environmental effects,
\( z_i \) is the additive genetic value of individual \( i \) which is \( N(0, \sigma^2) \) and conditionally on the
breeding values \( z_m \) and \( z_f \) of the parents of \( i \), \( z_i \) is \( N(\frac{1}{2}(z_m + z_f), \frac{1}{2}
\sigma^2) \). This representation lends itself to the EM algorithm (Dempster et al., 1977), with the
"complete data" consisting of the observed phenotypes \( y \) and the unobservable breeding
values \( z \). Although in (3) we have specified a constant mean \( \mu \) over all individuals, this is
for convenience only; differing means for individuals of different sex, or of different
generation, do not alter the method (see section 6). We note also that (3) is slightly
differently expressed from the standard form in the animal genetics literature (for
example, Schaeffer and Kennedy, 1986). Rather than the breeding values (additive
genetic effects) of the sires and dams used in those models, we consider the additive genetic effects of every individual, and the "design matrix" of other papers is thus incorporated into our matrix G. This form has also been used in the literature, first by Quaas and Pollak (1980), and more recently by Henderson (1986b) and Graser et al. (1987).

The basic form of the EM equations for the above model are well known (Hartley and Rao, 1967; Henderson 1986b). The natural sufficient statistics for the "complete data" (y,z) are $k^{-1} I' (y-z)$, $k^{-1} z' G^{-1} z$ and $k^{-1} e' e$ whose unconditioned expectations are $\mu$, $\sigma^2$ and $\tau^2$ respectively. Thus the iterative equations are formed by setting new values for these parameters equal to the conditional expectations of the statistics taken at current parameter values:

$$\mu^* = k^{-1} E_{\mu,\sigma^2,\tau^2} (I' (y-z) | y) = k^{-1} I' (y-a),$$  \hfill (4a)

$$\sigma^2 = k^{-1} E_{\mu,\sigma^2,\tau^2} (z' G^{-1} z | y) = k^{-1} (a' G^{-1} a + tr (G^{-1} W))$$ \hfill (4b)

and

$$\tau^2 = k^{-1} E_{\mu,\sigma^2,\tau^2} (e' e | y) = k^{-1} (h' h + tr (W))$$ \hfill (4c)

where

$$a = E_{\mu,\sigma^2,\tau^2} (z | y) = \tau^{-2} W x, \quad h = E_{\mu,\sigma^2,\tau^2} (e | y) = x - a, \quad x = y - \mu 1$$ \hfill (5)

and

$$W = var (z | y) = \sigma^2 \tau^2 G V^{-1} = \tau^2 (I - \tau^2 V^{-1})$$ \hfill (6)

These alternative equivalent representations of the conditional covariance matrix W result from the fact that we have only two random effects, whose conditional variances are thus equal. Both representations are used below. Note that whereas some authors (Henderson, 1986b; Gianola, 1986) include as components of $z$ the breeding values of all immediate members of the pedigree, whether or not the phenotypic value is recorded, our latent vector $z$ includes components only for the observed individuals. This is necessary for the identity (6). Although this reduction of the "missing" variables may decrease the simplicity of $G$, it will not improve the performance of the EM algorithm to introduce more latent variables than necessary, and where observations are sparse on an extended or complex pedigree, there may be many intermediate pedigree members which, if all were included, would make the dimensionality of $z$ very much greater than of $y$ or $e$.

The practical difficulty with the above EM equations, for those involved in the analysis of data in non-standard designs (on extended pedigree structures) and with small computers, is that all forms apparently require the computation of $V^{-1}$ at each iteration. Since this may be a large matrix, and since the EM algorithm may take a very large number of iterations to converge satisfactorily, the approach may be precluded for those without a Cray. The first requirement is thus the implementation of the above equations, and evaluation of the likelihood, without explicit inversion of $V$. 

4
3. Dissolving the matrices; useful formulae

We postpone to below the determination of \( a \), and show first that all other items of the above iteration are trivially computable, provided only the eigenvalues of \( G \) are known. This is a non-trivial requirement, but an infinitesimal one by comparison with repeated inversion of a varying matrix, being a single operation to be completed once for each pedigree structure; it can also be substantially simplified in many cases, as discussed below. Table 1(a) shows the steps of the EM iteration, and provides the equation references for implementation of each step.

