Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems

DAVID A. HARVILLE*

Recent developments promise to increase greatly the popularity of maximum likelihood (ML) as a technique for estimating variance components. Patterson and Thompson (1971) proposed a restricted maximum likelihood (REML) approach which takes into account the loss in degrees of freedom resulting from estimating fixed effects. Miller (1973) developed a satisfactory asymptotic theory for ML estimators of variance components. There are many iterative algorithms that can be considered for computing the ML or REML estimates. The computations on each iteration of these algorithms are those associated with computing estimates of fixed and random effects for given values of the variance components.

KEY WORDS: Variance component estimation; Maximum likelihood; Restricted maximum likelihood; Mixed linear models; Maximum likelihood computations.

1. INTRODUCTION

The testing and estimation procedures associated with the analysis of variance (ANOVA) and the underlying fixed, mixed, and random linear models have been widely used. A long-standing problem associated with the use of the mixed and random models has been the estimation of the variances of the random effects, i.e., the estimation of the variance components. For balanced data, it has been common practice to estimate these parameters by equating the mean squares in the ANOVA table to their expectations. Henderson (1953) developed analogous techniques for unbalanced data which, at least in terms of actual usage, have proved to be very popular. Recently, a bewildering variety of "new" approaches has been proposed. Simultaneously, there has been renewed interest in maximum likelihood techniques for estimating variance components.

A maximum likelihood approach to the estimation of variance components has some attractive features. The maximum likelihood estimators are functions of every sufficient statistic and are consistent and asymptotically normal and efficient (in the sense described by Miller (1973)). Certain deficiencies of various other methods are not shared by maximum likelihood. In particular, the maximum likelihood approach is "always" well-defined, even for the many useful generalizations of the ordinary ANOVA models, and, with maximum likelihood, nonnegativity constraints on the variance components or other constraints on the parameter space cause no conceptual difficulties. Moreover, the maximum likelihood estimates and the information matrix for a given parameterization of the model can be obtained readily from those for any other parameterization. Interest in the maximum likelihood estimators should be enhanced by the recent discovery by Olsen, Seely, and Birkes (1976) that, for at least some unbalanced designs, there exist estimators in the class of locally best translation-invariant quadratic unbiased estimators that have uniformly smaller variance than the Henderson estimators. These locally best estimators are related closely to maximum likelihood estimators (Hocking and Kutner 1975).

In spite of these properties, maximum likelihood estimators of variance components have not been used much in practice. There are several reasons for this neglect; the most important of which is, except in relatively simple settings, the computation of the maximum likelihood estimates requires the numerical solution of a constrained nonlinear optimization problem. Prior to the advent of the electronic computer, this requirement presented a virtually insurmountable barrier to their widespread use. Even after computers became commonplace, maximum likelihood was not much used to estimate variance components because effective computational algorithms were not readily available to practitioners. Recently, a number of results have come to light that promise to make the computation of maximum likelihood estimates of variance components practical in many settings where it was unfeasible before. Even in cases where their computation is still unfeasible, it may be possible to implement an approach similar to the one outlined in Section 7 which mimics maximum likelihood but is simpler computationally.

Two other problems that have kept maximum likelihood from becoming a more popular technique for estimating variance components are the following: (1) The maximum likelihood estimators of the variance components take no account of the loss in degrees of freedom resulting from the estimation of the model's fixed effects. (2) The maximum likelihood estimators are derived under the assumption of a particular parametric form, generally

* David A. Harville is Professor, Department of Statistics, Iowa State University, Ames, IA 50011. This work was supported in part by the Air Force Office of Scientific Research, under Grant No. AFOSR-76-3037. During the early stages of the work, which took place while he was a visitor at the Biometrics Unit of Cornell University, the author benefited substantially from informal conversations with S.R. Searle and C.R. Henderson. Material supplementary to that in the present paper can be found in Harville (1975) and Searle (1976).

June 1977, Volume 72, Number 358
Theory and Methods Section

normal, for the distribution of the data vector. The first of these problems has in effect been eliminated by Patterson and Thompson (1971) through their "restricted maximum likelihood" approach. With regard to the second problem, it will be argued in Section 8.1 that the maximum likelihood estimators derived on the basis of normality may be suitable even when the form of the distribution is not specified.

In the following, an attempt is made to present a unified review of the maximum likelihood approach to variance component estimation. Computational aspects are emphasized. The topics covered include: the current state of maximum likelihood theory as applied to the estimation of variance components, the relationship (shown to be intimate) between the maximum likelihood estimation of the variance components and the estimation of the model's fixed and random effects, the exploitation of that relationship for purposes of computation and approximation, the use of numerical algorithms for computing maximum likelihood estimates of variance components, the use of maximum likelihood as a vehicle for relating the various methods that have been proposed for estimating variance components, and a discussion of directions for future research.

The problem of estimating variance components can be regarded as a special case of a general linear model problem in which the elements of the covariance matrix are known functions of a parameter vector to be estimated. Throughout this article, an effort is made to promote this point of view. Many of the ideas that are discussed are applicable to the more general problem.

2. THE MODEL AND ITS APPLICABILITY

The models that underlie the analysis of variance can all be viewed as special cases of the general linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{Z}\mathbf{b} + \mathbf{e} \ . \tag{2.1}$$

In this model, $\mathbf{y}$ is a $n \times 1$ vector of random variables whose observed values comprise the data points; $\mathbf{X}$ and $\mathbf{Z}$ are matrices of "regressors" with dimensions $n \times p$ and $n \times q$, respectively; $\boldsymbol{\alpha}$ is a $p \times 1$ vector of unobservable parameters, which are called fixed effects; $\mathbf{b}$ is a $q \times 1$ vector of unobservable random effects; and $\mathbf{e}$ is a $n \times 1$ vector of unobservable random errors. Moreover, $E(\mathbf{b}) = \mathbf{0}$, $E(\mathbf{e}) = \mathbf{0}$, and cov $(\mathbf{b}, \mathbf{e}) = \mathbf{0}$. Put $\mathbf{D} = \text{var}\ (\mathbf{b})$, $\mathbf{R} = \text{var}\ (\mathbf{e})$, and $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}'$, so that var $(\mathbf{y}) = \mathbf{V}$. The matrix $\mathbf{X}$ is assumed to be known, but the elements of $\mathbf{D}$, $\mathbf{R}$, and possibly even $\mathbf{Z}$ may be functions of an unobservable parameter vector $\boldsymbol{\theta} = (\theta_1, \ \ldots, \ \theta_m)'$. The parameter space of $\boldsymbol{\alpha}$, $\boldsymbol{\theta}$, is taken to be $\{(\boldsymbol{\alpha}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega\}$, where Ω is some given subset of Euclidean m space such that $\mathbf{R}$ (and thus $\mathbf{V}$) is nonsingular for $\boldsymbol{\theta} \in \Omega$. Put $p^* = \text{rank}\ (\mathbf{X})$, and take $\mathbf{X}^*$ to be a $n \times p^*$ matrix whose columns are any p^* linearly independent columns of $\mathbf{X}$.

In the ordinary mixed and random ANOVA models, there is some number c of random factors, with the ith factor having q_i levels. The levels of each factor are generally taken to be uncorrelated with each other, the levels of the other factors, and the residual effects. Associated with the ith random factor is a parameter σ_i^2 which represents the common variance of its levels. The residual effects are taken to have common variance σ_{c+1}^2. The variances $\sigma_1^2, \ \ldots, \ \sigma_{c+1}^2$ are called variance components. In terms of (2.1), $\mathbf{b}' = (\mathbf{b}_1', \ \ldots, \ \mathbf{b}_c')$, where $\mathbf{b}_i$ is a $q_i \times 1$ vector whose elements are the levels of the ith random factor, $m = c + 1$,

$$\theta_i = \sigma_i^2 \ , \qquad (i = 1, \ \ldots, \ m) \ , \tag{2.2}$$

$$\mathbf{R} = \theta_m \mathbf{I} \ , \quad \mathbf{D} = \text{diag}\ [\theta_1 \mathbf{I}, \ \ldots, \ \theta_{m-1}\mathbf{I}] \ ,$$

$$\mathbf{V} = \theta_m I + \sum_{i=1}^{m-1} \theta_i \mathbf{Z}_i \mathbf{Z}_i' \ ,$$

where $\mathbf{Z}_i$ is a $n \times q_i$ matrix defined by the partitioning $\mathbf{Z} = (\mathbf{Z}_1, \ \ldots, \ \mathbf{Z}_{m-1})$, and

$$\Omega = \{\boldsymbol{\theta} : \theta_m > 0, \theta_i \geq 0 \ (i = 1, \ \ldots, \ m - 1)\} \ .$$

Generally, each row of $\mathbf{Z}_i$ has a single element equal to one, and its remaining elements are equal to zero; so that its jth row serves to indicate which level of the ith random factor enters the equation for the jth data point. At least some columns of $\mathbf{X}$ will usually also have only 0–1 entries. In particular, in the ordinary random ANOVA models, $\mathbf{X} = \mathbf{1}$. ($\mathbf{1}$ denotes a column vector each of whose elements is one.)

The ANOVA models are sometimes parameterized in terms of $\gamma_{c+1} = \sigma_{c+1}^2$ and $\gamma_i = \sigma_i^2/\sigma_{c+1}^2$ $(i = 1, \ \ldots, \ c)$ rather than in terms of $\sigma_1^2, \ \ldots, \ \sigma_{c+1}^2$. If we had taken $\theta_i = \gamma_i$ $(i = 1, \ \ldots, \ m)$, instead of taking $\boldsymbol{\theta}$ to be as specified by (2.2), we would have had

$$\mathbf{D} = \theta_m \ \text{diag}\ [\theta_1 \mathbf{I}, \ \ldots, \ \theta_{m-1}\mathbf{I}]$$

and

$$\mathbf{V} = \theta_m (\mathbf{I} + \sum_{i=1}^{m-1} \theta_i \mathbf{Z}_i \mathbf{Z}_i') \ .$$

A useful variation on the ordinary ANOVA models is obtained by taking the error variance to be heteroscedastic.

Descriptions of specific ANOVA models can be found in Searle (1971a). These models have been applied widely in the biological, agricultural, behavioral, and physical and engineering sciences. Still, it is a mistake to think of linear models only in terms of the ordinary regression and ANOVA models. To do so is to miss many potential applications. In particular, the observations may not all have been taken at the same time so that $\mathbf{y}$ or various subvectors of $\mathbf{y}$ are best regarded as time series, i.e., as having been generated by stochastic processes. Such data are common in many fields, e.g., in economics. Time-series data are often analyzed on the basis of linear models that can be viewed as special cases of (2.1) in which $q = 0$, and $\mathbf{e}$ or subvectors of $\mathbf{e}$ are generated by stochastic processes like autoregressive processes, moving average processes, or mixed autoregressive moving average processes (see Box and Jenkins (1970)). Useful extensions of these special cases can be obtained by supposing that the elements of $\mathbf{b}$ or of various of its subvectors are also

ordered by time and may have been generated by comparable stochastic processes. Models of this kind may be suitable for a wide variety of growth curve data. Also, many types of data ordinarily analyzed by the usual ANOVA models may be better fitted by these extended time-series models. The specification of $\mathbf{Z}$, m, $\mathbf{R}$, $\mathbf{D}$, and Ω and the interpretation of $\mathbf{b}$ and $\boldsymbol{\theta}$ for these special cases will depend, of course, on what is assumed about the underlying stochastic processes.

Note that multivariate linear models as well as univariate linear models are included in (2.1). While the particular models described above are essentially univariate models, i.e., models in which each component of $\mathbf{y}$ represents the same type of measurement, there is nothing in the general formulation (2.1) that excludes situations where different types of measurements are included among the components of $\mathbf{y}$, e.g., its first component might represent a height measurement and its second component a weight measurement. In fact, for each of our univariate examples, there is a corresponding multivariate example that is likewise a special case of (2.1). The ordinary ANOVA models generalize to the models that underlie the multivariate analysis of variance (MANOVA). The multivariate analogs of the extended time-series models form the basis for Kalman filtering techniques, which are much used in engineering applications (see, e.g., Duncan and Horn (1972)), and they could be applied to multivariate growth curve data.

Those special cases of the general linear model (2.1) discussed above are ones in which $\mathbf{Z}$ is a known matrix. There are also useful special cases in which at least some elements of $\mathbf{Z}$ are nontrivial functions of unobservable parameters. In particular, such formulations can be applied to factor analysis models with nonnull elements of $\mathbf{Z}$ representing the factor loadings.

Specialized results are available in the literature for that class of linear models characterized by $\mathbf{V}$ being linear in the parameters, i.e., for those linear models where

$$\mathbf{V} = \sum_{i=1}^{m} \theta_i \mathbf{G}_i \tag{2.3}$$

for $n \times n$ symmetric matrices $\mathbf{G}_1, \ldots, \mathbf{G}_m$ whose elements are known. As Anderson (1970) indicated, this class includes many useful special cases. In particular, $\mathbf{V}$ has the form (2.3) for the usual ANOVA models when we take $\theta_i = \sigma_i^2$ $(i = 1, \ldots, c + 1)$ though not when we take $\theta_i = \gamma_i$.

