Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model

R. R. Corbeil and S. R. Searle

Biometrics Unit
Cornell University
Ithaca, New York 14853

The maximum likelihood (ML) procedure of Hartley and Rao [2] is modified by adapting a transformation from Patterson and Thompson [7] which partitions the likelihood under normality into two parts, one being free of the fixed effects. Maximizing this part yields what are called restricted maximum likelihood (REML) estimators. As well as retaining the property of invariance under translation that ML estimators have, the REML estimators have the additional property of reducing to the analysis variance (ANOVA) estimators for many, if not all, cases of balanced data (equal subclass numbers).

A computing algorithm is developed, adapting a transformation from Hemmerle and Hartley [6], which reduces computing requirements to dealing with matrices having order equal to the dimension of the parameter space rather than that of the sample space. These same matrices also occur in the asymptotic sampling variances of the estimators.

KEY WORDS

Variance Components
Mixed Model
Restricted Maximum Likelihood
Maximum Likelihood
W-transformation

1. INTRODUCTION

The mixed model in the analysis of variance can be represented for a vector of observations $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{U}\boldsymbol{\beta} + \mathbf{e}$$

where $\boldsymbol{\mu}$ is a vector of all the fixed effects, $\boldsymbol{\beta}$ is a vector of all the random effects, $\mathbf{X}$ and $\mathbf{U}$ are the respective design matrices, and $\mathbf{e}$ is a vector of residual error terms having variance σ^2. This and the different variances of the elements of $\boldsymbol{\beta}$ are the variance components of the model. The maximum likelihood procedure of Hartley and Rao [2] yields simultaneous estimation of both the fixed effects and the variance components by maximizing the likelihood of $\mathbf{y}$ with respect to each element of $\boldsymbol{\mu}$ and with respect to each of the variance components.

Paper number BU-301 in the Biometrics Unit, Cornell University.

R. R. Corbeil's present address is University of San Diego, Alcala Park, San Diego, California, 92110.

Received March 1974; revised July 1975

In contrast, we develop estimators (and their large sample variances) which are free of the fixed effects in the sense that the likelihood does not contain $\boldsymbol{\mu}$; i.e., we maximize the likelihood over a restricted parameter set. This is a generalization of the procedure suggested by Thompson [9] who considered the problem only for balanced data and for the completely random model. The procedure developed here is applicable to unbalanced data generally (including, of course, balanced data which are just a special case), and it is also applicable to mixed models for any mix of fixed and random effects. This is achieved by adapting a transformation used by Patterson and Thompson [7] which leads to a partitioning of the likelihood function into two parts: one part is entirely free of the fixed effects, and maximization of this provides what is called restricted maximum likelihood (REML) estimators of the variance components. Adaptation of a transformation described by Hemmerle and Hartley [6] that simplifies computation of the Hartley-Rao estimators greatly aids the computing of the REML estimators and also simplifies derivation of their large-sample variances. Finally, maximizing that portion of the likelihood not used for the REML estimators provides estimation of the fixed effects based on the REML estimators.

The REML estimators are not only invariant to the fixed effects of the model but they are also free of the estimates of the fixed effects. Furthermore, in the several cases of balanced data (having equal numbers of observations in the subclasses) that we have investigated, the REML estimators are identical to the familiar analysis of variance (ANOVA) estimators for such data. This is not a property possessed by the maximum likelihood estimators of Hartley and Rao [2], and it is a useful one because of optimal properties of ANOVA estimators of variance components from balanced data.

2. The Model

2.1 The general case

The model for $\mathbf{y}$, a vector of N observations, is specified in terms that closely follow the notation of Hartley and Rao [2], Hartley and Vaughan [3] and Hemmerle and Hartley [6]. We take

$$\mathbf{y} = \mathbf{X}\boldsymbol{\mu} + \mathbf{U}_1\mathbf{b}_1 + \cdots + \mathbf{U}_c\mathbf{b}_c + \mathbf{e} \qquad (1)$$

where $\mathbf{y}$ is a vector of N observations,
$\boldsymbol{\mu}$ is a vector of k unknown constants, the fixed effects of the model,
$\mathbf{X}$ is an $N \times k$ incidence matrix, of full column rank, corresponding to $\boldsymbol{\mu}$ and with $k < N$,
$\mathbf{U}_i$ is an $N \times m_i$ design matrix associated with the ith random factor, with $\sum_{i=1}^{c} m_i + k < N$,
$\mathbf{b}_i$ is a vector of m_i random variables which are i.i.d. $N(0, \sigma_i^2)$, with the $\mathbf{b}_i$'s being mutually independent,
and $\mathbf{e}$ is a vector of N random variables which are i.i.d. $N(0, \sigma^2)$ and independent of the $\mathbf{b}_i$'s.

Hence $\mathbf{y}$ has a multivariate normal distribution with mean and variance

$$E(\mathbf{y}) = \mathbf{X}\boldsymbol{\mu} \quad \text{and} \quad \text{var}(\mathbf{y}) \equiv \mathbf{V} = \mathbf{H}\sigma^2 \qquad (2)$$

where

$$\mathbf{H} = \sum_{i=1}^{c} \gamma_i \mathbf{U}_i \mathbf{U}_i' + \mathbf{I}_N \quad \text{for} \quad \gamma_i = \sigma_i^2/\sigma^2. \qquad (3)$$

The symbol $\boldsymbol{\mu}$ for fixed effects emphasizes the generality of the model insofar as fixed effects are concerned. $\boldsymbol{\mu}$ is a vector of the maximum number of linearly independent estimable functions of the fixed effects. The simplest such vector has as its elements the population means of those of the sub-most cells of the fixed effects factors that contain data. The corresponding $\mathbf{X}$ of (1) then has a simple form. Define $\mathbf{y}$ as being the observations ordered so that all those within each sub-most cell of the fixed effects factors follow one another sequentially. If there are k such cells containing data, with the tth one having $n_t \neq 0$ observations, then

$$\mathbf{X} = \begin{bmatrix} \mathbf{1}_{n_1} & 0 & \cdots & 0 \\ 0 & \mathbf{1}_{n_2} & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & \cdots & & \mathbf{1}_{n_k} \end{bmatrix} = \sum_{t=1}^{k}{}^{+} \mathbf{1}_{n_t} \qquad (4)$$

where $\mathbf{1}_{n_t}$ is a vector of n_t ones and where $\sum^+$ represents a direct sum of matrices.