We have, from the alternative representations of \( W \) (equation (6)),

\[
tr(WG^{-1}) = \sigma^2 \tau^2 tr(V^{-1}) \quad \text{and} \quad tr(W) = \tau^2 (k - \tau^2 tr(V^{-1}))
\]

(7)

Now, since every eigenvector of \( G \) is also an eigenvector of \( V = \sigma^2 G + \tau^2 I \), the eigenvalues of \( G \) provide the eigenvalues of \( V \) and hence also of \( V^{-1} \). Thus the two trace terms in the above equations are easily computed. This eigenvalue decomposition is analogous to that noted for the case of nested analysis of variance by Patterson and Thompson (1971), and used by Dempster et al. (1984) to simplify their EM computations. We note further that

\[
a' G^{-1} a = \begin{bmatrix} \sigma^4 \\ \tau^4 \end{bmatrix} h' G h.
\]

(8)

Thus the inverse of the matrix \( G \) is nowhere required. In fact, Henderson (1976) has given an explicit simple algorithm for \( G^{-1} \), which again is a computation to be performed once only. If this is implemented the above equation provides instead a useful check on the iterative procedure. Another useful identity to which we return below is

\[
W^{-1} = \left[ \begin{array}{cc} I & G^{-1} \\ \tau^2 & \sigma^2 \end{array} \right]^{-1}
\]

(9)

Consider finally the computation of the value of the likelihood. An EM algorithm which provides maximum likelihood estimates is much diminished if it does not provide also a computable form of the likelihood (cf. Felsenstein, 1987). Not only are likelihood values required at the final estimates, to provide tests of hypotheses, but due to the slow convergence of EM it is helpful if intermediate evaluations can assess progress of the iterative procedure. Note that direct evaluation is again barred, requiring the matrix \( V^{-1} \). For polygenic models in general, evaluation via the "peeling algorithm" of Elston and Stewart (1971) is a computationally feasible approach even on a large pedigree, but in the present case a general formula of Dempster et al. (1977) also provides for easy evaluation. The required formula is

\[
\log L = - \log b(\mu, \sigma^2, \tau^2) + \log b(\mu, \sigma^2, \tau^2 | y)
\]

(10a)

where \( b \) is the inverse of the normalizing factor for the density of the complete data

\[
b(\sigma^2, \tau^2) = \int \int \exp(-\frac{1}{2}[\tau^2(y-z-\mu I)'(y-z-\mu I) + \sigma^2 z' G^{-1} z]) \, dz \, dy
\]

and \( b(\cdot | y) \) is the same integral only over the unobserved \( z \). Given the simple Normal
Table 1: Implementation of the EM equations, in both "full" and "marginal" versions.