3. ESTIMATION OF FIXED AND RANDOM EFFECTS

Corresponding to an actual data vector, i.e., an observed value of $\mathbf{y}$, is a realized or sample value of the vector $\mathbf{b}$ of random effects. This value will subsequently be denoted by $\boldsymbol{\beta}$ and can be thought of as a parameter vector just as $\boldsymbol{\alpha}$ is a parameter vector. The only distinction is that something is assumed to be known about the origin of $\boldsymbol{\beta}$. Estimating estimable functions of $\boldsymbol{\alpha}$ is a problem of great practical importance and has been dealt with in many articles. In contrast, the problem of estimating $\boldsymbol{\beta}$ or linear combinations of the components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ has not received much consideration (at least not from statisticians). Nevertheless, the latter problem arises in many applications as described, e.g., by Searle (1974), Henderson (1973c), and Harville (1975). In particular, the problem of estimating or predicting a future data point from data to which (2.1) applies can generally be formulated as a problem of estimating a linear combination of the components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

In the present section, we review some results on the estimation of linear combinations of the components of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. As will become evident in subsequent sections, these results are very relevant to the problem of estimating $\boldsymbol{\theta}$ by maximum likelihood techniques.

Subsequently, we take $\bar{\boldsymbol{\alpha}}$ to be any solution to the normal equations

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\bar{\boldsymbol{\alpha}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \ , \tag{3.1}$$

and put $\bar{\boldsymbol{\beta}} = \mathbf{D}\bar{\mathbf{v}}$, where $\bar{\mathbf{v}} = \mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\bar{\boldsymbol{\alpha}}) = \mathbf{Z}'\mathbf{P}\mathbf{y}$, with $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}$ (for any matrix $\mathbf{B}$, $\mathbf{B}^{-}$ will denote an arbitrary generalized inverse of $\mathbf{B}$, i.e., any solution to $\mathbf{B}\mathbf{B}^{-}\mathbf{B} = \mathbf{B}$).

Since the elements of $\mathbf{D}$, $\mathbf{R}$, and possibly $\mathbf{Z}$ are functions of the parameter vector $\boldsymbol{\theta}$, so are the elements of $\mathbf{V}$, $\bar{\boldsymbol{\alpha}}$, $\bar{\boldsymbol{\beta}}$, and $\bar{\mathbf{v}}$. When we wish to emphasize that the elements of a particular vector or matrix are functions of parameter vectors, we append the appropriate arguments. This notation facilitates the identification of the value of the vector or matrix for particular values of the parameter vectors. Thus, e.g., $\bar{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ is used interchangeably with $\bar{\boldsymbol{\alpha}}$, and, if $\boldsymbol{\theta}^*$ is a particular value of $\boldsymbol{\theta}$, $\bar{\boldsymbol{\alpha}}(\boldsymbol{\theta}^*)$ is the value of $\bar{\boldsymbol{\alpha}}$ for $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Subsequently, we denote the true values of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ by $\boldsymbol{\theta}^+$ and $\boldsymbol{\alpha}^+$.

By the expectation of a random variable, we shall mean its unconditional expectation with respect to the joint distribution of $\mathbf{b}$ and $\mathbf{e}$, as opposed say to its conditional expectation given $\mathbf{b} = \boldsymbol{\beta}$. We refer to an estimator $t(\mathbf{y})$ of a linear combination $\boldsymbol{\lambda}_1'\boldsymbol{\alpha} + \boldsymbol{\lambda}_2'\boldsymbol{\beta}$ of the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as (unconditionally) unbiased if $E[t(\mathbf{y})] \equiv \boldsymbol{\lambda}_1'\boldsymbol{\alpha}$, and call $E[t(\mathbf{y}) - \boldsymbol{\lambda}_1'\boldsymbol{\alpha} - \boldsymbol{\lambda}_2'\mathbf{b}]^2$ its (unconditional) mean squared error.

For the case where $\boldsymbol{\theta}^+$ is known, an answer to the question of how to estimate $\boldsymbol{\lambda}_1'\boldsymbol{\alpha} + \boldsymbol{\lambda}_2'\boldsymbol{\beta}$ is given below in Theorem 1. This estimator differs from that obtained by proceeding as though all the model's effects were fixed. The latter estimator is less efficient in a mean squared error sense.

Theorem 1: For the case where the true value $\boldsymbol{\theta}^+$ of $\boldsymbol{\theta}$ is known, the best (uniformly smallest mean squared error) linear unbiased estimator (BLUE) of a linear combination $\boldsymbol{\lambda}_1'\boldsymbol{\alpha} + \boldsymbol{\lambda}_2'\boldsymbol{\beta}$ of the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, where $\boldsymbol{\lambda}_1'\boldsymbol{\alpha}$ is estimable, is

$$\boldsymbol{\lambda}_1'\bar{\boldsymbol{\alpha}}(\boldsymbol{\theta}^+) + \boldsymbol{\lambda}_2'\bar{\boldsymbol{\beta}}(\boldsymbol{\theta}^+) \ . \tag{3.2}$$

(See Henderson (1973c) and (1975); Harville (1976).)

When $p = 0$ (so that the model contains no fixed effects) or equivalently when $\boldsymbol{\alpha}^+$ is known, Theorem 1

essentially reduces to the result described by Rao (1965, Sect. 4a.11). When $\lambda_2 = 0$, it reduces to the ordinary Gauss-Markov theorem and, when $\lambda_1 = 0$, to a result derived by Henderson (1963).

The following theorem is relevant to the computation of $\tilde{\alpha}$ and $\tilde{\nu}$ (and $\tilde{\beta}$).

Theorem 2: If $\tilde{\alpha}$ and $\tilde{\nu}$ are the $p \times 1$ and $q \times 1$ components of any solution to the linear system

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}ZD \\ Z'R^{-1}X & I + Z'R^{-1}ZD \end{bmatrix} \begin{bmatrix} \tilde{\alpha} \\ \tilde{\nu} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}, \quad (3.3)$$

then $\tilde{\alpha}$ is a solution to (3.1) and $\tilde{\nu} = \bar{\nu}$. Conversely, for any solution $\tilde{\alpha}$ to (3.1), the system (3.3) has a solution whose first component is $\tilde{\alpha}$.

We have, as a corollary to Theorem 2, that

$$\bar{\nu} \equiv (I + Z'R^{-1}ZD)^{-1}Z'R^{-1}(y - X\tilde{\alpha}) , \quad (3.4)$$

and that

$$\bar{\nu} \equiv (I + Z'SZD)^{-1}Z'Sy , \quad (3.5)$$

with $S \equiv R^{-1} - R^{-1}X(X'R^{-1}X)^{-}X'R^{-1}$. These two representations can also be obtained directly from the matrix identities

$$V^{-1} \equiv R^{-1} - R^{-1}ZD(I + Z'R^{-1}ZD)^{-1}Z'R^{-1} , \quad (3.6)$$

$$Z'V^{-1} \equiv (I + Z'R^{-1}ZD)^{-1}Z'R^{-1} ,$$

$$P \equiv S - SZD(I + Z'SZD)^{-1}Z'S ,$$

and

$$Z'P \equiv (I + Z'SZD)^{-1}Z'S , \quad (3.7)$$

which are derivable from results given and cited by Harville (1976).

In our formulation of the ordinary ANOVA models as special cases of the general linear model (2.1), the matrices R, D, and possibly Z and X exhibit relatively simple structures. Some or all of these matrices also have simple structures in other useful special cases of (2.1). The significance of Theorem 2 is that it provides us with the means for exploiting these structures, so as to simplify the computation of $\bar{\nu}$ (and thus $\bar{\beta}$) and a solution to (3.1). If we compute $\bar{\nu}$ and a solution to (3.1) by directly solving the system (3.3), these structures can be used to obvious advantage in computing the entries in the coefficient matrix and right side of (3.3) and again in the actual solution of the system. Likewise, we could exploit these structures by first solving the normal equations (3.1), using (3.6) in their formation, and then computing $\bar{\nu}$ on the basis of (3.4), which is equivalent to employing the system (3.3) after absorbing the $\tilde{\nu}$ equations into the $\tilde{\alpha}$ equations. Alternatively, we could start by computing $\bar{\nu}$ from (3.5) and then compute a solution to (3.1) from the first p equations in the system (3.3), which corresponds to absorbing the $\tilde{\alpha}$ equations of (3.3) into the $\tilde{\nu}$ equations. In carrying out the computations, advantage should be taken of the well-known fact (see, e.g., Westlake 1968) that $F^{-1}C$, where F and C are arbitrary except for obvious restrictions, is computed most efficiently by numerical techniques that solve the linear system $FB = C$ without explicitly forming F^{-1}.

Theorem 2 was presented by Harville (1976) as one in a class of modified versions of a result due to Henderson (1963). Henderson's result applied for $\theta \in \Omega$ such that D is nonsingular and states that, if $\tilde{\alpha}$ and $\tilde{\beta}$ are the components of any solution to

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & D^{-1} + Z'R^{-1}Z \end{bmatrix} \begin{bmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix}, \quad (3.8)$$

then $\tilde{\alpha}$ is a solution to (3.1) and $\tilde{\beta} = \bar{\beta}$. We choose to work with the system (3.3) rather than (3.8) because it is applicable for all $\theta \in \Omega$ and it has $\bar{\nu}$ imbedded in its solution instead of $\bar{\beta}$. Both of these features are useful in relating the maximum likelihood estimation of θ to the estimation of linear combinations of the elements of α and β. Another attractive feature of the system (3.3) is that its use does not require the inversion of D. However, the coefficient matrix of (3.8) is symmetric positive definite or semidefinite, which can be a useful property from a computational standpoint (see, e.g., Westlake 1968).

The elements of $\bar{\beta}$ belong to the class of estimators known as "shrinkers." For $p = 0$ and θ^+ known, $\bar{\beta}$ is formally the same as the Bayes estimator of β provided the distributions of b and e are multivariate normal, and, even in the absence of normality, is linear Bayes in the sense described by Hartigan (1969). For $p > 0$ but θ^+ known, the above approach coincides with what might be characterized as a partially Bayes approach (Harville 1976). If θ^+ is unknown as is being assumed here and is usually the case in practice, then in general (3.2) can no longer be regarded as an estimator of $\lambda_1'\alpha + \lambda_2'\beta$. One way to proceed when θ^+ is unknown is to use (3.2) as an estimator of $\lambda_1'\alpha + \lambda_2'\beta$ with θ^+ replaced by some value of θ. Depending on how this value is chosen, there is a formal resemblance to ridge regression or to the Stein-like or empirical Bayes estimators considered, e.g., by Efron and Morris (1973). In particular, a maximum likelihood estimate of θ can be substituted for θ^+.

4. THE MAXIMUM LIKELIHOOD APPROACH TO THE ESTIMATION OF θ

4.1 Definition

In discussing the maximum likelihood estimation of θ, we take the distribution of y to be of the multivariate normal form, so that the logarithm of the likelihood function differs by only an additive constant from the function

$$L(\theta, \alpha; y) = -(\tfrac{1}{2}) \log [\det (V)] - (\tfrac{1}{2})(y - X\alpha)'V^{-1}(y - X\alpha) ,$$

defined for θ and α such that $\theta \in \Omega$. By definition, maximum likelihood (ML) estimates of θ and α are values satisfying $\theta \in \Omega$ and $L(\theta, \alpha; y) = L_{\sup}(y)$, where

$$L_{\sup}(y) = \underset{\{(\theta, \alpha) : \theta \in \Omega\}}{\text{supremum}} L(\theta, \alpha; y) ,$$