2.2 Example

Illustration of this notation is given in terms of the numerical example taken from Bowker and Lieberman [1] as used by Hemmerle and Hartley [6]. It consists initially of 3 observations in each cell of a 2-way cross-classification with 3 rows and 2 columns. The model for y_{pqr}, the rth observation in the pth row and qth column is

$$y_{pqr} = \mu + \alpha_p + \beta_q + (\alpha\beta)_{pq} + e_{pqr}$$

for $p = 1, 2, 3$, $q = 1, 2$ and $r = 1, 2, 3$ where μ is a general mean, α_p is the effect due to the pth row, β_q is that due to the qth column, $(\alpha\beta)_{pq}$ is the interaction effect and e_{pqr} is the error term. Hemmerle and Hartley [6] amend the example to illustrate unbalanced data by dropping two observations so that the numbers of observations in the cells are as shown in Table 1. Then $r = 1, \cdots, n_{pq}$ for $n_{pq} = 2$ or 3. In considering data of this nature as being from a mixed model with row effects fixed, there are for model (1) $N = 16$ observations, with $c = 2$ random factors, namely columns with $m_1 = 2$ levels and interactions with $m_2 = 6$ levels. The variance components ratios for these factors are respectively $\gamma_1 = \sigma_\beta^2/\sigma^2$ and $\gamma_2 = \sigma_{\alpha\beta}^2/\sigma^2$ in accord with (3). The sub-most cells of the fixed effects factors are the rows, of which there are three, so that for $\mathbf{X}$ of (4), $k = 3$ and the values of n_t are $n_1 = 5$, $n_2 = 6$ and $n_3 = 5$.

3. The Estimators

The logarithm of the likelihood for $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\mu}, \mathbf{H}\sigma^2)$ of (2) is

$$\lambda = -\tfrac{1}{2}N \log 2\pi - \tfrac{1}{2}N \log \sigma^2 - \tfrac{1}{2} \log |\mathbf{H}| - \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu})'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu})/\sigma^2. \qquad (5)$$

To partition this into two parts one of which is free of $\boldsymbol{\mu}$, Patterson and Thompson [7] suggest the

Table 1—*Example: number of observations*

	Column 1	Column 2	Total
Row 1	3	2	5
Row 2	3	3	6
Row 3	2	3	5

singular transformation $\mathbf{y}'[\mathbf{S} \vdots \mathbf{H}^{-1}\mathbf{X}]$ where

$$\mathbf{S} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \sum_{t=1}^{k}{}^{+}(\mathbf{I}_{n_t} - n_t^{-1}\mathbf{J}_{n_t}), \tag{6}$$

is symmetric and idempotent, with $\mathbf{J}_{n_t}$ being an $n_t \times n_t$ matrix with every element unity. Since $\mathbf{SX}$ is null, $\mathbf{Sy}$ is distributed $N(\mathbf{0}, \mathbf{SHS}\sigma^2)$ independently of $\mathbf{X}'\mathbf{H}^{-1}\mathbf{y}$.

It is clear that the distribution of $\mathbf{Sy}$ is free of the fixed effects $\boldsymbol{\mu}$. Its likelihood function therefore forms the basis of our derivation of the estimators of the variance components involved in $\mathbf{H}\sigma^2$. However, to avoid the singularity of $\mathbf{SHS}$ arising from the form of $\mathbf{S}$ shown in (6), we use an alternative to $\mathbf{S}$ derived from it by deleting its n_1th, $(n_1 + n_2)$th, $(n_1 + n_2 + n_3)$th, $\cdots$, and $(n_1 + n_2 + \cdots + n_k)$th rows. Such a matrix has order $(N - k) \times N$, and denoting it by $\mathbf{T}$, we have

$$T = \sum_{t=1}^{k+} [(\mathbf{I}_{n_t-1} \vdots \mathbf{0}\mathbf{1}_{n_t-1}) - n_t^{-1}\mathbf{J}_{(n_t-1)\times n_t}] \tag{7}$$

$$= \sum_{t=1}^{k}{}^{+}(\mathbf{I}_{n_t-1} - n_t^{-1}\mathbf{J}_{n_t-1} \vdots -n_t^{-1}\mathbf{1}_{n_t-1}) \tag{8}$$

where $\mathbf{0}\mathbf{1}_a$ is a vector of zeros of order a, and $\mathbf{J}_{(n_t-1)\times n_t}$ is a matrix of order $(n_t - 1) \times n_t$ whose every element is unity. From $\mathbf{X}$ of (4) it is readily seen that

$$\mathbf{TX} = \mathbf{0} \tag{9}$$

and by the nature of $\mathbf{T}$ itself, it is easy to show that

$$\mathbf{T}'(\mathbf{TT}')^{-1}\mathbf{T} = \mathbf{S}. \tag{10}$$

The transformation now used is

$$\mathbf{z} = \begin{bmatrix} T \\ \mathbf{X}'\mathbf{H}^{-1} \end{bmatrix}\mathbf{y} = \begin{bmatrix} T\mathbf{y} \\ \mathbf{X}'\mathbf{H}^{-1}\mathbf{y} \end{bmatrix} \tag{11}$$

with, in view of (9), its distribution being

$$\mathbf{z} \sim N\left[\begin{pmatrix} \mathbf{0} \\ \mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \mathbf{THT}'\sigma^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\sigma^2 \end{pmatrix}\right]. \tag{12}$$

This transformation is non-singular, because $\mathbf{X}'$ and $\mathbf{T}$ of (4) and (7) each have full row rank, and from (9) the rows of $\mathbf{T}$ are linearly independent of those of $\mathbf{X}'$.