<table>
<thead>
<tr>
<th>(a) &quot;True&quot; EM (equ. (4)-(6)) Operation equations</th>
<th>(b) Partitioned EM (equ. (16)-(18)) Operation equations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read pedigree and data</td>
<td>Read pedigree and data</td>
</tr>
<tr>
<td>Compute G and eigenvalues</td>
<td>Compute G and eigenvalues</td>
</tr>
<tr>
<td>Set initial estimates of $\mu, \sigma^2, \tau^2$</td>
<td>Set initial estimates of $\mu, \sigma^2, \tau^2$</td>
</tr>
<tr>
<td>START ITERATION</td>
<td>START ITERATION; part 1</td>
</tr>
<tr>
<td>$x = y - \mu 1$</td>
<td>$x = y - \mu 1$</td>
</tr>
<tr>
<td>(5)</td>
<td>(5)</td>
</tr>
<tr>
<td>Compute $a$, $h = x - a$</td>
<td>Compute $a$, $h = x - a$</td>
</tr>
<tr>
<td>evaluate likelihood (^)</td>
<td>evaluate likelihood (^)</td>
</tr>
<tr>
<td>(10)</td>
<td>(10)</td>
</tr>
<tr>
<td>Compute $\mu^*$</td>
<td>evaluate $aG^{-1}a$</td>
</tr>
<tr>
<td>evaluate $tr(V^{-1})$</td>
<td>evaluate $tr(V^{-1})$</td>
</tr>
<tr>
<td>hence $tr(W)$, $tr(WG^{-1})$</td>
<td>hence $tr(W)$, $tr(WG^{-1})$</td>
</tr>
<tr>
<td>(7)</td>
<td>(7)</td>
</tr>
<tr>
<td>Compute $\sigma^{*2}, \tau^{*2}$</td>
<td>Compute $\sigma^{*2}, \tau^{*2}$</td>
</tr>
<tr>
<td>(4b),(4c)</td>
<td>(16a),(16b)</td>
</tr>
<tr>
<td>REPEAT to convergence</td>
<td>REPEAT to convergence</td>
</tr>
<tr>
<td>Print current estimates (^)</td>
<td>Print current estimates (^)</td>
</tr>
<tr>
<td>REPEAT to convergence</td>
<td>REPEAT to convergence</td>
</tr>
<tr>
<td>Compute $a$, $h = x - a$</td>
<td>Compute $a$, $h = x - a$</td>
</tr>
<tr>
<td>evaluate likelihood</td>
<td>evaluate likelihood</td>
</tr>
<tr>
<td>(5),(15)</td>
<td>(5),(15)</td>
</tr>
<tr>
<td>(10)</td>
<td>(10)</td>
</tr>
</tbody>
</table>

(\^): optional

forms of our distributions, and the independence of $z$ and $e = y - z - \mu 1 = x - z$, the above expression is readily found to be (up to an additive constant $(k/2) log(2\pi)$)

$$-rac{1}{2}k \log \sigma^2 + \log |G| + \frac{1}{2} \left[ - \log |W^{-1}| - \frac{(x'x-a'a)}{\tau^2} + \frac{a'G^{-1}a}{\sigma^2} \right], \tag{10b}$$

or, using (8),

$$-rac{1}{2} \log |V| + \frac{(x'x-a'a)}{\tau^2} - \frac{\sigma^2 h'Gh}{\tau^4}, \tag{10c}$$

which again requires only the eigenvalues of $G$, and hence of $V$.

4. The eigenvalues of $G$

A pedigree is defined by a list of triples, giving for every individual, its identifier, its mother's identifier, and its father's identifier. It is conventional in pedigree analysis to identify either both parents of an individual, or neither. This does not mean that all parents have observed data variables, merely that an identifier is assigned to a parent. Individuals with neither parent identified are known as founders, and are (by definition)
unrelated relative to the pedigree as specified. Individuals with both parents specified are known as *non-founders*.

The matrix $G$, sometimes called the "numerator relationship matrix", has $(i,j)$ component $G_{ij}$ equal to twice the kinship coefficient between individuals $i$ and $j$. Thus, if $i$ is not $j$ nor an ancestor of $j$, and if $m$ and $f$ are the parents of $i$,

$$G_{ij} = \frac{1}{2}(G_{mj} + G_{fj}), \quad (11a)$$

and

$$G_{ii} = (1 + G_{mf}) = (1 + F_i), \quad (11b)$$

where $F_i$ is the inbreeding coefficient of individual $i$. We assume here that $i$, $m$ and $f$ are distinct; no selfing allowed. If $i$ is a founder, and not $j$ nor an ancestor of $j$, then $G_{ij} = 0$.

Despite the structured form of $G$ and the explicit algorithm of Henderson (1976) for $G^{-1}$, the eigenvalues of $G$ appear less immediate. However, some simplifying comments can be made. First, for a non-inbred pedigree, the trace of $G = \sum G_{ii}$, which is of course the sum of the eigenvalues, is simply the number of individuals $k$. More generally, equation (11b) gives the trace of $G$ since $F_i$ is easily found by a recursive algorithm, even on large and complex pedigrees. Second, the trace of $G^{-1}$, which is the sum of the inverses of the eigenvalues of $G$ can also be readily determined, using the algorithm of Henderson (1976). For a non-inbred pedigree it is

$$n_0 + \frac{5}{3} n_1 + 3 n_2$$

where $n_j$ is the number of individuals with $j$ parents in the pedigree. Thirdly, $|G|$ is also simply determined:

(i) addition of a founder to the pedigree leaves $|G|$ unchanged, since $G_{ii} = 1$ and $G_{ij} = 0$ for $j \neq i$.