i.e., values at which L assumes a maximum for those $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ such that $\boldsymbol{\theta} \in \Omega$. It is well-known that, for fixed $\boldsymbol{\theta}$, L is maximized with respect to $\boldsymbol{\alpha}$ by taking $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}(\boldsymbol{\theta})$. Thus, putting $L_1^*(\boldsymbol{\theta}; \mathbf{y}) = L[\boldsymbol{\theta}, \tilde{\boldsymbol{\alpha}}(\boldsymbol{\theta}); \mathbf{y}]$, $\tilde{\boldsymbol{\theta}}$ is a ML estimate of $\boldsymbol{\theta}$ if and only if $\tilde{\boldsymbol{\theta}} \in \Omega$ and $L_1^*(\tilde{\boldsymbol{\theta}}; \mathbf{y}) = L_{\sup}(\mathbf{y})$ (i.e., if and only if L_1^* assumes a maximum at $\tilde{\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in \Omega$, in which case a ML estimate of $\boldsymbol{\alpha}$ is $\tilde{\boldsymbol{\alpha}}(\tilde{\boldsymbol{\theta}})$). Similarly, for fixed values of some number a of components of $\boldsymbol{\theta}$, which without loss of generality we take to be the first a components, it may be possible to determine analytically values $\tilde{\theta}_{a+1}(\theta_1, \ldots, \theta_a), \ldots, \tilde{\theta}_m(\theta_1, \ldots, \theta_a)$ that maximize L_1^* for $\theta_{a+1}, \ldots, \theta_m$ such that $\boldsymbol{\theta} \in \Omega$. Then, putting $L_2^*(\theta_1, \ldots, \theta_a; \mathbf{y}) = L_1^*[\{\theta_1, \ldots, \theta_a, \tilde{\theta}_{a+1}(\theta_1, \ldots, \theta_a), \ldots, \tilde{\theta}_m(\theta_1, \ldots, \theta_a)\}; \mathbf{y}]$, $\tilde{\theta}_1, \ldots, \tilde{\theta}_a$ are ML estimates of $\theta_1, \ldots, \theta_a$ if and only if they maximize L_2^* for those $\theta_1, \ldots, \theta_a$ that satisfy $\boldsymbol{\theta} \in \Omega$ for some $(\theta_{a+1}, \ldots, \theta_m)$ value, in which case ML estimates of $\theta_{a+1}, \ldots, \theta_m$ are $\tilde{\theta}_{a+1}(\tilde{\theta}_1, \ldots, \tilde{\theta}_a), \ldots, \tilde{\theta}_m(\tilde{\theta}_1, \ldots, \tilde{\theta}_a)$. In particular, in our alternate formulation of the ordinary ANOVA models as special cases of (2.1), $\theta_i = \gamma_i$ $(i = 1, \ldots, m)$, and, for fixed values of $\theta_1, \ldots, \theta_{m-1}$, L_1^* is maximized for $\theta_m > 0$ by taking

$$\theta_m = (1/n)[\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}}(\theta)]' \cdot [\mathbf{I} + \sum_{i=1}^{m-1} \theta_i \mathbf{Z}_i \mathbf{Z}_i']^{-1}[\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}}(\boldsymbol{\theta})] , \quad (4.1)$$

unless $\mathbf{y}$ lies in the column space of $\mathbf{X}$ (an event of probability zero when $n > p^*$). (The right side of (4.1) does not depend on θ_m because $\tilde{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ does not depend on θ_m in this setting.) Except in certain fairly simple situations, it will be the case $a \geq 1$; i.e., while analytical techniques can often be used to reduce the dimensions of the problem, numerical techniques will ordinarily have to be employed at some point in order to effect a final solution.

Under what conditions does a ML estimate of $\boldsymbol{\theta}$ exist; i.e., under what conditions is there a value of $\boldsymbol{\theta}$ satisfying $\boldsymbol{\theta} \in \Omega$ and $L_1^*(\boldsymbol{\theta}; \mathbf{y}) = L_{\sup}(\mathbf{y})$? Hartley and Rao (1967, Sect. 2) gave conditions which were claimed to insure the existence of ML estimates for the variance components associated with the ordinary ANOVA models. Miller (1973, Appendix D) found a deficiency in the Hartley-Rao conditions and showed how to "fix them up." Their conditions are quite unrestrictive.

4.2 Asymptotic Properties

Anderson (1971) considered the special case of (2.1) where $\mathbf{y}$ is made up of s $(= n/r)$ r-variate vectors that are independently and identically distributed. He showed that, for $s \to \infty$ with r fixed, the usual asymptotic properties of the maximum likelihood method can be extended.

Asymptotic properties are of value in a particular application only if there is reason to believe that the data are extensive enough that the properties hold. Anderson's asymptotic results can be applied with confidence if s is sufficiently large. However, for many useful models of the form (2.1), $\mathbf{y}$ cannot be partitioned into independently and identically distributed subvectors (except trivially by taking $s = 1$) even though n may be very large; so that the above asymptotic formulation is inappropriate. In particular, it is inappropriate for the ordinary ANOVA models (except for relatively simple cases like the balanced random one-way classification).

Hartley and Rao (1967) were the first to attempt an asymptotic theory that would be truly appropriate for the more complicated of the ordinary ANOVA models. They derived the limiting properties of the ML estimators of $\boldsymbol{\alpha}, \gamma_1, \ldots, \gamma_{c+1}$ as $n \to \infty$ and $q_i \to \infty$ $(i = 1, \ldots, c)$ simulataneously in such a way that the number of observations falling into any particular level of any random factor stays below some universal constant. However, Miller (1973) pointed out that the latter restriction greatly limits the applicability of the Hartley-Rao results. For example, it rules out any sequence of increasingly larger balanced random two-way cross-classifications. Miller developed an asymptotic theory for the ordinary ANOVA models which, while it is similar to that presented by Hartley and Rao, does not exclude any cases of real interest. Miller (like Hartley and Rao) required that $p^* = p$ (which causes no real loss of generality) and that the matrix $\mathbf{Z}_i$ consist only of zeroes and ones with exactly one 1 in each row and at least one 1 in each column $(i = 1, \ldots, c)$. He introduced a quantity η_i that can be regarded as the effective number of levels for the ith random factor $(i = 1, \ldots, c)$, defined another quantity η_{c+1} by $\eta_{c+1} = n - \text{rank } (\mathbf{Z})$, and assumed the existence of a function η_0 of n such that the matrix

$$\lim_{n \to \infty} \eta_0^{-1} \mathbf{X}'[\mathbf{V}(\boldsymbol{\theta}^+)]^{-1}\mathbf{X} \quad (4.2)$$

exists and is positive definite. (Our notation differs from Miller's.) Miller showed, under fairly unrestrictive additional assumptions, that, for sequences of designs for which $n \to \infty$ and $\eta_i \to \infty$ $(i = 0, \ldots, c + 1)$ simultaneously in an "orderly way," the likelihood equations for $\boldsymbol{\alpha}$ and $\boldsymbol{\theta} = (\sigma_1^2, \ldots, \sigma_{c+1}^2)$ have a root with probability one (provided the true value $(\sigma_i^2)^+$ of σ_i^2 is greater than zero $(i = 1, \ldots, c)$), and such a root is consistent and asymptotically efficient. Furthermore, denoting the σ_i^2 component of this root by $\hat{\sigma}_i^2$ $(i = 1, \ldots, c + 1)$ (implying that the $\boldsymbol{\alpha}$ component is $\tilde{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}})$ where $\hat{\boldsymbol{\theta}}' = (\hat{\sigma}_1^2, \ldots, \hat{\sigma}_{c+1}^2)$), the limiting distribution of $\sqrt{\eta_0}[\tilde{\boldsymbol{\alpha}}(\hat{\boldsymbol{\theta}}) - \boldsymbol{\alpha}^+]$,

$$\sqrt{\eta_1}[\hat{\sigma}_1^2 - (\sigma_1^2)^+], \ldots, \qquad \sqrt{\eta_{c+1}}[\hat{\sigma}_{c+1}^2 - (\sigma_{c+1}^2)^+]$$

is normal with mean vector $\mathbf{0}$ and covariance matrix diag $(\mathbf{J}_1^{-1}, \mathbf{J}_2^{-1})$, where $\mathbf{J}_1$ is the matrix given by (4.2) and $\mathbf{J}_2$ is the $(c + 1) \times (c + 1)$ matrix with ijth element

$$(\tfrac{1}{2}) \lim (\eta_i \eta_j)^{-\frac{1}{2}} \operatorname{tr} [\mathbf{Z}_i'\{\mathbf{V}(\boldsymbol{\theta}^+)\}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\{\mathbf{V}(\boldsymbol{\theta}^+)\}^{-1}\mathbf{Z}_i] .$$

4.3 Restricted Maximum Likelihood

One criticism of the ML approach to the estimation of $\boldsymbol{\theta}$ is that the ML estimator of that vector takes no account of the loss in degrees of freedom that results from estimating $\boldsymbol{\alpha}$. For the ordinary ANOVA models, the variance-component estimators obtained by solving the likelihood equations do not in general coincide with those obtained

by ANOVA methods (not even in the case of balanced data), and, unlike the latter estimators, they are generally biased (in a downward direction (Patterson and Thompson 1974)), sometimes severely so (Corbeil and Searle 1976b; Patterson and Thompson 1974). In particular, for the ordinary fixed ANOVA or regression models, which collectively comprise the special case of (2.1) where

$$q = 0\ , \quad m = 1\ , \quad \mathbf{V} = \theta_1\mathbf{I}\ , \quad \Omega = \{\theta_1 : \theta_1 > 0\}\ , \tag{4.3}$$

the ML estimator (4.1) of the single "variance component" θ_1 has expectation $\theta_1(n - p^*)/n$, so that it is biased downward by an amount $\theta_1 p^*/n$, which can be significant if the number of degrees of freedom $n - p^*$ is sufficiently small.

These "deficiencies" are eliminated in the restricted maximum likelihood (REML) approach which was developed for specific balanced ANOVA models by several scholars including Russell and Bradley (1958) and Anderson and Bancroft (1952). It was extended to "all" balanced ANOVA models by W.A. Thompson (1962), and was set forth in general form by Patterson and R. Thompson (1971 and 1974). For balanced ANOVA models, the equations that are the likelihood equations in the REML approach have as their solution the standard ANOVA estimates.

By an error contrast, we shall mean a linear combination $\mathbf{u}'\mathbf{y}$ of the observations such that $E(\mathbf{u}'\mathbf{y}) \equiv 0$, i.e., such that $\mathbf{u}'\mathbf{X} = \mathbf{0}$ (where $\mathbf{u}$ does not depend on $\boldsymbol{\theta}$ or $\boldsymbol{\alpha}$). The maximum possible number of linearly independent error contrasts in any set of error contrasts is $n - p^*$. A particular set of $n - p^*$ linearly independent error contrasts is given by $\mathbf{A}_1\mathbf{y}$ where $\mathbf{A}_1$ is a $(n - p^*) \times n$ matrix whose rows are any $n - p^*$ linearly independent rows of the matrix $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^-\mathbf{X}'$. In the REML approach, inferences for $\boldsymbol{\theta}$ are based on the likelihood function associated with $n - p^*$ linearly independent error contrasts rather than on that associated with the full data vector $\mathbf{y}$. It makes no difference which $n - p^*$ linearly independent contrasts are used because the likelihood function for any such set differs by no more than an additive constant (which varies with which error constrasts are included but does not depend on $\boldsymbol{\theta}$ or $\boldsymbol{\alpha}$) from the function

$$\begin{aligned} L_1(\boldsymbol{\theta}; \mathbf{y}) = &-(\tfrac{1}{2})\log[\det(\mathbf{V})] \\ &-(\tfrac{1}{2})\log[\det(\mathbf{X}^{*\prime}\mathbf{V}^{-1}\mathbf{X}^*)] \\ &-(\tfrac{1}{2})(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})\ , \end{aligned}$$

defined for $\boldsymbol{\theta} \in \Omega$ (Harville 1974). (Here, $\mathbf{X}^*$ is as defined in Section 2.) A REML estimator is a value of $\boldsymbol{\theta}$ that maximizes L_1 for $\boldsymbol{\theta} \in \Omega$. As in full ML, numerical techniques must ordinarily be employed to determine the estimate, though sometimes analytical techniques can be used to reduce the numerical problem to that of maximizing a function L_2 (defined analogously to L_2^*) involving only a components of $\boldsymbol{\theta}$.

It might seem that some information is lost by basing inferences for $\boldsymbol{\theta}$ on L_1 instead of L. Patterson and Thompson (1971) argued to the contrary. Suppose that $\mathbf{A}_1\mathbf{y}$ is any vector of $n - p^*$ linearly independent error contrasts, and that $\mathbf{A}_2$ is any $p^* \times n$ matrix of constants such that $\mathbf{A}' = (\mathbf{A}_1', \mathbf{A}_2')$ is nonsingular. The likelihood function for the transformed data vector $\mathbf{A}\mathbf{y}$ is proportional to that for $\mathbf{y}$ so that we can just as well base our inferences on $\mathbf{A}\mathbf{y}$ as on $\mathbf{y}$. Since $E(\mathbf{A}_2\mathbf{y})$ consists of linearly independent estimable functions of $\boldsymbol{\alpha}$, Patterson and Thompson maintained that, in the absence of outside knowledge of $\boldsymbol{\alpha}$, $\mathbf{A}_2\mathbf{y}$ contains no information about $\boldsymbol{\theta}$ and therefore inferences for $\boldsymbol{\theta}$ should be based only on $\mathbf{A}_1\mathbf{y}$. More precisely, $\mathbf{A}_1\mathbf{y}$ would seem to be marginally sufficient for $\boldsymbol{\theta}$ in the sense described by Sprott (1975). A similar argument, pertaining specifically to estimation, is that the full ML estimator of $\boldsymbol{\theta}$ necessarily depends on $\mathbf{y}$ only through a set of $n - p^*$ linearly independent error contrasts, i.e., as a function of $\mathbf{A}\mathbf{y}$ it depends on $\mathbf{A}_1\mathbf{y}$ but not on $\mathbf{A}_2\mathbf{y}$, so that the REML estimator does not ignore any information that is actually *used* by the full approach. (The fact that the full ML estimator of $\boldsymbol{\theta}$ depends only on error contrasts follows upon observing that it depends on $\mathbf{y}$ only through the function L_1^* which, like L_1, depends on $\mathbf{y}$ only through $\mathbf{A}_1\mathbf{y}$.) A related observation has to do with translation invariance. (An estimator $\mathbf{T}(\mathbf{y})$ of a scalar- or vector-valued function of $\boldsymbol{\theta}$ will be called translation invariant if $\mathbf{T}(\mathbf{y} + \mathbf{X}\mathbf{a}) = \mathbf{T}(\mathbf{y})$ for all $\mathbf{y}$ and all $p \times 1$ vectors $\mathbf{a}$.) It is well-known that the REML estimator of $\boldsymbol{\theta}$ is translation invariant. However, the translation invariance of this estimator is not a valid reason for preferring it to the full ML estimator (as has sometimes been maintained), because the full ML estimator is also translation invariant.