Now consider the log likelihood of $\mathbf{z}$. It is, from (11) and (12), the log likelihoods of $\mathbf{Ty}$ and of $\mathbf{X}'$-$\mathbf{H}^{-1}\mathbf{y}$ which we denote by λ_1 and λ_3 respectively:

$$\lambda_1 = -\tfrac{1}{2}(N - k)\log 2\pi - \tfrac{1}{2}(N - k)\log \sigma^2 - \tfrac{1}{2}\log|\mathbf{THT}'| - \tfrac{1}{2}\mathbf{y}'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{Ty}/\sigma^2 \tag{13}$$

and

$$\lambda_2 = -\tfrac{1}{2}k\log 2\pi - \tfrac{1}{2}k\log\sigma^2 - \tfrac{1}{2}\log|\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}| - \tfrac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu})'\mathbf{H}^{-1} \cdot\mathbf{X}(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\mu})/\sigma^2. \tag{14}$$

With λ_1 not involving $\boldsymbol{\mu}$ the estimators of σ^2 and the γ_i's, called restricted maximum likelihood (REML) estimators are, following the method of Patterson and Thompson [7], those values of σ^2 and the γ_i's that maximize λ_1. Differentiation of (13) gives

$$\frac{\partial\lambda_1}{\partial\sigma^2} = -\tfrac{1}{2}(N - k)/\sigma^2 + \tfrac{1}{2}\mathbf{y}'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{Ty}/\sigma^4 \tag{15}$$

and

$$\frac{\partial\lambda_1}{\partial\gamma_i} = -\tfrac{1}{2}\,\mathrm{tr}\,[\mathbf{U}_i'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{TU}_i] + \tfrac{1}{2}\mathbf{y}'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{TU}_i\mathbf{U}_i'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{Ty}/{}^2\sigma \quad \text{for} \quad i = 1, 2, \cdots, c \tag{16}$$

where tr $(\mathbf{Q})$ is the trace of a matrix $\mathbf{Q}$.

Equating (15) and (16) to zero gives the REML estimators. The resulting equations clearly have no analytic solution and have to be solved numerically. An iterative procedure is to first assign initial values to $\boldsymbol{\gamma}' = \{\gamma_1, \cdots, \gamma_c\}$ and then (i) solve

$$\hat{\sigma}^2 = \mathbf{y}'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{Ty}/(N - k) \tag{17}$$

based on (15), and (ii) use the γ-values, and $\hat{\sigma}^2$ from (17), to calculate new γ-values that make (16) closer to zero. Repetition of (i) and (ii), ending at (i), is continued until a desired degree of accuracy is attained.

4. Computing Procedures

Although Patterson and Thompson [7] give a procedure based on Fisher's iterative method for $c = 1$ and suggest how to use it for $c > 1$, the Newton-Raphson technique is well suited to the problem of finding successive values of $\boldsymbol{\gamma}$ that zeroize (16), and has been effectively applied by Hemmerle and Hartley [6] to similar equations of the Hartley and Rao [2] maximum likelihood method. We use their application here, first adapting a transformation they use, which simplifies notation and computing procedures.

4.1 The W-transformation

The Newton-Raphson technique for finding values of the elements of $\boldsymbol{\gamma}$ that zeroize (16) utilizes the second-order partial derivatives of λ_1 with respect to the γ_i's. These are, using (16)

$$\frac{\partial^2\lambda_1}{\partial\gamma_i\,\partial\gamma_j} = \tfrac{1}{2}\,\mathrm{tr}\,[\mathbf{U}_i'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{TU}_j\mathbf{U}_j'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{TU}_i] - \mathbf{y}'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{TU}_i\mathbf{U}_i'\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{TU}_j\mathbf{U}_j'\mathbf{T}' \cdot(\mathbf{THT}')^{-1}\mathbf{Ty}/\sigma^2 \quad \text{for} \quad i, j = 1, 2, \cdots, c. \tag{18}$$

The matrix products in (16) and (18) are submatrices of the following transformation to $\mathbf{W}$

suggested by Hemmerle and Hartley [6]:

$$\mathbf{W} = \{\mathbf{W}_{ij}\} \quad \text{for} \quad i, j = 1, 2, \cdots, c + 1$$
$$= \begin{bmatrix} \mathbf{U}' \\ \mathbf{y}' \end{bmatrix} \mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{T}[\mathbf{U} \ \mathbf{y}]. \tag{19}$$

Thus for $\mathbf{W}_{i,c+1} \equiv \mathbf{w}_i$ (16) and (18) are

$$\frac{\partial \lambda_1}{\partial \gamma_i} = -\tfrac{1}{2} \operatorname{tr}(\mathbf{W}_{ii}) + \tfrac{1}{2}\mathbf{w}_i'\mathbf{w}_i/\sigma^2 \tag{20}$$

and

$$\frac{\partial^2 \lambda_1}{\partial \gamma_i \, \partial \gamma_j} = \tfrac{1}{2} \operatorname{tr}(\mathbf{W}_{ij}\mathbf{W}_{ij}') - \mathbf{w}_i'\mathbf{W}_{ij}\mathbf{w}_j/\sigma^2 \tag{21}$$

for $i, j = 1, \cdots, c$, and (17) is

$$\hat{\sigma}^2 = W_{c+1,c+1}/(N - k). \tag{22}$$

The elements of $\mathbf{W}$ in (20)–(22) require, from (19), computing $(\mathbf{THT}')^{-1}$, of order $N - k$ which is less than N, the order of the matrix to be inverted for maximum likelihood. For many data sets this will be impossibly large computationally, but the inversion can be reduced to that of a matrix of order $m. = \sum_{i=1}^{c} m_i$, the total number of levels of all random effects in the model. Although for some data sets this will also be too large, it is always less than $N - k$, frequently much less, and in many instances will be such that the inversion can be computed. To achieve this reduction note from (3) that