(ii) addition of a non-founder $i$ with both parents $m$ and $f$ observed multiplies $|G|$ by

$$(1 + F_i)^{-\frac{1}{2}}(G_{im} + G_{if}) = \frac{1}{2} - (F_f + F_m)/4 \quad (12a)$$

from equation (11) noting that $|G|$ is unchanged by taking linear combinations of columns, and $G_{ij} = \frac{1}{2}(G_{mj} + G_{fj}) = 0$ for $j \neq i$.

(iii) similarly, addition of non-founder $i$ whose parent $f$ is not an observed member of the pedigree and is a founder, multiplies $|G|$ by

$$1 - \frac{G_{im}}{2} = \frac{3}{4} - \frac{F_m}{4} \quad (12b)$$

In the simplest cases of a non-inbred pedigree,

$$|G| = \left(\frac{1}{2}\right)^{n_2} \left(\frac{3}{4}\right)^{n_1} \quad (12c)$$

where again $n_j$ is the number of individuals with $j$ parents present in the pedigree.
Although these results provide useful checks on the eigenvalues of G, the major simplification rests on the following observation. Where any set of l individuals (l>1) all share equal relationship to every other member of the observed pedigree, there are \( l-1 \) normalised orthogonal contrasts of their breeding values, each of the form

\[
\sum_{j=1}^{l} w_j z_j \text{ where } \sum_{j} w_j = 0, \sum_{j} w_j^2 = 1.
\]

These contrasts are independent of each other, of their average \( l^{-1} \sum_{j=1}^{l} z_j \), and of all other breeding values in the pedigree. Since G is the variance of the scaled breeding values, \( z/\sigma \), the variances of these normalised contrasts are eigenvalues of G, and for the determination of the remaining eigenvalues of G these l rows and columns of this matrix may be replaced by a single row and column. The off-diagonal terms of this row are the terms of each replaced row (all these rows being identical), and the diagonal term is the variance of the average of the replaced breeding values which is the sum of all elements in the eliminated diagonal block divided by \( l \). For example, every founder couple gives rise to an eigenvalue of

\[
\text{var}\left(\frac{z_1-z_2}{\sqrt{2}\sigma}\right) = 1.
\] (13a)

Every set of l full sibs without descendants (regardless of whether they have additional sibs with descendants), provides \( l-1 \) eigenvalues

\[
\sigma^{-2}\text{var}\left(\sum_{i=1}^{l} w_i z_i \right) = \sigma^{-2}\left(\text{var}(z_i) - \text{cov}(z_i,z_j)\right)
\]

\[
= (1 + F_i) - ((1+F_m) + (1 + F_f) + 2(2F_i))/4 = 1 - (F_m+F_f)/4 \quad \text{(13b)}
\]

where \( F_m \) and \( F_f \) are the inbreeding coefficients of the parents of the sibship, and \( F_i \) that of any offspring individual in it. Every set of l half-sibs, whose non-common parents are not observed and are unrelated to other members of the pedigree, give rise to \( l-1 \) eigenvalues

\[
\sigma^{-2}\text{var}\left(\sum_{i} w_i z_i \right) = (1 - \sigma^{-2}\text{cov}(z_i,z_j)) = \frac{3}{4} \quad \text{(13c)}
\]

Thus an extended pedigree may be very substantially reduced, many of the eigenvalues of G being simply determined from the pedigree structure, and those of the remaining core quite quickly computed. For example, for the 45-member test pedigree of figure 1, with 37 observed individuals, there are two founder couples, 16 sibs without descendants in 5 sibships, and (if individuals 10, 19, 20, 26, 27 and 28 are unobserved) 6 additional half sibs in 2 half-sibships. Thirteen of the eigenvalues of G are thus immediate. The remaining 20 are easily determined.
5. Peeling the breeding values

We have not yet determined the conditional expected breeding values \( \mathbf{a} \), other than as an unacceptable matrix formula. We present now a method for this, completing all items required for implementation of the steps of Table 1(a). For simplicity (although this is not essential) we shall augment the pedigree by adding members so that every individual is either a founder or has two parents present in the pedigree. Although we determine conditional expected breeding values for this larger set of individuals, note that only those for the observed individuals will enter into the EM iterative formulae (see note following these formulae; equations (4)-(6)).