How does the REML estimator of $\boldsymbol{\theta}$ compare with the ML estimator with regard to mean squared error (MSE)? In general, the answer depends on the specifics of the underlying model and possibly on $\boldsymbol{\theta}^+$. For models satisfying the conditions (4.3) (i.e., ordinary fixed ANOVA or regression models) the ML estimator of the variance θ_1 has uniformly smaller MSE than the REML estimator when $p^* \leq 4$; however, the REML estimator has the smaller MSE when $p^* \geq 5$ and $n - p^*$ is sufficiently large ($n - p^* > 2$ suffices if $p^* \geq 13$). MSE comparisons between variance-component estimators obtained by solving the likelihood equations for ML and variance-component estimators obtained by solving the likelihood equations for REML were made by Corbeil and Searle (1976b) and by Hocking and Kutner (1975) for several mixed and random ANOVA models.

5. DERIVATION AND COMPUTATION OF DERIVATIVES AND OTHER RELEVANT ITEMS

Various procedures for computing ML or REML estimates of $\boldsymbol{\theta}$ will be discussed in Section 6. These procedures are iterative, requiring the repeated evaluation of L, L_1^*, or L_1; their first- or second-order partial derivatives; the expected values of their second-order partial derivatives; and/or related quantities. In deciding on a procedure for a given application and in implementing that procedure,

it is imperative to know how to evaluate the required items efficiently.

Using well-known results on matrix differentiation (Graybill 1969, Sect. 10.8; Nering 1970, Chap. 6, Sect. 7), we find

$$\begin{aligned}\partial L_1/\partial\theta_i &= -(\tfrac{1}{2})\operatorname{tr}[\mathbf{P}(\partial\mathbf{V}/\partial\theta_i)] \\ &\quad + (\tfrac{1}{2})(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\alpha}})'\mathbf{V}^{-1}(\partial\mathbf{V}/\partial\theta_i)\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\alpha}}) , \\ \partial^2 L_1/\partial\theta_i\partial\theta_k &= -(\tfrac{1}{2})\operatorname{tr}[\mathbf{P}\{(\partial^2\mathbf{V}/\partial\theta_i\partial\theta_k) \\ &\quad - (\partial\mathbf{V}/\partial\theta_i)\mathbf{P}(\partial\mathbf{V}/\partial\theta_k)\}] \\ &\quad + (\tfrac{1}{2})(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\alpha}})'\mathbf{V}^{-1}[(\partial^2\mathbf{V}/\partial\theta_i\partial\theta_k) \\ &\quad - 2(\partial\mathbf{V}/\partial\theta_i)\mathbf{P}(\partial\mathbf{V}/\partial\theta_k)]\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\alpha}}) ,\end{aligned}$$

and

$$E(\partial^2 L_1/\partial\theta_i\partial\theta_k) = -(\tfrac{1}{2})\operatorname{tr}[\mathbf{P}(\partial\mathbf{V}/\partial\theta_i)\mathbf{P}(\partial\mathbf{V}/\partial\theta_k)] .$$

Expressions for the first- and second-order partial derivatives of L with respect to the components of $\boldsymbol{\theta}$ and for the expected values of the latter partials can be obtained from the above expressions by first putting $\mathbf{X} = \mathbf{0}$ and then substituting $\mathbf{y} - \mathbf{X}\boldsymbol{\alpha}$ for $\mathbf{y}$. Expressions for all of the first- and second-order partials and expected second-order partials of L and $L_1{}^*$ are given by Harville (1975).

As observed by Searle (1970), the information matrix associated with the full likelihood function is the matrix diag $[\mathbf{B}^*, (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})]$, where $\mathbf{B}^*$ is the $m \times m$ matrix with ikth element $(\frac{1}{2})\operatorname{tr}[\mathbf{V}^{-1}(\partial\mathbf{V}/\partial\theta_i)\mathbf{V}^{-1}(\partial\mathbf{V}/\partial\theta_k)]$. The information matrix associated with L_1 is the $m \times m$ matrix $\mathbf{B}$ whose ikth element is $(\frac{1}{2})\operatorname{tr}[\mathbf{P}(\partial\mathbf{V}/\partial\theta_i)\mathbf{P}(\partial\mathbf{V}/\partial\theta_k)]$.

In practice, it is generally inefficient and possibly unfeasible computationally to evaluate L, $L_1{}^*$, or L_1; their partial derivatives; their expected second-order partials; or related quantities directly from expressions like those given previously. As noted earlier, the matrices $\mathbf{R}$, $\mathbf{D}$, and possibly $\mathbf{Z}$ and $\mathbf{X}$ often have relatively simple structures. Formulas were given in Section 3 that made it clear how to exploit these structures for purposes of computing BLUE's of linear combinations of the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. We now demonstrate how, by making use of the results of Section 3 and several related identities, comparable formulas can be obtained for the preceding items.

Taking $\mathbf{C}$ to be the coefficient matrix of the linear system (3.3) and $\mathbf{C}^*$ to be the matrix that results from substituting $\mathbf{X}^*$ for $\mathbf{X}$ in $\mathbf{C}$, we find

$$\begin{aligned}&\det(\mathbf{V}) \equiv \det(\mathbf{R})\cdot\det(\mathbf{I}+\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{D}) , \\ &\det(\mathbf{V})\cdot\det(\mathbf{X}^{*\prime}\mathbf{V}^{-1}\mathbf{X}^*) \equiv \det(\mathbf{R})\cdot\det(\mathbf{C}^*) \\ &\quad \equiv \det(\mathbf{R})\cdot\det(\mathbf{X}^{*\prime}\mathbf{R}^{-1}\mathbf{X}^*)\cdot\det(\mathbf{I}+\mathbf{Z}'\mathbf{S}\mathbf{Z}\mathbf{D}) , \quad (5.1)\end{aligned}$$

$$\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\alpha}}) \equiv \mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\alpha}}-\mathbf{Z}\tilde{\boldsymbol{\beta}}) \equiv \mathbf{S}(\mathbf{y}-\mathbf{Z}\tilde{\boldsymbol{\beta}}) , \quad (5.2)$$

$$\begin{aligned}&\mathbf{P} \equiv \mathbf{R}^{-1} - \mathbf{R}^{-1}(\mathbf{X},\mathbf{Z}\mathbf{D})\mathbf{C}^{-}(\mathbf{X},\mathbf{Z})'\mathbf{R}^{-1} , \\ &\mathbf{Z}'\mathbf{P} \equiv (\mathbf{0},\mathbf{I}_{q\times q})\mathbf{C}^{-}(\mathbf{X},\mathbf{Z})'\mathbf{R}^{-1} . \quad (5.3)\end{aligned}$$

Proofs of these identities are outlined by Harville (1975).

To illustrate how the various results can be combined to produce formulas of the desired kind, we note that the representations,

$$\begin{aligned}L_1 &= -(\tfrac{1}{2})\log[\det(\mathbf{R})] - (\tfrac{1}{2})\log[\det(\mathbf{C}^*)] \\ &\quad - (\tfrac{1}{2})\mathbf{y}'R^{-1}(\mathbf{y}-\mathbf{X}\tilde{\boldsymbol{\alpha}}-\mathbf{Z}\tilde{\boldsymbol{\beta}}) \\ &= -(\tfrac{1}{2})\log[\det(\mathbf{R})] - (\tfrac{1}{2})\log[\det(\mathbf{X}^{*\prime}\mathbf{R}^{-1}\mathbf{X}^*)] \\ &\quad - (\tfrac{1}{2})\log[\det(\mathbf{I}+\mathbf{Z}'\mathbf{S}\mathbf{Z}\mathbf{D})] - (\tfrac{1}{2})\mathbf{y}'\mathbf{S}(\mathbf{y}-\mathbf{Z}\tilde{\boldsymbol{\beta}}) ,\end{aligned}$$

follow immediately from (5.1) and (5.2). Next, consider $\partial L_1/\partial\theta_i$. If $\mathbf{D}$ depends on θ_i but $\mathbf{R}$ and $\mathbf{Z}$ do not (as is the case for $i = 1, \ldots, m-1$ in our formulations of the ordinary ANOVA models), then, using (3.7), we have

$$\begin{aligned}\partial L_1/\partial\theta_i &= -(\tfrac{1}{2})\operatorname{tr}[(\mathbf{I}+\mathbf{Z}'\mathbf{S}\mathbf{Z}\mathbf{D})^{-1}\mathbf{Z}'\mathbf{S}\mathbf{Z}(\partial\mathbf{D}/\partial\theta_i)] \\ &\quad + (\tfrac{1}{2})\tilde{\mathbf{v}}'(\partial\mathbf{D}/\partial\theta_i)\tilde{\mathbf{v}} ,\end{aligned}$$

where, if we wish, we could substitute the right side of (5.3) for $(\mathbf{I}+\mathbf{Z}'\mathbf{S}\mathbf{Z}\mathbf{D})^{-1}\mathbf{Z}'\mathbf{S}$. Finally, consider $\partial L_1/\partial\theta_i$, $E(\partial^2 L_1/\partial\theta_i\partial\theta_k)$, and $\partial^2 L_1/\partial\theta_i\partial\theta_k$ specifically for the case of the ordinary ANOVA models, taking the parameterization to be $\theta_i = \gamma_i$ $(i = 1, \ldots, m)$. Upon defining the partitioned matrix

$$\mathbf{T} = \begin{bmatrix}\mathbf{T}_{11} & \cdots & \mathbf{T}_{1c} \\ \vdots & & \vdots \\ \mathbf{T}_{1c} & \cdots & \mathbf{T}_{cc}\end{bmatrix} ,$$

where $\mathbf{T}_{ij}$ is $q_i \times q_j$, to be the lower right corner of any generalized inverse of $\mathbf{C}$ (which is necessarily equal to $(\mathbf{I}+\mathbf{Z}'\mathbf{S}\mathbf{Z}\mathbf{D})^{-1}$, see Harville (1976, p. 392)) and observing that (for the ordinary ANOVA models)

$$\begin{aligned}\mathbf{T}_{ik} + \gamma_{c+1}\gamma_k\sum_{j=1}^{c}\mathbf{T}_{ij}\mathbf{Z}_j'\mathbf{S}\mathbf{Z}_k &= \mathbf{I} , \quad \text{if} \quad k = i , \\ &= \mathbf{0} , \quad \text{if} \quad k \neq i ,\end{aligned}$$

we find

$$\begin{aligned}\partial L_1/\partial\gamma_i &= -(\tfrac{1}{2})\gamma_i^{-1}[q_i - \operatorname{tr}(\mathbf{T}_{ii})] \\ &\quad + (\tfrac{1}{2})\gamma_{c+1}\tilde{\mathbf{v}}_i'\tilde{\mathbf{v}}_i , \quad (5.4)\end{aligned}$$

$$\begin{aligned}\partial L_1/\partial\gamma_{c+1} &= -(\tfrac{1}{2})\gamma_{c+1}{}^{-1}[(n-p^*) \\ &\quad - \mathbf{y}'\mathbf{S}(\mathbf{y}-\mathbf{Z}\tilde{\boldsymbol{\beta}})] , \quad (5.5)\end{aligned}$$