$$\mathbf{H} = \mathbf{I} + \mathbf{UDU}' \tag{23}$$

for

$$\mathbf{U} = [\mathbf{U}_1 \ \mathbf{U}_2 \ \cdots \ \mathbf{U}_c] \tag{24}$$

and

$$\mathbf{D} = \sum_{i=1}^{c} {}^{+}\gamma_i \mathbf{I}_{m_i} . \tag{25}$$

Then (see appendix)

$$\mathbf{T}'(\mathbf{THT}')^{-1}\mathbf{T} = \mathbf{S} - \mathbf{SUM}^{-1}\mathbf{U}'\mathbf{S} \tag{26}$$

for

$$\mathbf{M} = \mathbf{D}^{-1} + \mathbf{U}'\mathbf{SU}, \quad \text{of order } m. = \sum_{i=1}^{c} m_i . \tag{27}$$

Now define

$$\mathbf{W}_0 = \begin{bmatrix} \mathbf{U}' \\ \mathbf{y}' \end{bmatrix} \mathbf{S}[\mathbf{U} \ \mathbf{y}] = \begin{bmatrix} \mathbf{W}_{00} & \mathbf{w}_0 \\ \mathbf{w}_0' & w_0 \end{bmatrix} \tag{28}$$

which, because of (19) is $\mathbf{W}$ with $\mathbf{H}$ replaced by $\mathbf{I}$. Then on using (26)–(28) in (19), $\mathbf{W}$ becomes

$$\mathbf{W} = \mathbf{W}_0 - \begin{bmatrix} \mathbf{U}' \\ \mathbf{y}' \end{bmatrix} \mathbf{SUM}^{-1}\mathbf{U}'\mathbf{S}[\mathbf{U} \ \mathbf{y}] \tag{29}$$

$$= \begin{bmatrix} \mathbf{W}_{00} - \mathbf{W}_{00}\mathbf{M}^{-1}\mathbf{W}_{00} & \mathbf{w}_0 - \mathbf{W}_{00}\mathbf{M}^{-1}\mathbf{w}_0 \\ \mathbf{w}_0' - \mathbf{w}_0'\mathbf{M}^{-1}\mathbf{W}_{00} & w_0 - \mathbf{w}_0'\mathbf{M}^{-1}\mathbf{w}_0 \end{bmatrix}. \tag{30}$$

4.2 Implementation

Calculating $\mathbf{W}$ of (30) requires calculating (27) and (28). First (28): from $\mathbf{S}$ of (6), the leading term of (28) is

$$\mathbf{W}_{00} = \mathbf{U}'\mathbf{SU} = \mathbf{U}'\mathbf{U}$$
$$- \mathbf{U}'\mathbf{X} \operatorname{diag}\{1/n_1 \cdots 1/n_k\} \mathbf{X}'\mathbf{U} \tag{31}$$

where $\mathbf{U}'\mathbf{U}$ is the familiar "coefficient matrix" for the random effects; (i.e., if $\boldsymbol{\mu}$ were null and the random effects were in fact fixed, the normal equations for them would be $\mathbf{U}'\mathbf{U}\mathbf{b}^0 = \mathbf{U}'\mathbf{y}$). In (31) a typical sub-matrix of $\mathbf{U}'\mathbf{X}$ is

$\mathbf{U}_i'\mathbf{X} = \{n_{i(j),t}\}$ for $j = 1, \cdots, m_i$ and $t = 1, \cdots, k$, and $m_i \times k$ matrix whose typical element $n_{i(j),t}$ is the number of observations in the jth level of the ith random effects factor and the tth sub-most cell of the fixed effects factors. (32)

In the example of Table 1, the rows are considered fixed, with $k = 3$ levels. The first random factor is columns, with $m_1 = 2$ levels, and for (32)

$$\mathbf{U}_1'\mathbf{X} = \{n1_{(j),t}\} \quad \text{for} \quad j = 1, 2, \quad \text{and} \quad t = 1, 2, 3$$
$$= \begin{bmatrix} 3 & 3 & 2 \\ 2 & 3 & 3 \end{bmatrix}.$$

The second random factor is interactions, with $m_2 = 6$ levels and in (32)

$$\mathbf{U}_2'\mathbf{X} = \{n_{2(j),t}\} \text{ for } j = 1, \cdots, 6 \text{ and } t = 1, 2, 3$$
$$= \begin{bmatrix} 3 & 0 & 0 \\ 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 3 \end{bmatrix}.$$

A second term in (28) is

$$\mathbf{w}_0 = \mathbf{U}'\mathbf{Sy} = \{\mathbf{U}_i'\mathbf{Sy}\}, \quad m_i \times 1, \quad \text{for } i = 1, \cdots, c.$$

From (6), $\mathbf{Sy} = \mathbf{x}$ is the vector $\mathbf{y}$ with each observation replaced by its deviation from the cell mean of the sub-most cell of the fixed effects factors in which it occurs:

$$\mathbf{x} = \mathbf{Sy} = \mathbf{y} - \left(\sum_{t=1}^{k} {}^{+} n_t^{-1} \mathbf{J}_{n_t}\right)\mathbf{y}$$
$$= \mathbf{y} - [\{\bar{y}_t.\mathbf{1}_{n_t}\} \text{ for } t = 1, \cdots, k].$$

Hence

$$\mathbf{w}_0 = \{\mathbf{U}_i'\mathbf{x}\} = \{\text{an } m_i \times 1 \text{ vector of totals of the } x\text{'s, totalled over each level of the } i\text{th random factor}\} \text{ for } i = 1, \cdots, c. \quad (33)$$