Consider now the conditional breeding value of a single individual in the pedigree given those of all of his mates \( (s_j) \), his offspring \( (c_{jl}) \), and his parents \( (m \text{ and } f) \) (if present in the pedigree), and conditional on his phenotypic value \( y_i \) (if observed) (Figure 2). (For ease of notation we shall use \( s_j, c_{jl}, m \) and \( f \) to denote both the individual and the currently estimated breeding value thereof.) This conditional expectation is equal to the expectation given all observations and other breeding values on the pedigree; this fundamental conditional dependence is of course also the basis of "peeling" methods for computation of likelihoods (Elston and Stewart, 1971; Cannings, Thompson and Skolnick 1978). That is,

\[
E_{\mu, \sigma^2, \tau^2}(z_i | y, z^{(i)}) = E_{\mu, \sigma^2, \tau^2}(z_i | y_i, m, f, \{s_j\}, \{c_{jl}\})
\]

where \( z^{(i)} \) denotes all components of \( z \) other than \( z_i \). Now, given \( m \) and \( f \), \( z_i \) is \( N(\frac{1}{2}(m+f), \frac{1}{2} \sigma^2) \), or if \( i \) is a founder \( z_i \) is \( N(0, \sigma^2) \). Given \( z_i \), \( (y_i - \mu) \) is \( N(z_i, \tau^2) \) and given \( z_i \) and mate value \( s_j \), the offspring values \( c_{jl} \) are independent \( N(\frac{1}{2}(z_i + s_j), \frac{1}{2} \sigma^2) \). Thus the above expectation is readily found to be

\[
\left( (J_p + 1)\sigma^2 + J_d \tau^2 + \frac{1}{2} J_s N_s \sigma^2 \right)^{-1} \left[ J_p \sigma^2 (m+f) + J_d \tau^2 (y_i - \mu) + J_s \sigma^2 \sum_{j} (c_{jl} - \frac{1}{2} s_j) \right]
\]

\[
\cdots
\]

(15)

where \( J_p = 0 \) or 1 as the individual is/is not a founder, where \( J_d = 1 \) or 0 as the individual is/is not observed, where \( J_s = 1 \) or 0 as the individual has/has not mates and offspring present, and \( N_s \) is the total number of offspring. Thus a founder/parent term in the denominator is required, but observation, spouse, and offspring terms are included only if the relevant observations/individuals exist. Now this equation may be iterated through the pedigree, adjusting one breeding value at a time, until convergence is achieved. In fact convergence will be rapidly achieved, and to the required solution \( \mathbf{a} = \tau^{-2} \mathbf{Wx} \) (equation (5)). This is seen by noting that from equation (9)

\[
\tau^{-2} \mathbf{y} = \mathbf{W}^{-1} \mathbf{a} = \left[ \frac{1}{\tau^2} + \frac{\mathbf{G}^{-1}}{\sigma^2} \right] \mathbf{a}
\]

and that the above iteration (15) is no more than the iterative Gauss-Seidel inversion of (an augmented version of) this positive definite matrix equation with dominant diagonal terms. This (and other) iterative inversion routines have been used by several authors to
determine expected breeding values (for example, Van Vleck and Dwyer, 1985). However, the "peeling" and conditional independence approach gives a more intuitive justification for the result. In implementing the above iteration it is normally most efficient to work up, rather than down, an extended pedigree, since it is the younger members of the structure whose phenotypes are normally observed, whereas, for the earlier generations, expected breeding values are influenced via data on their descendants.