$$\begin{aligned}(-2)E(\partial^2 L_1/\partial\gamma_i\partial\gamma_i) &= \gamma_i{}^{-2}\operatorname{tr}[(\mathbf{I}-\mathbf{T}_{ii})^2] , \\ (-2)E(\partial^2 L_1/\partial\gamma_i\partial\gamma_k) &= \gamma_i{}^{-1}\gamma_k{}^{-1}\operatorname{tr}[\mathbf{T}_{ik}\mathbf{T}_{ki}] , \\ (-2)E(\partial^2 L_1/\partial\gamma_i\partial\gamma_{c+1}) &= \gamma_i{}^{-1}\gamma_{c+1}{}^{-1}[q_i - \operatorname{tr}(\mathbf{T}_{ii})] , \\ (-2)E(\partial^2 L_1/\partial\gamma_{c+1}\partial\gamma_{c+1}) &= \gamma_{c+1}{}^{-2}(n-p^*) , \\ (-2)\partial^2 L_1/\partial\gamma_i\partial\gamma_i &= -\gamma_i{}^{-2}\operatorname{tr}[(\mathbf{I}-\mathbf{T}_{ii})^2] \\ &\quad + 2\gamma_{c+1}\gamma_i{}^{-1}\tilde{\mathbf{v}}_i(\mathbf{I}-\mathbf{T}_{ii})\tilde{\mathbf{v}}_i , \\ (-2)\partial^2 L_1/\partial\gamma_i\partial\gamma_k &= -\gamma_i{}^{-1}\gamma_k{}^{-1}\operatorname{tr}[\mathbf{T}_{ik}\mathbf{T}_{ki}] \\ &\quad - 2\gamma_{c+1}\gamma_k{}^{-1}\tilde{\mathbf{v}}_i'\mathbf{T}_{ik}\tilde{\mathbf{v}}_k , \\ (-2)\partial^2 L_1/\partial\gamma_i\partial\gamma_{c+1} &= \tilde{\mathbf{v}}_i'\tilde{\mathbf{v}}_i , \\ (-2)\partial^2 L_1/\partial\gamma_{c+1}\partial\gamma_{c+1} &= -\gamma_{c+1}{}^{-2}[(n-p^*) \\ &\quad - 2\mathbf{y}'\mathbf{S}(\mathbf{y}-\mathbf{Z}\tilde{\boldsymbol{\beta}})] ,\end{aligned}$$

for $k \neq i = 1, \ldots, c$, where $\tilde{\mathbf{v}}_i$ is the $q_i \times 1$ vector defined by $\tilde{\mathbf{v}}' = (\tilde{\mathbf{v}}_1', \ldots, \tilde{\mathbf{v}}_c')$. (For those of the expressions that involve partial differentiation with respect to γ_i, we require $\gamma_i > 0$, $i = 1, \ldots, c+1$.)

The above representations are closely related to various of the representations given by Patterson and Thompson (1971), Henderson (1973a and 1973b), Hemmerle and Hartley (1973), Thompson (1975), and

Corbeil and Searle (1976a). Depending upon the particulars of the model being considered, it may be possible to further simplify these representations by algebraic means. For example, "complete" simplification is possible in the case of the random two-way nested ANOVA model as evidenced by Searle's results (1970).

Note that first- and second-order partial derivatives and expected second-order partial derivatives for L, L_1^*, and L_1 for one parameterization of (2.1) can be obtained from those for a second parameterization by making use of the chain rule of calculus (see, e.g., Nering (1970, Chap. 6, Sect. 5)). In particular, in the case of the ordinary ANOVA models, expressions for partial derivatives taken with respect to $\sigma_1^2, \ldots, \sigma_{c+1}^2$ can be obtained from those taken with respect to $\gamma_1, \ldots, \gamma_{c+1}$ and vice versa. General formulas for going from one information matrix to the other are given by Zacks (1971, p. 227). The chain rule can also be used to obtain the partial derivatives of L_2^* or L_2 from the partial derivatives of L_1^* or L_1.

The above results completely link the problem described in Section 3 of estimating linear combinations of the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ when $\boldsymbol{\theta}^+$ is known to the problem of evaluating L, L_1^*, and L_1, their first- and second-order partial derivatives, and their expected second-order partial derivatives. For each of the approaches given in Section 3 to the first problem, they point the way to a corresponding approach to the second problem. Note, however, there is a consideration in the second problem not ordinarily present in the first. In the estimation of linear combinations of the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ when $\boldsymbol{\theta}^+$ is known, there is only a single set of computations, while, in the iterative procedures for the ML or REML estimation of $\boldsymbol{\theta}$, similar computations must be performed for each of a sequence of $\boldsymbol{\theta}$ values. When the computations must be carried out for more than one $\boldsymbol{\theta}$ value, they should be accomplished such that, to the greatest extent possible, those operations that depend on $\boldsymbol{\theta}$ are segregated from those that do not, so the latter operations need be performed only once. Hemmerle and Hartley (1973) discuss this point in the context of the ML estimation of variance components, and Corbeil and Searle (1976a) describe the analogous considerations for REML estimation.

In general, the evaluation of first-order partial derivatives can require considerable computations beyond those necessary to evaluate L, L_1^*, or L_1; the evaluation of the expected values of the second-order partial derivatives can require many computations additional to those needed to evaluate the first-order partial derivatives; and the evaluation of the second-order partial derivatives themselves can require still more extensive computations. However, judging from the preceding representations, it would appear, in the case of the ordinary ANOVA models, first- and even second-order derivative information can be had rather cheaply. In assessing the relative difficulty of the computations for any particular application, information on the numerical solution of linear equations such as that provided by Westlake (1968) can be invaluable.

6. NUMERICAL PROCEDURES FOR MAXIMUM LIKELIHOOD ESTIMATION

Ordinarily, we must resort to an iterative numerical procedure to obtain a ML or REML estimate of $\boldsymbol{\theta}$. However, there are simple cases where the estimate can be found by analytical means. Herbach (1959) derived explicit expressions for the ML estimators of the parameters (the mean and two variance components) of the balanced one-way random-effects model. The results of W.A. Thompson (1962) can be used to obtain explicit expressions for the REML estimators of the variance components of any balanced ANOVA model. Thompson worked these out himself for several models including the balanced two-way crossed random-effects ANOVA model (both with and without interaction). While the standard ANOVA estimators of variance components comprise a solution to the likelihood equations for REML in the case of the balanced ANOVA models, the likelihood equations for full ML do not admit an explicit solution for all such models. Explicit solutions to the latter equations exist for the balanced two-way nested ANOVA models, though not for the balanced two-way crossed random-effects ANOVA model with interaction (Hartley and Rao 1967; Miller 1973; Herbach 1959).

There are many iterative numerical algorithms that can be regarded as candidates for computing ML or REML estimates of $\boldsymbol{\theta}$. Some were developed specifically for special cases; e.g., for computing ML estimates of variance components. Others are general procedures for the numerical solution of broad classes of constrained nonlinear optimization problems. There is no real hope for finding a single iterative numerical algorithm for the ML or REML estimation of $\boldsymbol{\theta}$ that will be best, or perhaps even satisfactory, for every application. An algorithm that requires relatively few computations to converge to a ML or REML estimate in one setting may converge slowly or even fail to converge in another. In deciding which among available algorithms to try in a particular application, we must make some judgments about their computational requirements and their other properties as applied to a given setting. This section is devoted to describing the various algorithms and their characteristics. The initial descriptions, given in Subsections 6.1 and 6.2, ignore any complications brought about by constraints on the parameter space, i.e., by $\boldsymbol{\theta}$ being confined to Ω when Ω is a proper subset of Euclidean m space. In Subsection 6.3, several techniques are considered for modifying the various algorithms to cope with constraints.

6.1 Specialized Algorithms

On the kth iteration of an iterative algorithm for producing a ML or REML estimate of $\boldsymbol{\theta}$, the current value for the estimate is converted into a new value. In the following, we denote by $\tilde{\boldsymbol{\theta}}^{(k)}$ the value produced by the algo-

rithm on its kth iteration, and, for any quantity f which is a function of $\boldsymbol{\theta}$, we use $f^{(k)}$ to represent the value of f at $\boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}^{(k)}$, e.g., $\mathbf{V}^{(k)} = \mathbf{V}\{\tilde{\boldsymbol{\theta}}^{(k)}\}$. The value $\tilde{\boldsymbol{\theta}}^{(0)}$ used to start the algorithm must be supplied by the user.

Anderson (1973) and Henderson (1973a) proposed iterative algorithms designed specifically for handling certain special cases of the problem of computing ML estimates of $\boldsymbol{\theta}$. Their approaches are in effect based on manipulating the equation $\partial L_1^*/\partial\boldsymbol{\theta} = \mathbf{0}$ into the form $\boldsymbol{\theta} = \mathbf{g}(\boldsymbol{\theta}; \mathbf{y})$ for some $m \times 1$ vector $\mathbf{g}$ of functions of $\boldsymbol{\theta}$. Nonlinear equations of this form can be solved by the method of successive approximations, which consists of taking $\tilde{\boldsymbol{\theta}}^{(k+1)} = \mathbf{g}\{\tilde{\boldsymbol{\theta}}^{(k)}; \mathbf{y}\}$ (see, e.g., Beltrami (1970, Sect. 1.2)).

Anderson's iterative algorithm for computing a ML estimate of $\boldsymbol{\theta}$ is for the special case where $\mathbf{V}$ has the representation (2.3). Anderson found in effect that $\partial L_1^*/\partial\boldsymbol{\theta} = \mathbf{0}$ can be rewritten as $\mathbf{B}^*\boldsymbol{\theta} = \mathbf{d}$, where $\mathbf{B}^*$ is as defined in Section 5 and $\mathbf{d}$ is the $m \times 1$ vector whose ith element is

$$(\tfrac{1}{2})(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})'\mathbf{V}^{-1}(\partial\mathbf{V}/\partial\theta_i)\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}}) \ .$$

For fixed $\boldsymbol{\theta}$ such that $\mathbf{V}$ is nonsingular, $\mathbf{B}^*$ is necessarily positive semidefinite and the linear system $\mathbf{B}^*\tilde{\boldsymbol{\theta}} = \mathbf{d}$ is consistent for $\tilde{\boldsymbol{\theta}}$ (see LaMotte 1973). When $\mathbf{B}^*$ is nonsingular (and thus positive definite), which is the case if and only if $\mathbf{G}_1, \ldots, \mathbf{G}_m$ are linearly independent matrices, the equation $\mathbf{B}^*\boldsymbol{\theta} = \mathbf{d}$ is equivalent to the equation $\boldsymbol{\theta} = \mathbf{B}^{*-1}\mathbf{d}$. The method of successive approximations as applied to the latter equation is to take the $(k+1)$st iterate to be $\tilde{\boldsymbol{\theta}}^{(k+1)} = [\mathbf{B}^{*(k)}]^{-1}\mathbf{d}^{(k)}$. In the event that sufficient conditions given by Anderson (1969) for the existence of an explicit solution to the likelihood equations are met, the iterative procedure converges in one iteration from any starting value (Miller 1973).

A similar iterative algorithm can be constructed for computing a REML estimate of $\boldsymbol{\theta}$ for the case where $\mathbf{V}$ has the representation (2.3). The likelihood equations for REML can be put in the form $\mathbf{B}\boldsymbol{\theta} = \mathbf{d}$. For fixed $\boldsymbol{\theta}$ with $\mathbf{V}$ nonsingular, $\mathbf{B}$ is positive semidefinite and the linear system $\mathbf{B}\tilde{\boldsymbol{\theta}} = \mathbf{d}$ is consistent for $\tilde{\boldsymbol{\theta}}$ (LaMotte 1973, pp. 316 and 327–8). The matrix $\mathbf{B}$ is nonsingular if and only if θ_i is estimable in the class of quadratic translation-invariant estimators for $i = 1, \ldots, m$ (again see LaMotte 1973), in which case $\mathbf{B}\boldsymbol{\theta} = \mathbf{d}$ is equivalent to $\boldsymbol{\theta} = \mathbf{B}^{-1}\mathbf{d}$. Applying the method of successive approximations to the latter equation yields an iterative algorithm for computing a REML estimate of $\boldsymbol{\theta}$ analogous to Anderson's procedure for computing a ML estimate.