To illustrate $\mathbf{x}$ from the example we use the familiar dot and bar notation for totals and means; e.g., $y_{11.} = \sum_{r=1}^{3} y_{11r}$ and $\bar{y}_{11.} = y_{11.}/3$. Then $\mathbf{x} = \{y_{pqr} - \bar{y}_{p..}\}$ for $p = 1, 2, 3$, $q = 1, 2$ and $r = 1, \cdots, n_{pq}$ in lexicon order and $\mathbf{w}_0$ of (33) is

$$\mathbf{w}_0 = \begin{bmatrix} y_{.1.} - (3\bar{y}_{1..} + 3\bar{y}_{2..} + 2\bar{y}_{3..}) \\ y_{.2.} - (2\bar{y}_{1..} + 3\bar{y}_{2..} + 3\bar{y}_{3..}) \\ y_{11.} - 3\bar{y}_{1..} \\ y_{12.} - 2\bar{y}_{1..} \\ y_{21.} - 3\bar{y}_{2..} \\ y_{22.} - 3\bar{y}_{2..} \\ y_{31.} - 2\bar{y}_{3..} \\ y_{32.} - 3\bar{y}_{3..} \end{bmatrix},$$

The final term for (28) is

$$w_0 = \mathbf{y}'\mathbf{S}\mathbf{y} = \mathbf{x}'\mathbf{x} = \text{total sum of squares of the } x\text{'s} = \text{within cell sum of squares of the } y\text{'s for the } k \text{ sub-most cells of the fixed effects factors.} \quad (34)$$

and in the example this is

$$w_0 = \sum_{p=1}^{3} \sum_{q=1}^{2} \sum_{r=1}^{n_{pq}} (y_{pqr} - \bar{y}_{p..})^2.$$

Having calculated $\mathbf{W}_{00}$, $\mathbf{w}_0$ and w_0 for $\mathbf{W}_0$ of (28), $\mathbf{M}^{-1}$ required for $\mathbf{W}$ in (30) comes from (27) and (28) as

$$\mathbf{M} = \mathbf{D}^{-1} + \mathbf{W}_{00}. \quad (35)$$

Hence, as suggested by R. Thompson (personal communication) the matrix and vector terms of $\mathbf{W}$ are

$$\mathbf{W}_{00} - \mathbf{W}_{00}\mathbf{M}^{-1}\mathbf{W}_{00} = \mathbf{D}^{-1} - \mathbf{D}^{-1}\mathbf{M}^{-1}\mathbf{D}^{-1}$$

and (36)

$$\mathbf{w}_0 - \mathbf{W}_{00}\mathbf{M}^{-1}\mathbf{w}_0 = \mathbf{D}^{-1}\mathbf{M}^{-1}\mathbf{w}_0,$$

the scalar term, $w_0 - \mathbf{w}_0'\mathbf{M}^{-1}\mathbf{w}_0$ remaining as is. As Thompson points out, the computational advantages of the right-hand sides of (36) over the left are those of multiplication of $\mathbf{M}^{-1}$ by diagonal rather than symmetric matrices.

With these expressions, implementation of the iterative technique can be carried out exactly as suggested by Hemmerle and Hartley [6]. First, from the data, calculate $\mathbf{W}_{00}$ using (31)–(34). Then assign an initial set of values to the γ_i's, $i = 1, \cdots, c$, and use them in $\mathbf{D}$ of (25), thence in $\mathbf{M}$ of (35), and calculate $\mathbf{M}^{-1}$. Then use $\mathbf{M}^{-1}$ in (36) and (30) to obtain $\mathbf{W}$. Elements of $\mathbf{W}$ are then used in (20), (21) and (22) for obtaining, by the Newton-Raphson procedure, solutions to equations formed by equating (15) and (16) to zero. At each iteration, it is $\mathbf{D}$ of (25) and, through (35), $\mathbf{M}$ that change, and hence through (36) the terms of $\mathbf{W}$.

5. Estimation of Fixd Effects

In the model described in (1) and (2), the maximum likelihood estimator of $\boldsymbol{\mu}$ on assuming $\mathbf{H}$ known is

$$\hat{\boldsymbol{\mu}} = (\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}^{-1}\mathbf{y}, \quad (37)$$

where from (23)

$$\mathbf{H}^{-1} = (\mathbf{I} + \mathbf{U}\mathbf{D}\mathbf{U}')^{-1} = \mathbf{I} - \mathbf{U}(\mathbf{D}^{-1} + \mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'.$$

On replacing the γ_i in $\mathbf{D}^{-1}$ by their REML estimators and denoting the corresponding value of $\mathbf{H}$ by $\tilde{\mathbf{H}}$, an estimator of $\boldsymbol{\mu}$ suggested by (37) is

$$\tilde{\boldsymbol{\mu}} = (\mathbf{X}'\tilde{\mathbf{H}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{H}}^{-1}\mathbf{y}. \quad (38)$$

Development of the covariance matrix of this estimator would involve acknowledging that $\tilde{\mathbf{H}}$ is only an estimator of $\mathbf{H}$; ignoring this complexity var $(\tilde{\boldsymbol{\mu}})$ is

$$\text{var}(\tilde{\boldsymbol{\mu}}) = (\mathbf{X}'\tilde{\mathbf{H}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{H}}^{-1}\mathbf{H}\tilde{\mathbf{H}}^{-1}\mathbf{X}(\mathbf{X}'\tilde{\mathbf{H}}^{-1}\mathbf{X})^{-1}\sigma^2$$

and on again using $\tilde{\mathbf{H}}$ and $\tilde{\sigma}^2$ in place of $\mathbf{H}$ and σ^2 we have an estimator of this variance as

$$\widehat{\text{var}}(\tilde{\boldsymbol{\mu}}) = (\mathbf{X}'\tilde{\mathbf{H}}^{-1}\mathbf{X})^{-1}\tilde{\sigma}^2.$$

No optimum properties are known for this estimator or for its covariance matrix. It is known, as may be readily observed, that $\hat{\boldsymbol{\mu}}$ does maximize λ_2 of (14), for known $\mathbf{H}$.