6. Extension to more complex models; backfitting algorithm.

Before proceeding to more complex models we note first an alternative modification of the EM equations above. Instead of simultaneous updating of $\mu$, $\sigma^2$ and $\tau^2$, we can instead use simpler 2-parameter EM equations for two sub-models. First, with $\mu$ fixed we have

$$\sigma^{*2} = k^{-1}E_{\mu, \sigma^2, \tau^2}(x'G^{-1}z \mid y) \quad \text{(16a)}$$

and

$$\tau^{*2} = k^{-1}E_{\mu, \sigma^2, \tau^2}(e' \mid y) \quad \text{(16b)}$$

Then, using the procedure of (14) and (15), we update

$$a = E_{\mu, \sigma^2, \tau^2}(z \mid y). \quad \text{(17)}$$

Then, holding $\sigma^2$ at its current value $\sigma^{*2}$;

$$\mu^{**} = k^{-1}E_{\mu, \sigma^{*2}, \tau^{*2}}(1' (y - z) \mid y) = k^{-1}1'(y - a), \quad \text{(18a)}$$

and

$$\tau^{**2} = k^{-1}E_{\mu, \sigma^{*2}, \tau^{*2}}(e' \mid y), \quad \text{(18b)}$$

giving a new set of estimates $(\mu^{**}, \sigma^{*2}, \tau^{**2})$.

Table 1(b) shows the steps of this iteration, comparing the procedure with the "full" EM procedure (Table 1(a)). Like the full EM, each step of this two-stage iteration yields an increasing likelihood, provided the necessary updating of $a$ is done at every step and convergence to a maximum of the likelihood surface is usually obtained; the method is analogous to that implemented by Felsenstein (1987). In the current simple model, the partitioning of the EM iteration in this way is not efficient. Extra computation is involved in recomputation of $a$ (equation (17)). This recomputation is necessary in theory, although if parameter values are changing slowly it has little effect on the performance of the algorithm. There is also scarcely any difference in practice between the convergence properties of (16)-(18) and the "true" EM, equations (4)-(6), although again the latter seems to usually provide marginally faster convergence.

However, this partitioning of the iteration becomes computationally effective when we extend our matrix-free iterative procedure to more general models of the form:

$$y = \sum_{j=1}^{p} \mu_j r^{(j)} + z + e, \quad \text{(19)}$$
where \( r_{ij}^{(l)} = 0 \text{ or } 1, \ 1 \leq i \leq k, \ 1 \leq j \leq p \). That is, we now have numerous fixed effects \( \mu_j \) each being present/absent for each of the \( k \) individuals observed. For example, we could incorporate sex effects and/or a range of environmental effects (location, year, soil fertility, population density). In our approach we estimate all effects, as, for example, do Hemmerle and Lorens (1976). Many other authors "absorb" these effects (Dempster et al., 1984), and estimate only the variance components.

With all the fixed effects remaining constant, we may apply equations (16) for the updating of \( (\sigma^2, \tau^2) \); the subtraction of the fixed \( \mu \) from each observation in forming \( a \) is replaced by the subtraction of all fixed effects applicable to that observation. Then each \( \mu_j \) may be successively updated via the analogue of equation (18a), \( \sigma^2 \) and all \( \{\mu_l; l \neq j\} \) remaining fixed:

\[
\mu_j^{**} = r_j^{(l)}^{-1} (y-a-\sum_{i \neq j} \mu_i r_{ij}^{(l)})/r_j^{(l)} \cdot 1.
\]  

(20)

The parameter \( \tau^2 \) is then updated via equation (18b). Thus, as in equations (16)-(18) we may successively update all parameters. Because of the linearity of the natural sufficient statistics for the fixed effects, each part iteration must increase the likelihood, provided only the expected random effects are recomputed at every stage. Again we note that for this model of two random effects (\( z \) and \( e \)) and multiple fixed effects, this partitioning of the EM-iterations is analogous to that of Felsenstein (1987). An alternative would be to implement the analogues of equations (4)-(6) for \( (\mu_j, \sigma^2, \tau^2) \) keeping fixed \( \{\mu_l; l \neq j\} \) but if \( p \) is large this will entail more computation since \( \sigma^2 \) will be updated \( p \) times in each full cycle.