Anderson's iterative algorithm and its REML analog can of course be used to compute ML and REML estimates of the variance components $\sigma_1^2, \ldots, \sigma_{c+1}^2$ associated with the ordinary ANOVA models. However, Anderson's algorithm differs from the iterative algorithm proposed by Henderson (1973a). Henderson's algorithm, which is the same in principle as a procedure discussed by Hartley and Rao (1967, Sect. 5), is designed specifically for computing ML estimates of variance components. By using representations for $\partial L_1^*/\partial\gamma_i$ and $\partial L_1^*/\partial\gamma_{c+1}$ analogous to (5.4) and (5.5), the equations $\partial L_1^*/\partial\gamma_i = 0$ and $\partial L_1^*/\partial\gamma_{c+1} = 0$ can be put in the form

$$\sigma_i^2 = [\tilde{\boldsymbol{\beta}}_i'\tilde{\boldsymbol{\beta}}_i + \sigma_i^2 \operatorname{tr}(\mathbf{T}_{ii}^*)]/q_i \ , \quad i = 1, \ldots, c \ , \tag{6.1}$$

$$\sigma_{c+1}^2 = \mathbf{y}'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}} - \mathbf{Z}\tilde{\boldsymbol{\beta}})/n \ , \tag{6.2}$$

where

$$\tilde{\boldsymbol{\beta}} = \begin{bmatrix} \tilde{\boldsymbol{\beta}}_1 \\ \vdots \\ \tilde{\boldsymbol{\beta}}_c \end{bmatrix} \quad \text{and} \quad (\mathbf{I} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{D})^{-1} = \begin{bmatrix} \mathbf{T}_{11}^* & \cdots & \mathbf{T}_{1c}^* \\ \vdots & & \vdots \\ \mathbf{T}_{c1}^* & \cdots & \mathbf{T}_{cc}^* \end{bmatrix} .$$

Henderson's algorithm consists of applying the method of successive approximations to (6.1) and (6.2). An analogous algorithm for computing REML estimates of $\sigma_1^2, \ldots, \sigma_{c+1}^2$ is obtained by applying the method of successive approximations to

$$\sigma_i^2 = [\tilde{\boldsymbol{\beta}}_i'\tilde{\boldsymbol{\beta}}_i + \sigma_i^2 \operatorname{tr}(\mathbf{T}_{ii})]/q_i \ , \quad i = 1, \ldots, c \ , \tag{6.3}$$

$$\sigma_{c+1}^2 = \mathbf{y}'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}} - \mathbf{Z}\tilde{\boldsymbol{\beta}})/(n - p^*) \ . \tag{6.4}$$

Note that (6.1) and (6.3) can be rewritten as

$$\sigma_i^2 = [\tilde{\boldsymbol{\beta}}_i'\tilde{\boldsymbol{\beta}}_i]/[q_i - \operatorname{tr}(\mathbf{T}_{ii}^*)] \ , \quad i = 1, \ldots, c \ , \tag{6.5}$$

$$\sigma_i^2 = [\tilde{\boldsymbol{\beta}}_i'\tilde{\boldsymbol{\beta}}_i]/[q_i - \operatorname{tr}(\mathbf{T}_{ii})] \ , \quad i = 1, \ldots, c \ . \tag{6.6}$$

Possibly interesting modifications of Henderson's procedure and its REML analog are obtained by applying the method of successive approximations to (6.5) and (6.2), rather than (6.1) and (6.2), and to (6.6) and (6.4), rather than (6.3) and (6.4).

The following lemma was proven by Harville (1975).

Lemma 1: For the ordinary ANOVA models (with $\mathbf{R}$, $\mathbf{D}$, and Ω as specified in Section 2), (i) $\operatorname{tr}(\mathbf{T}_{ii}^*) > 0$; (ii) $\operatorname{tr}(\mathbf{T}_{ii}) > 0$; (iii) $q_i \geq \operatorname{tr}(\mathbf{T}_{ii}^*)$ provided $\sigma_i^2 > 0$, with strict inequality holding if $\mathbf{Z}_i \neq \mathbf{0}$; and (iv) $q_i \geq \operatorname{tr}(\mathbf{T}_{ii})$ provided $\sigma_i^2 > 0$, with strict inequality holding if $\operatorname{rank}(\mathbf{X}, \mathbf{Z}_i) > p^*$.

The lemma implies that Henderson's iterative procedure for computing ML estimates of variance components and its REML analog have an apparently pleasing property (which is not shared by Anderson's algorithm). Suppose that $\mathbf{y}$ does not lie in the column space of $\mathbf{X}$, which is the case with probability one when, e.g., $\mathbf{y}$ is normally distributed. If the algorithms are started with strictly positive values for the variance components, then at no point can the values for the variance components ever become negative. In fact, starting from strictly positive values, they can never reach zero values either, though it is possible for them to attain values arbitrarily close to zero. Note that these algorithms ordinarily should not be started with a zero value for any variance component, since the value for that component would then continue to be zero throughout the iterative procedure. A further implication of the lemma is that the modified Henderson procedure, which is based on (6.5) or (6.6) rather than on (6.1) or (6.3), is well defined, unless at

some point a zero value is attained for some variance component. That is, the denominators of (6.5) and (6.6) are strictly positive unless $\sigma_i^2 = 0$, in which case the denominators are zero. The latter phenomenon causes no difficulty if we agree to take the value $\{\tilde{\sigma}_i^2\}^{(k+1)}$ for σ_i^2 on the $(k+1)$st iteration to be zero whenever $\{\tilde{\sigma}_i^2\}^{(k)} = 0$. The modified algorithms, like the originals, can never reach negative values, nor should they ordinarily be started with zero values for any of the variance components since, once a zero value is inserted, it is never changed. The iterates derived from (6.5) or (6.6) have an intuitively appealing form. On each iteration, σ_i^2 is "estimated" by computing the sum of squares of the "BLUE's" of the components of the $q_i \times 1$ vector $\boldsymbol{\beta}_i$ defined by $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_c')$ and by then dividing by a number between q_i and zero.

6.2 General Algorithms

To locate a ML or REML estimate of $\boldsymbol{\theta}$, we can, in the special cases where they apply, try one of the iterative numerical algorithms described in Section 6.1. We can also consider iterative numerical algorithms developed for the general problem of maximizing an arbitrary function. Moreover, when confronted with a situation for which there is no specialized algorithm, we are forced to use one of the general procedures. In this subsection, several general algorithms and their properties are described and references are indicated where more complete information can be found. The discussion will be in terms of the problem of computing a REML estimate of $\boldsymbol{\theta}$, i.e., the problem of computing a value of $\boldsymbol{\theta}$ that maximizes L_1. This will cause no real loss of generality, since the extensions to the problems of maximizing L with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$, maximizing L_1^* with respect to $\boldsymbol{\theta}$, and more generally maximizing an arbitrary function will be obvious.

The $(k+1)$st iterate of an iterative maximization algorithm has the representation $\tilde{\boldsymbol{\theta}}^{(k+1)} = \tilde{\boldsymbol{\theta}}^{(k)} + \rho_k \mathbf{w}_k$, where $\mathbf{w}_k$ is a vector that serves to identify the direction of search and ρ_k is a positive scalar that determines the distance to be traversed in the indicated search direction. Many of the proposed algorithms are gradient algorithms. A gradient algorithm, as applied to the maximization of L_1, is one where $\mathbf{w}_k$ has a convenient representation of the form $\mathbf{w}_k = \mathbf{N}_k\{\partial L_1/\partial\boldsymbol{\theta}\}^{(k)}$ for some $m \times m$ matrix $\mathbf{N}_k$. The various gradient methods are characterized by different choices for $\mathbf{N}_k$ and ρ_k. If $\mathbf{N}_k$ is chosen to be positive definite, then necessarily there exists a ρ_k $(\rho_k > 0)$ such that $L_1\{\tilde{\boldsymbol{\theta}}^{(k+1)}; \mathbf{y}\} > L_1\{\tilde{\boldsymbol{\theta}}^{(k)}; \mathbf{y}\}$, unless of course $\{\partial L_1/\partial\boldsymbol{\theta}\}^{(k)} = \mathbf{0}$. In fact, this inequality holds for positive definite $\mathbf{N}_k$ if ρ_k is taken to be sufficiently close to zero (see, e.g., Beltrami 1970). With regard to the choice of ρ_k, the methods fall into three categories: (i) $\rho_k = 1$; (ii) $\rho_k = \mu_k$ (to some degree of approximation), where μ_k is the value of the scalar ρ that maximizes $f_k(\rho) = L_1[\tilde{\boldsymbol{\theta}}^{(k)} + \rho\mathbf{w}_k; \mathbf{y}]$ for $\rho > 0$, i.e., ρ_k is chosen so as to maximize progress in the indicated search direction; and (iii) ρ_k is taken to be any positive value of ρ for which $f_k(\rho) > f_k(0)$, i.e., we merely require that the $(k+1)$st iteration produce some progress in the indicated search direction. In (iii), some effort (short of that required to approximate μ_k) may be expended to find a value of ρ for which $f_k(\rho)$ is "large." The determination of the ρ_k specified by (ii) or (iii) is a one-dimensional search problem. Suitable algorithms for one-dimensional searches are discussed by Murray (1972). They require at a minimum that $f_k(\rho)$ be evaluated at various trial values of ρ, and thus can be very time consuming in instances where the evaluation of L_1 involves extensive computations.

Two of the oldest and best known of the gradient algorithms are the steepest ascent algorithm and the Newton-Raphson algorithm. In the method of steepest ascent, $\mathbf{N}_k = \mathbf{I}$ for all k and customarily $\rho_k = \mu_k$. The steepest ascent algorithm is one of the few that is supported by convergence theorems (see, e.g., Beltrami (1970) and Powell (1970)). Unfortunately, its rate of convergence is often found to be intolerably slow (Powell 1970). Bard (1974, p. 88) states, "the method is not recommended for practical applications." Hartley and Vaughn (1972) describe, in the context of the ML estimation of variance components, a variation of the steepest ascent algorithm. Their approach requires that a system of c differential equations be solved numerically on each iteration. In the Newton-Raphson procedure, $\mathbf{N}_k = \{\mathbf{J}^{(k)}\}^{-1}$ and $\rho_k = 1$, where $\mathbf{J}$ is the $m \times m$ matrix whose ijth element is $-\partial^2 L_1/\partial\theta_i\partial\theta_j$. (It is assumed here that $\mathbf{J}^{(k)}$ is invertible.) Unlike the steepest ascent method, the Newton-Raphson procedure utilizes second-order partial derivatives. When applied to a quadratic function that has a negative definite Hessian matrix, the Newton-Raphson procedure will converge to the maximizing value in a single iteration from any starting point. Even when it is applied to a function like L_1 which is not quadratic, the Newton-Raphson algorithm can be expected to locate a maximizing value in relatively few iterations provided it is started within a sufficiently small neighborhood of that value (see, e.g., Bard 1974). However, if the starting value is poor, it may converge to a stationary point which is not a local or global maximum and often does not converge at all (Powell 1970). This difficulty is overcome in the extended or modified Newton-Raphson procedure. The extended procedure uses the same search direction as the original, but ρ_k is determined so that $L_1\{\tilde{\boldsymbol{\theta}}^{(k+1)}; \mathbf{y}\}$ is at least somewhat larger than $L_1\{\tilde{\boldsymbol{\theta}}^{(k)}; \mathbf{y}\}$. (If the directional derivative is negative in the direction $\{\mathbf{J}^{(k)}\}^{-1}\{\partial L_1/\partial\boldsymbol{\theta}\}^{(k)}$, the search direction $-\{\mathbf{J}^{(k)}\}^{-1}\{\partial L_1/\partial\boldsymbol{\theta}\}^{(k)}$, can be used instead.)

The method of scoring is a gradient procedure that applies when the function to be maximized depends on data points (observed values of random variables). It is identical to the Newton-Raphson procedure except that the role of the second-order partial derivatives is played instead by their expected values. As applied to the maximization of L_1, the $(k+1)$st iterate of the method of scoring is defined by putting $\mathbf{N}_k = \{\mathbf{B}^{(k)}\}^{-1}$ and $\rho_k = 1$. Note that $\mathbf{N}_k$ in this method coincides with

the large-sample covariance matrix of the REML estimator of θ evaluated at $\theta = \tilde{\theta}^{(k)}$, which illustrates a general property of the method when it is applied directly to the maximization of a likelihood function. The method of scoring as defined above can also be applied to the problem of maximizing "reduced" likelihood functions like L_1^*, L_2^*, or L_2. It is to be expected that this will produce iterates for the remaining parameters different from those produced by applying the method directly to the relevant likelihood function. The advantage of the method of scoring over the Newton-Raphson method is that, since the expected values of second-order partial derivatives are ordinarily easier to compute than the second-order partial derivatives themselves (refer to Section 5), it will generally require less computer time per iteration, though possibly at the expense of an increased number of iterations to convergence. In the case of the ML or REML estimation of variance components, this advantage may, however, be fairly insignificant (again refer to Section 5). The method of scoring can be extended or modified in the same way as the Newton-Raphson procedure by considering values for ρ_k different from one. Note, when the method of scoring is applied to the maximization of L_1, it defines a search direction in which at least some increase in L_1 can be achieved (provided $\{\partial L_1/\partial\theta\}^{(k)} \neq \mathbf{0}$) since ordinarily $\mathbf{B}^{(k)}$ will be positive definite (see Section 6.1). This again illustrates a general property of the method when applied directly to the maximization of a likelihood function. The $(k+1)$st iterate generated by applying the method of scoring to the maximization of L with respect to θ and by then substituting $\bar{\alpha}\{\tilde{\theta}^{(k)}\}$ for α is, in the case where $\mathbf{V}$ has the linear representation (2.3), the same as the $(k+1)$st iterate defined by T.W. Anderson's iterative ML algorithm. This observation was first made by J.N.K. Rao (see Miller 1973). Moreover, the iterates produced by the REML analog of Anderson's algorithm are identical to those defined by applying the method of scoring to the maximization of L_1 with respect to θ (Hocking and Kutner 1975). Thus this procedure can be viewed as a special case of the method of scoring.