6. Large Sample Variances

6.1 Derivation

Searle [8] gives, for a model of $\mathbf{y}$ as described here in (1), (2) and (3), an expression for the information matrix $\mathbf{I}(\boldsymbol{\sigma}^2)$ of the vector of parameters

$$\boldsymbol{\sigma}^{2\prime} = [\sigma_1^{\,2}\sigma_2^{\,2} \cdots \sigma_c^{\,2}\sigma^2]. \quad (39)$$

The same procedure can be used for obtaining the information matrix $\mathbf{I}(\boldsymbol{\varphi})$ of

$$\boldsymbol{\varphi}' = [\varphi_1\varphi_2 \cdots \varphi_{c+1}] = [\boldsymbol{\gamma}'\sigma^2] \quad (40)$$

for

$$\boldsymbol{\gamma}' = [\gamma_1\gamma_2 \cdots \gamma_c] \quad \text{for} \quad \gamma_i = \sigma_i^{\,2}/\sigma^2, \quad (41)$$

using **Ty** as the data vector. The result (see appendix) is

$$\mathbf{I}(\varphi) = \frac{1}{2}\begin{bmatrix} \{\mathrm{tr}\,(\mathbf{W}_{ij}\mathbf{W}_{ij}')\} & \{\mathrm{tr}\,(\mathbf{W}_{ii})\}/\sigma^2 \\ \{\mathrm{tr}\,(\mathbf{W}_{ii})\}'/\sigma^2 & (N-k)/\sigma^2 \end{bmatrix}$$

$$i, j = 1, 2, \cdots, c \qquad (42)$$

for **W** of (19); and, of course, the large-sample covariance matrix of the REML estimator $\tilde{\varphi}$ of φ (i.e., of the maximum likelihood estimator of γ' and σ^2 from **Ty**) is

$$\mathrm{var}\,(\tilde{\varphi}) = [\mathbf{I}(\varphi)]^{-1}. \qquad (43)$$

Should the large sample covariance matrix of the REML estimator of $\tilde{\boldsymbol{\sigma}}^2$ be required we use the relationship between the two information matrices given in Zacks [10], namely

$$\mathbf{I}(\varphi) = \mathbf{B}'\mathbf{I}(\boldsymbol{\sigma}^2)\mathbf{B}, \qquad (44)$$

where **B** is the Jacobian matrix $\{\partial\sigma_i^2/\partial\varphi_j\}$ for $i, j = 1, 2, \cdots, c+1$. In our case this reduces (see appendix) to

$$\begin{aligned} v(\tilde{\sigma}_i^2) &= \gamma_i^2 v(\tilde{\sigma}^2) + 2\sigma^2\gamma_i \,\mathrm{cov}\,(\tilde{\sigma}^2, \tilde{\gamma}_i) + \sigma^4 v(\tilde{\gamma}_i), \\ \mathrm{cov}\,(\tilde{\sigma}_i^2, \tilde{\sigma}_j^2) &= \gamma_i\gamma_j v(\tilde{\sigma}^2) + \sigma^2\gamma_i \,\mathrm{cov}\,(\tilde{\sigma}^2, \tilde{\gamma}_j) \\ &\quad + \sigma^2\gamma_j \,\mathrm{cov}\,(\tilde{\sigma}^2, \tilde{\gamma}_i) \qquad (45) \\ &\quad + \sigma^4 \,\mathrm{cov}\,(\tilde{\gamma}_i, \tilde{\gamma}_j) \end{aligned}$$

and

$$\mathrm{cov}\,(\tilde{\sigma}^2, \tilde{\sigma}_i^2) = \gamma_i v(\tilde{\sigma}^2) + \sigma^2 \,\mathrm{cov}\,(\tilde{\sigma}^2, \tilde{\gamma}_i),$$

for $i, j = 1, \cdots, c$.

6.2 *Computing estimated variances*

The use in (42) of **W** computed in the final round of iteration for the REML estimate $\tilde{\varphi}$ gives an estimate $\tilde{\mathbf{I}}(\varphi)$ of $\mathbf{I}(\varphi)$. The occurrence of elements of **W** in (42) makes this particularly easy to compute because they are an integral part of computing REML estimates. An estimate of var $(\tilde{\varphi})$ of (43) is then obtained as $\mathrm{var}\,(\tilde{\varphi}) = [\tilde{\mathbf{I}}(\varphi)]^{-1}$. Using the elements of this matrix in place of those of var $(\tilde{\varphi})$ in (45) and also replacing σ^2 and γ by their REML estimates yields a corresponding estimate $\widehat{\mathrm{var}}\,(\tilde{\boldsymbol{\sigma}}^2)$ of var $(\tilde{\boldsymbol{\sigma}}^2)$.

7. A Numerical Example

Estimates computed from the 2-way cross classification example described in Sections 2.2 and 4.2 are shown in Table 2. This is the example used by Hemmerle and Hartley [6] who drop 2 observations to exemplify unbalanced data in which they then treat the rows as fixed effects, in order to also have a mixed model. We here consider both the balanced data case and the unbalanced data for the mixed model. The results for the balanced data prompts three comments.

TABLE 2—*Example: estimates from different methods of estimation*

Method of estimation	Estimates of $\gamma_1 = \sigma_\beta^2/\sigma^2$	$\gamma_2 = \sigma_{\alpha\beta}^2/\sigma^2$	σ^2
	Balanced data (all $n_{pq} = 3$)		
ANOVA	21.55	.32	69.78
ML	10.81	0	69.51*
REML	21.55	.32	69.78
	Unbalanced data (n_{pq} as in Table 1)		
Fitting Constants	18.42	.35	78.63
ML	9.33	0	77.53*
REML	18.57	.34	78.84

* See text.

First, the REML and ANOVA estimates are identical; this has been found true of four balanced data cases we have considered analytically (the 1-way layout, the 2-way hierarchical and the 2-way crossed classification with one factor treated as a fixed effects factor, both with and without interaction). Numerical examples of other models are in keeping with these results, suggesting that REML and ANOVA estimators are identical for balanced data generally.