For multiple random effects the situation is more complex;

a) the quadratic nature of the natural sufficient statistics for the variance components means that the EM procedure cannot be partitioned as for the fixed effects above;

b) as commented by Dempster et al. (1984) the eigenvalue decomposition following equation (7) above is in general "applicable only to mixed models with exactly two variance components"; and

c) equations (10) for the likelihood evaluation no longer apply in simple form.

Despite this, there are cases where the methods of this paper extend with surprising ease, the most important being the inclusion of a dominance variance or common-sib random component (for many typical data structures these being confounded). We note that the variance matrix of the dominance effects \( (\sigma_D^2 \mathbf{D}) \) often has simple form:

\[
\mathbf{D}_{ii} = 1, \quad \mathbf{D}_{ij} = 1/4 \quad \text{for full sibs},
\]

(21)

and \( \mathbf{D}_{ij} = 0 \) for any pairs of individuals not bilaterally related.

This simple block form of \( \mathbf{D} \) ensures that \( \mathbf{G} \) and \( \mathbf{D} \) share many eigenvectors, in particular those arising from founder couples (13a) and from sets of sibs without offspring (13b). Again, in figure 1, there are only 20 observed individuals with full sibs, in 6 sibships, and the 11 sib contrasts which are eigenvectors of \( \mathbf{G} \) are also eigenvectors of \( \mathbf{D} \). For the 17 observed individuals without observed bilateral relatives, the dominance contribution is
an individual effect; all vectors are eigenvectors of \( I \). Thus, in large part, the eigenvalue decomposition referred to in (b) can be extended, and then the EM equations implemented.

The difficulty (c) with computation of the value of the log-likelihood is also avoided in the cases where \( D \) and \( G \) share eigenvectors, for then \(|V|\) is easily computed, and the method of equation (10c) may be extended. Alternatively, for any model providing an overall multivariate Normal trait distribution the peeling method for likelihood computation (Elston and Stewart, 1971) as extended by Ott(1979) and by Cannings, Thompson and Skolnick (1980) applies regardless of the number of variance parameters. Our program has implemented this latter approach for likelihood evaluation.

7. An example; unbalanced diallel cross.

The procedures of the preceding sections have been implemented, both in C on a SUN3/160 (E.A.T) and in Pascal on a MicroVAXII (R.G.S), and have been tested on hypothetical data on multigeneration pedigrees (for example, Figure 1). Here, they are illustrated by the following analysis of real data on seed weight in 476 individuals of \textit{Salvia lyrata} (Lamiaceae) taken from an incomplete diallel cross of fourteen parent plants (Figure 3). The seed weights of parents were not recorded. Thus, the data are measures on a single generation containing 25 full sibships with multiple half-sib relationships among them. (These data are not ideally suited to our method, which is designed to be applied to data on extended pedigrees; for this particular structure, the EM-algorithm of Thompson and Meyer(1986), which use only parental breeding values, could perhaps be as efficiently employed.) Although the data contain information regarding effects of genetic interaction and of maternal and paternal cytoplasm on offspring phenotype (Cockerham and Weir, 1977) and, hence, the components of variance of these effects, we here analyse the data using a simple model (3) involving only additive genetic and environmental effects. These data have also been subjected to likelihood analysis using a scoring algorithm (Shaw, in preparation), but analysis of the full data set by that method is impossible on a small computer; only results for a subpedigree consisting of two offspring from each of the 25 sibships have been obtained. In contrast, with the EM-procedure of this paper there is no difficulty in analysing the full data set, under the model of equation (3). Since eight of the 25 sibships arise as four reciprocal pairs, for the purposes of the model (3) there are 21 sibships, ranging in size from 4 to 50.