The extended or modified Newton-Raphson procedure represents an attempt at retaining the good performance of the Newton-Raphson procedure when it is started close to a maximizing value while improving on its performance when it is started with a poor estimate. A similar philosophy underlies the gradient method described by Bard (1974, Sect. 5-8), which is based on the work of Levenberg (1944), Marquardt (1963), and Goldfeld, Quandt, and Trotter (1966). As applied to the maximization of L_1, the $(k+1)$st iterate of the latter method is obtained by taking $\rho_k = 1$ and $\mathbf{N}_k = [\mathbf{A}_k + \lambda_k\mathbf{M}_k]^{-1}$. Here, $\mathbf{M}_k$ is a positive definite matrix, λ_k is a scalar that ordinarily is taken to be positive, and $\mathbf{A}_k$ either equals $J^{(k)}$ (the negative of the Hessian matrix at $\theta = \tilde{\theta}^{(k)}$) or is some approximation to it. The matrix $\mathbf{A}_k + \lambda_k\mathbf{M}_k$ will be positive definite provided λ_k is taken to be sufficiently large (even if $\mathbf{A}_k$ is indefinite). When $\mathbf{A}_k = \mathbf{J}^{(k)}$ and $\mathbf{M}_k = \mathbf{I}$, the search direction employed in this method can be regarded as a compromise between the steepest ascent direction and the Newton-Raphson direction. Based on scaling considerations, a good choice for $\mathbf{M}_k$ is to take it to be the diagonal matrix whose diagonal elements are the absolute values of the diagonal elements of $\mathbf{A}_k$, except that zeros are replaced by ones (Bard 1974, Sect. 5-8). The scalar λ_k should be chosen so $L_1\{\tilde{\theta}^{(k+1)}; \mathbf{y}\} > L_1\{\tilde{\theta}^{(k)}; \mathbf{y}\}$. Algorithms for determining a suitable λ_k are discussed by Bard (1974); Goldfeld, Quandt, and Trotter (1966); Marquardt (1963); and Powell (1970). The reason for taking $\rho_k = 1$ in this approach is that the step size as well as the search direction are taken into account in choosing λ_k. Just as the computations per iteration can ordinarily be decreased by going from the Newton-Raphson algorithm to the method of scoring, we can expect the computations per iteration in the above method to be reduced by taking $\mathbf{A}_k = \mathbf{B}^{(k)}$ rather than $\mathbf{A}_k = \mathbf{J}^{(k)}$. When $\mathbf{A}_k = \mathbf{B}^{(k)}$, this method can be regarded as one natural extension of Marquardt's highly successful algorithm for solving nonlinear least-squares problems.

The search for improved optimization algorithms is an ongoing process. Recent progress was reviewed by Powell (1970) and Murray (1972). One relatively new class of methods consists of the gradient methods known as variable-metric methods. These methods use an $\mathbf{N}_k$ whose construction does not require second-order partial derivatives (or their expected values) but which nevertheless approximates $\mathbf{J}^{(k)}$ for sufficiently large k. A valuable feature in many instances is second-order partial derivatives need not be computed. Perhaps the best known of the variable-metric algorithms is the Davidon-Fletcher-Powell algorithm described by Powell (1970). It has been widely used and has been very successful.

6.3 Modifications to Accommodate Constraints on θ

Ordinarily, the space Ω to which θ is constrained has a representation of the form

$$\Omega = \{\theta : r_1(\theta)[>, \geq]0, \ldots, r_d(\theta)[>, \geq]0\} \quad , \tag{6.7}$$

for some functions $r_1, \ldots, r_d$. Here, $[>, \geq]$ is used to indicate that the inequality can be a strict inequality or not. For example, in our formulations of the ordinary ANOVA models, we can let $d = c + 1$ and $r_i(\theta) = \theta_i$ $(i = 1, \ldots, c+1)$, and take the last inequality in (6.7) to be a strict inequality.

As noted in Section 6.2, Henderson's iterative algorithm for computing ML estimates of variance components and its REML analog are not affected by the constraints on the parameter space; i.e., by the nonnegativity constraints on the variance components. With it, negative components are simply never encountered. Unfortunately, none of the gradient algorithms discussed in Section 6.3 have this kind of property. When they are applied, e.g., to the maximization of L, L_1^*, or L_1, any of them can in general produce an iterate that lies outside the constraint space. In particular, in the case of the ordinary mixed ANOVA models, they can produce an

iterate with negative values for one or more of the variance components. (This is also true of Anderson's iterative procedure—see Miller (1973)). Hemmerle and Hartley (1973) encountered this difficulty in applying the Newton-Raphson method to the problem of computing ML estimates for σ_{c+1}^2 and for the positive square roots of the ratios $\gamma_1, \ldots, \gamma_c$. When an iterate was obtained with negative or nearly negative values for one or more elements, they set those elements equal to zero and in effect constrained them to be zero on subsequent iterations. This approach to the problem is not satisfactory because it can cause the procedure to converge to a point that is not even a constrained local maximum of L_1^*, let alone a constrained global maximum, i.e., a ML estimate. (See the discussion by Bard 1974, Sect. 6-3.) The same criticism applies, though to a lesser extent, to the approach taken by Miller (1973) in using Anderson's procedure to compute ML estimates of variance components. He disregarded the nonnegativity constraints, unless the procedure converged to a vector having one or more negative components, in which case he restarted the algorithm with those components subsequently constrained to remain at zero. He continued to fix any zero components and to restart the algorithm until no negative values were obtained. Because iterates are permitted that can lie outside the parameter space, there is an additional difficulty with Miller's approach. The procedure may on occasion call for evaluating items that depend on $\boldsymbol{\theta}$ at points at which they are ill-conditioned or even undefined.

Satisfactory techniques for modifying unconstrained maximization algorithms so as to take into account inequality constraints are discussed by Bard (1974), Beltrami (1970), and Gill and Murray (1974). At least three of these techniques more or less meet our needs: (i) the penalty technique, (ii) the gradient projection technique, and (iii) the transformation technique.

There are actually many penalty techniques. Among these, the interior techniques, which cause each iterate to lie in the interior of the parameter space, are the best suited for the ML or REML estimation of $\boldsymbol{\theta}$. One interior technique is that proposed by Carroll (1961). In terms of the maximization of L_1, Carroll's technique is to apply the unconstrained maximization algorithm to the function

$$\psi(\boldsymbol{\theta}) = L_1(\boldsymbol{\theta}; \mathbf{y}) - \sum_{j=1}^{d} \phi_j / r_j(\boldsymbol{\theta}) \ ,$$

where $\phi_1, \ldots, \phi_d$ are small positive constants, rather than to L_1 itself. The algorithm is started in the interior of the constraint space and, at each iteration, the distance traversed in the indicated search direction is limited (if necessary) so that the resulting iterate is again interior to the constraint space. The underlying philosophy is that the function ψ is close to L_1 except in the neighborhood of boundaries where it assumes very large negative values, which serve as barriers that deflect the algorithm. Ordinarily, the maximization of ψ should be carried out for more than one set of values of $\phi_1, \ldots, \phi_d$. When convergence is obtained for one set, the values of $\phi_1, \ldots, \phi_d$ are reduced and the algorithm is applied to ψ again starting at the point of convergence in the previous application. The process is terminated when reductions in $\phi_1, \ldots, \phi_d$ no longer produce significant changes in the point of convergence.

The gradient projection technique can be used whenever all of the inequality constraints are linear constraints (as in computing ML or REML estimates of variance components), i.e., whenever $r_i(\boldsymbol{\theta}) = \mathbf{u}_i'\boldsymbol{\theta} - c_i$, for some $m \times 1$ vector $\mathbf{u}_i$ and some scalar c_i $(i = 1, \ldots, d)$. In conjunction with the gradient projection technique, we suppose that none of the inequality constraints are strict inequalities, i.e., the constraints are $\mathbf{u}_i'\boldsymbol{\theta} \geq c_i$ $(i = 1, \ldots, d)$, and $\mathbf{u}_i$ has been normalized so that $\mathbf{u}_i'\mathbf{u}_i = 1$. (A strict inequality $\mathbf{u}_i'\boldsymbol{\theta} > c_i$ can be approximated by the constraint $\mathbf{u}_i'\boldsymbol{\theta} \geq c_i + \epsilon_i$, where ϵ_i is some small positive constant.) Any of the unconstrained gradient algorithms can be modified by the gradient projection technique. At the completion of the kth iteration of the modified algorithm, we have $\mathbf{u}_i'\boldsymbol{\theta}^{(k)} > c_i$ for some values of i and $\mathbf{u}_i'\boldsymbol{\theta}^{(k)} = c_i$ for the remainder. On the $(k+1)$st iteration, certain of the latter constraints are treated as active constraints. The active constraints are selected in accordance with algorithms like those discussed by Gill and Murray (1974). Put $\mathbf{U}_k = (\mathbf{u}_{j(1)}, \ldots, \mathbf{u}_{j(\xi)})$ where constraints $j_{(1)}, \ldots, j_{(\xi)}$ are the active constraints, and denote by $\boldsymbol{\Lambda}_k$ the choice for $\mathbf{N}_k$ in the unconstrained gradient algorithm. The gradient projection technique is to use the gradient procedure which has

$$\mathbf{N}_k = [\mathbf{I} - \boldsymbol{\Lambda}_k\mathbf{U}_k(\mathbf{U}_k'\boldsymbol{\Lambda}_k\mathbf{U}_k)^{-1}\mathbf{U}_k']\boldsymbol{\Lambda}_k \ ,$$

and which restricts ρ_k so that no constraint is violated but which otherwise determines ρ_k as in the unconstrained case. The gradient projection technique thus modifies the the original search direction $\boldsymbol{\Lambda}_k\{\partial L_1/\partial\boldsymbol{\theta}\}^{(k)}$ by projecting it into the space determined by those vectors $\boldsymbol{\theta}$ satisfying $\mathbf{U}_k'\boldsymbol{\theta} = \mathbf{0}$. This technique is considered to be superior to the penalty technique for handling linear constraints, especially when it is suspected that the maximizing value may be located on a boundary.

Sometimes a constrained maximization problem can be transformed into an unconstrained maximization problem by a change of variables (Bard 1974). For example, the ordinary ANOVA models can be parameterized so that $\mathbf{R} = \theta_m^2\mathbf{I}$, $\mathbf{D} = \text{diag}\,[\theta_1^2\mathbf{I}, \ldots, \theta_{m-1}^2\mathbf{I}]$, and $\Omega = \{\boldsymbol{\theta} : \theta_m \neq 0\}$. Here, $\sigma_1^2 = \theta_1^2, \ldots, \sigma_m^2 = \theta_m^2$. To obtain a ML or REML estimate of $\sigma_1^2, \ldots, \sigma_m^2$, we can now maximize L_1^* or L_1 with respect to $\boldsymbol{\theta}$, subject only to the constraint $\theta_m \neq 0$, and then transform the maximizing vector by squaring its elements. This maximization problem is, for all practical purposes, unconstrained because the constraint, $\theta_m \neq 0$, ordinarily never comes into play. This kind of approach was used by Hartley and Vaughn (1972). One possible drawback in using this technique, to compute REML estimates of variance components for example, is that additional stationary points of L_1 are introduced, i.e., for $\theta_i = 0$, $\partial L_1/\partial\theta_i = 0$ even though $\partial L_1/\partial\sigma_i^2 \neq 0$. Thus, we should use this technique only in conjunction with

algorithms that guarantee at least some increase in the value of the objective function on each iteration.

6.4 Discussion

In a given application, it may be possible to improve the performance of the various iterative optimization algorithms by first transforming the variables. Most of these algorithms are at their best when applied to functions that are at least approximately quadratic. Thus any transformation that makes the function more closely resemble a quadratic function over the relevant region should be helpful. In particular, in using the Newton-Raphson algorithm to compute ML estimates for the ordinary ANOVA models, Hemmerle and Hartley (1973) found that the behavior of the algorithm could be improved significantly by parameterizing in terms of $\sqrt{\gamma_1}, \ldots, \sqrt{\gamma_c}, \gamma_{c+1}$ rather than in terms of $\gamma_1, \ldots, \gamma_{c+1}$.

In general, it will be more efficient to compute ML estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ by applying the various iterative optimization algorithms to the "reduced" function L_1^* rather than to L itself. Similarly, when the problem of maximizing L_1^* or L_1 can be reduced in dimension by analytical means to the problem of maximizing L_2^* or L_2 (refer to Sections 4.1 and 4.3), it is to be expected that it will be more efficient to compute ML or REML estimates of $\boldsymbol{\theta}$ by applying the iterative algorithms to the latter functions. Analytical reductions have proved to be useful in nonlinear least-squares problems (see, e.g., Lawton and Sylvestre 1971).