Second, the ML estimate for γ_2 is zero, meaning that the estimate for the interaction variance component $\sigma_{\alpha\beta}^2$ is also zero. This is a consequence of the computing algorithm: when the computed value of an ML estimate is negative that value is put equal to zero. The consequence of this is important: it implies not just that a variance component is going to be estimated as zero but that the factor corresponding to that component (in this case the interaction factor) is not to be in the model—and so the model is changed to exclude it. From the changed model the other variance components are then estimated. This accounts for the σ^2 estimate under ML being different from that under ANOVA in the balanced data case, and different from the fitting constants estimator in the unbalanced data case. In view of the fact that, for balanced data, ANOVA estimators are known to have desirable minimum variance properties, and since in these data the ANOVA estimator of $\sigma_{\alpha\beta}^2 = .32(69.78) = 22.32$, which is far from being zero, the change brought about by ML estimation of eliminating the interaction factor from the model therefore seems radical. The same situation occurs with the unbalanced data, although the fitting constants estimators there do not have minimum variance properties as do ANOVA estimators in balanced data. Nevertheless, ML estimation from these data again deletes the interaction factor. In neither case does REML estimation.

A third feature of the results is that the ML estimate of $\gamma_1 = \sigma_\beta^2/\sigma^2$ in both the balanced and unbalanced data is very close to half that of the REML and ANOVA or fitting constants estimator. This is not just a consequence of the changed model arising from the negative γ_2 estimate already referred to. It also occurs in the with-interaction model itself. In both cases, for balanced data, it results from "degrees of freedom" divisors in the solutions of the ML equations being b instead of $b - 1$. The effect of this is to change the estimator in approximately the ratio $(b - 1)/b$ which equals $\frac{1}{2}$ in this case, since $b = 2$, the number of columns. A similar effect occurs with the unbalanced data. But it does not occur with the REML estimates.

Estimated sampling variances of the estimates are not shown in Table 2 because the basis for these is large sample theory and the data of our example are not a large sample. This position is exacerbated here because estimated sampling variances are based upon the parameter estimates obtained and these we have seen in the case of γ_1 differ greatly as between ML and REML (and ANOVA). Thus in the balanced data case the ML estimate is $\tilde{\gamma}_1 = 10.81$ with $\tilde{v}(\tilde{\gamma}_1) = 134.26$ whereas the REML estimate is $\tilde{\gamma}_1 = 21.55$ with $\tilde{v}(\tilde{\gamma}_1) = 1024.72$. Since there are no simple relationships between estimates and their estimated sampling variances, not even with balanced data, let alone unbalanced data, it is quite impractical to attempt drawing conclusions from such as these.

8. Acknowledgements

Thanks go to numerous colleagues with whom we discussed early drafts of this paper.

9. Appendix

A.1 Generalizations

The structure of $\mathbf{T}$ in (8) is appealing from the point of view of familiar ANOVA designs, but it can be generalized (Harville [4], [5]) as follows. Following Patterson and Thompson [7], define the $N \times (N - k)$ matrix $\mathbf{A}'$ by

$$\mathbf{S} = \mathbf{A}'\mathbf{A} \quad \text{and} \quad \mathbf{A}\mathbf{A}' = \mathbf{I}. \tag{A1}$$

The likelihood associated with any $(N - k)$ linearly independent combinations of the observations (error contrasts), including that associated with $\mathbf{Ty}$, is necessarily proportional to the likelihood associated with $\mathbf{Ay}$. Denoting the probability density function of the random vector $\mathbf{v} = \mathbf{Ay}$ by f_v, then $\log f_v(\mathbf{Ay})$ is (13) with $\mathbf{T}$ replaced by $\mathbf{A}$. For making this replacement observe that by using (23)

$$\mathbf{AHA}' = \mathbf{I} + \mathbf{AUDU}'\mathbf{A}'$$

and

$$(\mathbf{AHA}')^{-1} = \mathbf{I} - \mathbf{AU}(\mathbf{D}^{-1} + \mathbf{U}'\mathbf{SU})^{-1}\mathbf{U}'\mathbf{A}',$$

as may be verified by multiplication; and also

$$\mathbf{A}'(\mathbf{AHA}')^{-1}\mathbf{A} = \mathbf{S} - \mathbf{SU}(\mathbf{D}^{-1} + \mathbf{U}'\mathbf{SU})^{-1}\mathbf{U}'\mathbf{S} \tag{A2}$$

and

$$|\mathbf{AHA}'| = |\mathbf{I} + \mathbf{AUDU}'\mathbf{A}'| = |\mathbf{I} + \mathbf{DU}'\mathbf{SU}| = |\mathbf{D}|\,|\mathbf{D}^{-1} + \mathbf{U}'\mathbf{SU}|, \tag{A3}$$

as may be seen by expanding the determinant of

$$\begin{bmatrix} \mathbf{I} & \mathbf{AU} \\ \mathbf{U}'\mathbf{A}' & -\mathbf{D}^{-1} \end{bmatrix}$$

in two different ways. Even more generally, suppose that $\mathbf{H} = \mathbf{R} + \mathbf{UDU}'$, where $\mathbf{R}$ is symmetric positive definite with $\mathbf{L}$, of order $N \times N$, defined by $\mathbf{R} = \mathbf{LL}'$. Then

$$\mathbf{L}^{-1}\mathbf{y} = \mathbf{L}^{-1}\mathbf{X}\boldsymbol{\mu} + \mathbf{L}^{-1}\mathbf{UB} + \mathbf{L}^{-1}\mathbf{e}$$

and the likelihood associated with any $(N - k)$ "error contrasts" in $\mathbf{y}$ is necessarily proportional to the likelihood associated with any $(N - k)$ error contrasts in $\mathbf{y}^* = \mathbf{L}^{-1}\mathbf{y}$ and is thus proportional to

$$(2\pi\sigma^2)^{\frac{1}{2}(N-k)}\,|\mathbf{D}|^{-\frac{1}{2}}\,|\mathbf{D}^{-1} + \mathbf{U}'\mathbf{S}^*\mathbf{U}|^{-\frac{1}{2}} \cdot \exp\{-\tfrac{1}{2}\mathbf{y}'[\mathbf{S}^* - \mathbf{S}^*\mathbf{U}(\mathbf{D}^{-1} + \mathbf{U}'\mathbf{S}^*\mathbf{U})^{-1}\mathbf{U}'\mathbf{S}^*]\mathbf{y}/\sigma^2\}$$

where $\mathbf{S}^* = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}$.