For comparison, we first analyse a subset of the data consisting of measures of 50 seeds, 2 from each of the crosses. First, the eigenvalues of \( G \), required in equations (7) and (10), are obtained. There are 50-21=29 eigenvalues of 0.5 (from (13b)), and the remaining 21 eigenvalues of the reduced matrix range from 1.0 (four times) to 3.874. The EM-equations of table 1(a) are then performed. Within each inner iteration for breeding values, initial values are taken as one half of the difference between observed data values (zero, for unobserved individuals) and the current \( \mu \) estimate, to avoid any possible degenerate points of the iteration. For this inner iteration, ten cycles was always more than sufficient for convergence. Given starting values of \( \mu = 0.0, \sigma^2 = 0.1 \) and
\[ \tau^2 = 0.9, \] the main iterative procedure converged in 28 cycles, the log-likelihood increasing by 10.63. The procedure yielded estimates for the mean \( \mu \) of 2.575 mg and for additive (\( \sigma^2 \)) and environmental (\( \tau^2 \)) variances of 0.160 and 0.204. Varying starting values had little effect on the final estimates or log likelihood; final estimates never differed from the values reported above by more than 3%. Choice of starting values closer to the estimates (i.e. within 15%) did reduce the number of iterations to fewer than 10.

For the full data there are 476-21=455 eigenvalues of value 0.5, and the remaining 21 computed from the reduced matrix are found to range in value from 2.438 to 25.553. From starting values of \( \mu = 0.0, \sigma^2 = \tau^2 = 0.2 \), estimates of mean seed weight \( \mu = 2.677 \text{ mg}, \sigma^2 = 0.071 \text{ and } \tau^2 = 0.13 \) were obtained in 41 iterations. The log likelihood increased by 36.14. Choice of starting values closer to these estimates led within 6 iterations to a slightly higher (by 2.07) log likelihood and the following estimates: \( \mu = 2.697, \sigma^2 = 0.053, \tau^2 = 0.145 \). In all these cases, the likelihood increased every iteration.

8. Discussion

We have proposed a method of analysing quantitative genetic data that employs the EM algorithm (Dempster, et al. 1977). While we have developed our algorithm to maximise the joint likelihood of fixed effects and variance components, it is straightforward to modify the EM equations to maximise the likelihood in terms of the variance components alone thereby obtaining REML estimates (Laird, 1982; Henderson, 1986a). These have often been preferred in the context of quantitative genetics, because the estimates of variance components are unbiased. We have developed the algorithm for estimation of components of variance of a single trait. In many instances, however, estimates of genetic and environmental components of covariance between traits are of equal or greater interest, because they may affect correlated responses to selection (Falconer, 1981); for empirical examples, see Schaeffer, Wilton and Thompson (1978), Tong, et al. (1979), and Garland (1987). Extension of our algorithm to the analysis of a set of \( q \) traits may be accomplished by considering \( q \times q \) matrices of quadratic forms of random and residual effects. Although the formulae become cumbersome, the methods remain computationally feasible, since matrices of order \( k \) (the number of observed individuals) are not involved.

The EM approach has been widely developed and applied in the animal breeding literature (see references cited in Introduction). The advantage of the method as presented in this paper is that it not only avoids cumbersome matrix computations and may therefore be implemented on small computers, but also is sufficiently general that it may be readily applied to data on extended, multigenerational, complex pedigrees. Both of these features make our algorithm particularly valuable in quantitative genetic studies of natural populations, in which standard breeding designs are often not practicable.
REFERENCES


**Figures.**

**Figure 1:** Example pedigree of 45 individuals, of whom 37 are observed, for a quantitative trait. This pedigree can be readily analysed by the methods of this paper; for details see text.

**Figure 2:** Nuclear family structure, showing the values required for the conditional independence of the breeding value of i of other breeding values in the pedigree. (For details see text, equations (14) and (15).)

**Figure 3:** Data structure for the *Salvia lyrata* seed weight data, analysed in section 7.
Figure 1: Example pedigree.
Figure 2: Nuclear family neighbourhood.
Figure 3: Number of progeny in diallel cross, of Salvia lyrata, used as an example in the text.

<table>
<thead>
<tr>
<th>Female parent</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td></td>
<td></td>
<td>25*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>13*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>25*</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td>18*</td>
<td>25*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>11*</td>
<td>17</td>
<td>16</td>
<td>16</td>
<td>24*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td>25</td>
<td>25</td>
<td></td>
<td>8*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td>19</td>
<td></td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
<td></td>
<td>21</td>
<td></td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td>25</td>
<td></td>
<td>25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* denotes one of the eight part-sibships of the four reciprocal crosses.
(Total progeny = 476)