Note that for $\mathbf{T}$ of (8), used in the body of the paper, $(\mathbf{TT}')^{-1} = \mathbf{K}'\mathbf{K}$ for some non-singular $\mathbf{K}$. Then $\mathbf{A} = \mathbf{KT}$ satisfies (A1), and (A2) yields (26).

The transformation from $\mathbf{y}$ to $\mathbf{z}$ in (11) has been made, and the distribution of $\mathbf{z}$ in (12) has been derived, for a particular $\mathbf{T}$ based on full column rank and incidence properties of $\mathbf{X}$. However, for any $\mathbf{X}_{N\times p}^*$ that spans the same space as any full column rank matrix $X_{N\times k}$, there exists a matrix $\mathbf{M}$ such that $\mathbf{X}_{N\times p}^* = \mathbf{X}_{N\times k}\mathbf{M}_{k\times p}$. In general, $\mathbf{T}$ can then be chosen such that $\mathbf{TX} = \mathbf{0}$ and hence $\mathbf{TX}^* = 0$. Thus the transformation (11) will then have the distribution (12). Generalization of the REML procedure then follows.

A.2 Large sample variances

The information matrix given in Searle [8] for $\hat{\boldsymbol{\sigma}}^2$ of (39) based on

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\mu}, \mathbf{V}) \quad \text{with} \quad \sum_{i=1}^{c+1} \sigma_i^2\mathbf{U}_i\mathbf{U}_i'$$

as in (1), (2) and (3) is

$$\mathbf{I}^*(\boldsymbol{\sigma}^2) = \tfrac{1}{2}\{\text{tr}\,(\mathbf{V}^{-1}\mathbf{U}_i\mathbf{U}_i'\mathbf{V}^{-1}\mathbf{U}_j\mathbf{U}_j')\} \quad \text{for} \quad i, j = 1, 2, \cdots, c+1.$$

Similarly, that for estimating $\boldsymbol{\sigma}^2$ based on $\mathbf{Ty} \sim N(\mathbf{0}, \mathbf{TVT}')$ with $\mathbf{TX} = \mathbf{0}$, is

$$\mathbf{I}(\boldsymbol{\sigma}^2) = \tfrac{1}{2}\{\text{tr}\,[(\mathbf{TVT}')^{-1}\mathbf{TU}_j\mathbf{U}_i'\mathbf{T}'(\mathbf{TVT}')^{-1}\mathbf{TU}_j\mathbf{U}_j'\mathbf{T}']\} \quad \text{for} \quad i, j = 1, 2, \cdots, c+1. \tag{A4}$$

Derivation of (42) and (45) is based on (A4) and requires two uses of (44).

When var $(\tilde{\boldsymbol{\sigma}}^2) = [\mathbf{I}(\tilde{\boldsymbol{\sigma}}^2)]^{-1}$ is needed in practice, it is obtained from (45) by way of var $(\tilde{\boldsymbol{\varphi}})$ from $\mathbf{I}(\boldsymbol{\varphi})$ of (42). But to derive (42) we start with $\mathbf{I}(\boldsymbol{\sigma}^2)$ of (A4) and use (44). In fact, with $\boldsymbol{\varphi}'$ of (40) and $\boldsymbol{\gamma}'$ of (41) we have

$$B = \left\{\frac{\partial \sigma_i^{\,2}}{\partial \varphi_j}\right\}, \qquad i, j = 1, 2, \cdots, c+1$$
$$= \begin{bmatrix} \sigma^2\mathbf{I} & \boldsymbol{\gamma} \\ \mathbf{0} & 1 \end{bmatrix} \tag{A5}$$

and substitution of this and (A4) into (44) leads directly, after some simplification, to (42). Then, with $\mathbf{I}(\boldsymbol{\varphi})$ of (42) we use (44) in reverse to derive (45) as

$$\text{var}\,(\tilde{\boldsymbol{\sigma}}^2) = [\mathbf{I}(\boldsymbol{\sigma}^2)]^{-1} = \mathbf{B}[\mathbf{I}(\boldsymbol{\varphi})]^{-1}\mathbf{B}' = \mathbf{B}\,\text{var}\,(\tilde{\boldsymbol{\varphi}})\mathbf{B}'.$$

This, for $\mathbf{B}$ of (A5) and $\boldsymbol{\varphi}$ of (40), leads to the results shown in (45).

References

[1] Bowker, A. H. and Lieberman, G. J. (1963). *Engineering Statistics*. Prentice-Hall, Englewood Cliffs, N. J.

[2] Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika, 54*, 93–108.

[3] Hartley, H. O. and Vaughan, W. K. (1972). A computer program for the mixed analysis of variance model based on maximum likelihood. In *Statistical Papers in Honor of George W. Snedecor*, Ed. T. A. Bancroft, 129–144, Iowa State University Press, Ames.

[4] Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika, 61*, 383–385.

[5] Harville, D. A. (1974). Personal communication.

[6] Hemmerle, W. J. and Hartley, H. O. (1973). Computing maximum likelihood estimates for the mixed A.O.V. model using the *W*-transformation. *Technometrics, 15*, 819–832.

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

[8] Searle, S. R. (1970). Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrics, 26*, 505–524.

[9] Thompson, W. A., Jr. (1962). The problem of negative estimates of variance components. *Ann. Math. Statist., 33*, 273–289.

[10] Zacks, S. (1971). *The Theory of Statistical Inference*. Wiley, N. Y.