There is ordinarily no assurance that a value of $\boldsymbol{\theta}$ obtained by applying an iterative maximization algorithm to L, L_1^*, or L_2^* or to L_1 or L_2 is a ML or REML estimate of $\boldsymbol{\theta}$. Even if such a $\boldsymbol{\theta}$ value is obtained by starting the algorithm with what is thought to be an excellent guess or estimate, it is good practice to apply the algorithm several more times, using a different starting point on each occasion. If these repetitions all yield the same $\boldsymbol{\theta}$ value, we can be more confident that we have located a ML or REML estimate.

Actual numerical experience in using the various iterative algorithms to compute ML or REML estimates of variance components seems to be very limited and is largely confined to a variation of the method of steepest ascent (Hartley and Vaughn 1972), the Newton-Raphson procedure (Hemmerle and Hartley 1973; Corbeil and Searle 1976a; Jennrich and Sampson 1976), the method of scoring (Jennrich and Sampson 1976), and to Anderson's method (Miller 1973).

7. APPROXIMATING THE RESTRICTED MAXIMUM LIKELIHOOD APPROACH

Efficient computational algorithms for producing ML or REML estimates of $\boldsymbol{\theta}$ can be devised by making use of the results outlined in Sections 3 through 6. As the speed of electronic computers increases, the number of settings feasible to compute ML or REML estimates also increases. Nevertheless, there remain numerous situations where their computation is unthinkable. The latter situations are essentially those where the computations necessary to form and solve the linear system (3.3) are too extensive. In this section, we outline an approach to the estimation of $\boldsymbol{\theta}$ that can be viewed as an approximation to the REML approach. This approximate approach can be used when the computation of the exact ML or REML estimate is too demanding.

In cases where the function L_1 is known to have a unique stationary point which lies in the constraint space Ω and which corresponds to a global maximum, the problem of computing a REML estimate of $\boldsymbol{\theta}$ is essentially that of forming the system of nonlinear equations $\partial L_1/\partial\boldsymbol{\theta} = \mathbf{0}$ and solving it for $\boldsymbol{\theta}$. If the computations required to evaluate $\partial L_1/\partial\boldsymbol{\theta}$ are too extensive, REML estimation cannot be undertaken. The equations $\partial L_1/\partial\boldsymbol{\theta} = \mathbf{0}$ consist in effect of m translation-invariant quadratic forms set equal to their expectations. This observation suggests when the REML approach is unfeasible computationally, we take our estimate of $\boldsymbol{\theta}$ to be the solution $\tilde{\boldsymbol{\theta}}$ to $\mathbf{G}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{0}$, where $\mathbf{G}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{Q}(\boldsymbol{\theta}; \mathbf{y}) - E(\mathbf{Q})$, and where $\mathbf{Q}(\boldsymbol{\theta}; \mathbf{y})$ is a vector whose elements $Q_i = \mathbf{y}'\boldsymbol{\Gamma}_i\mathbf{y}$ $(i = 1, \ldots, m)$ consist of m translation-invariant quadratic forms which resemble those used in REML estimation but which are easier to evaluate. The quadratic forms used in REML are

$$(\tfrac{1}{2})(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})'\mathbf{V}^{-1}(\partial\mathbf{V}/\partial\theta_i)\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})$$
$$= (\tfrac{1}{2})(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}} - \mathbf{Z}\tilde{\boldsymbol{\beta}})'\mathbf{R}^{-1}(\partial\mathbf{V}/\partial\theta_i)\mathbf{R}^{-1} \cdot (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}} - \mathbf{Z}\tilde{\boldsymbol{\beta}}) \quad (7.1)$$

$$= (\tfrac{1}{2})(\mathbf{y} - \mathbf{Z}\tilde{\boldsymbol{\beta}})'\mathbf{S}(\partial\mathbf{V}/\partial\theta_i)\mathbf{S}(\mathbf{y} - \mathbf{Z}\tilde{\boldsymbol{\beta}}) \ , \quad (7.2)$$

$i = 1, \ldots, m$. For i such that $\mathbf{D}$ depends on θ_i but $\mathbf{R}$ and $\mathbf{Z}$ do not, these quadratic forms have the additional representations

$$-(\tfrac{1}{2})\tilde{\boldsymbol{\beta}}'(\partial\mathbf{D}^{-1}/\partial\theta_i)\tilde{\boldsymbol{\beta}} \ , \quad (7.3)$$

provided $\boldsymbol{\theta}$ is such that $\mathbf{D}$ is nonsingular. One technique for "approximating" the quadratic forms used in REML is to replace $\tilde{\boldsymbol{\alpha}}$ and/or $\tilde{\boldsymbol{\beta}}$ in (7.1), (7.2), or (7.3) with $\tilde{\boldsymbol{\alpha}} = \mathbf{H}\mathbf{y}$ and $\tilde{\boldsymbol{\beta}} = \mathbf{A}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})$, where $\mathbf{H}$ is a $p \times n$ matrix such that $E(\mathbf{X}\tilde{\boldsymbol{\alpha}}) = \mathbf{X}\boldsymbol{\alpha}$ and $\mathbf{A}$ is a $q \times n$ matrix, both of which must be specified. The elements of $\mathbf{H}$ and $\mathbf{A}$ may be functions of $\boldsymbol{\theta}$. The matrices $\mathbf{H}$ and $\mathbf{A}$ should be chosen so that, for the case where $\boldsymbol{\theta}^+$ is known, $\mathbf{X}\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ with $\boldsymbol{\theta} = \boldsymbol{\theta}^+$ are good estimators of $\mathbf{X}\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, but, at the same time, they must be such that $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ are computable for any given $\boldsymbol{\theta}$ value. In cases where $\mathbf{R}$ is hard to invert, we could replace $\mathbf{R}^{-1}$ in (7.1) or (7.2) with some "approximation," as well as substituting $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$.

The expression $[E(\mathbf{K})]^{-1}[\text{var}\ (\mathbf{Q})][E(\mathbf{K}')]^{-1}$, where $\mathbf{K}(\boldsymbol{\theta}; \mathbf{y})$ is the $m \times m$ matrix whose jth column is $-\partial\mathbf{G}/\partial\theta_j$, may furnish a useful approximation to var $(\tilde{\boldsymbol{\theta}})$ for "large" samples (Harville 1975). When $Q_1, \ldots, Q_m$ are the quadratic forms actually used in REML rather than approximations, this expression simplifies to $\mathbf{B}^{-1}$.

There are several hazards in estimating $\boldsymbol{\theta}$ by the approximate REML approach described above, i.e., by solving the equations $\mathbf{G}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{0}$, where Q_i is given by (7.1),

(7.2), or (7.3) with $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ substituted for $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\beta}}$. In REML, the likelihood equations may not have a solution that lies in the constraint space Ω, and, even if there are solutions in Ω, some or all of them may not correspond to maximizing values of L_1, i.e., to REML estimates. Similarly, in the approximate REML approach, there may not exist a solution to $\mathbf{G}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{0}$ that lies in Ω and, even if such a solution does exist, it may not be a desirable estimate. In implementing the REML approach, we were able to circumvent these difficulties, at least to some extent, by using "hill-climbing" techniques, which force increases in L_1 at each iteration, preventing convergence to undesirable stationary points, and which can be modified to accommodate constraints. This observation points the way to what may be a useful modification of the approximate REML approach. Instead of merely solving the equations $\mathbf{G}(\boldsymbol{\theta}; \mathbf{y}) = \mathbf{0}$, we could proceed just as though we were maximizing a function whose gradient vector is $G(\boldsymbol{\theta}; \mathbf{y})$. We could use various of the gradient algorithms described in Section 6.2 to maximize this function, with appropriate modification for constraints as described in Section 6.3. (See Harville 1975 for further information.) The final iterate would comprise our estimate of $\boldsymbol{\theta}$.

In applying various of the gradient algorithms in our approximate REML approach, we must, on each iteration, evaluate Q_i, $E(Q_i)$, $\partial Q_i/\partial\theta_j$, $\partial[E(Q_i)]/\partial\theta_j$, and/or $E(\partial Q_i/\partial\theta_j)$, $(i, j = 1, \ldots, m)$. We have

$$E(Q_i) = \text{tr}\,(\boldsymbol{\Gamma}_i\mathbf{V}) = \text{tr}\,(\mathbf{R}^{\frac{1}{2}}\boldsymbol{\Gamma}_i\mathbf{R}^{\frac{1}{2}}) + \text{tr}\,(\mathbf{D}^{\frac{1}{2}}\mathbf{Z}'\boldsymbol{\Gamma}_i\mathbf{Z}\mathbf{D}^{\frac{1}{2}}) \ ,$$

so that $E(Q_i)$ can be evaluated by performing the same operations on each column of $\mathbf{R}^{\frac{1}{2}}$ and each column of $\mathbf{Z}\mathbf{D}^{\frac{1}{2}}$ as are performed on $\mathbf{y}$ in computing Q_i. This technique is essentially Hartley's method of synthesis (see Rao 1968). The computation of the other required items can be approached in a similar manner.

8. RELATIONSHIPS OF MAXIMUM LIKELIHOOD AND RESTRICTED MAXIMUM LIKELIHOOD TO OTHER METHODS

8.1 MIVQUE's and MINQUE's

Much of the recent literature on the problem of estimating variance components, and more generally on the problem of estimating $\boldsymbol{\theta}$ when $\mathbf{V}$ has the representation (2.3), has centered on the derivation of estimators that have minimum MSE at some point in the parameter space, i.e., that are locally best when attention is restricted to estimators satisfying various conditions. The initial work was done by Townsend (1968) and by Townsend and Searle (1971). They derived exact expressions for the locally best quadratic unbiased estimators of the two variance components associated with the unbalanced one-way random ANOVA model, under the assumptions that $\mathbf{y}$ is normal and the mean vector is $\mathbf{0}$. Harville (1969a) considered the same setting but dropped the assumption that the mean vector is null. Harville gave some results on estimators that are locally best in the class of quadratic unbiased estimators and in the class of translation-invariant quadratic unbiased estimators, though his results were left in very inconvenient form. These early efforts were generalized and greatly improved upon by LaMotte (1970, 1971, and 1973). LaMotte's results apply to all linear models for which $\mathbf{V}$ has the representation (2.3), though he did assume normality. He considered several classes of estimators for a linear function $\boldsymbol{\lambda}'\boldsymbol{\theta}$, and, for each class, produced convenient representations for the locally best estimators. In particular, he showed that, when attention is restricted to translation-invariant quadratic unbiased estimators, the estimator that is locally best at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is $\boldsymbol{\lambda}'\hat{\boldsymbol{\theta}}$ where $\hat{\boldsymbol{\theta}}$ is any solution to the linear system

$$[\mathbf{B}(\boldsymbol{\theta}^*)]\hat{\boldsymbol{\theta}} = [\mathbf{d}(\boldsymbol{\theta}^*)] \qquad (8.1)$$

(provided that $\boldsymbol{\lambda}'\boldsymbol{\theta}$ is estimable in the class of translation-invariant quadratic estimators, which is the case if and only if the equations $[\mathbf{B}(\boldsymbol{\theta}^*)]\boldsymbol{\tau} = \boldsymbol{\lambda}$ have a solution for $\boldsymbol{\tau}$). Rao (1971b and 1972) independently obtained similar results and, in addition, indicated extensions to non-normal cases. Following Rao, we use MIVQUE as an abbreviation for locally best (minimum variance) translation-invariant quadratic unbiased estimator. (The E in MIVQUE can also stand for estimation.)

In general, quadratic unbiased estimators of $\boldsymbol{\lambda}'\boldsymbol{\theta}$ (including MIVQUE's) can yield estimates that violate the constraints on the parameter space, so that strictly speaking they are not estimators at all. Nevertheless, as observed by Kempthorne (Searle 1968, p. 783), they can be regarded as useful condensations of the data, just as true estimators are. What is questionable is the practice of comparing these pseudo-estimators on the basis of their MSE's. For reasons discussed by Harville (1969b, Sect. 3.3.5), such comparisons are potentially misleading.

The traditional approach to the estimation of $\boldsymbol{\theta}$, when $\mathbf{V}$ has the representation (2.3) as in the case of the ordinary ANOVA models, is to equate m translation-invariant quadratic forms (that are not functionally dependent on $\boldsymbol{\theta}$) to their expectations and solve the resulting linear system for $\boldsymbol{\theta}$. The ith of the likelihood equations $\partial L/\partial\boldsymbol{\theta} = \mathbf{0}$ depends on the data only through the quadratic form $(\frac{1}{2})(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})'\mathbf{V}^{-1}\mathbf{G}_i\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\alpha})$. Suppose that, in this quadratic form, we substitute $\bar{\boldsymbol{\alpha}}(\boldsymbol{\theta})$ for $\boldsymbol{\alpha}$ and then replace $\boldsymbol{\theta}$ with a fixed value $\boldsymbol{\theta}^*$. The result is a translation-invariant quadratic form that is functionally independent of $\boldsymbol{\theta}$. LaMotte (1970) considered estimating $\boldsymbol{\theta}$ by equating the m translation-invariant quadratic forms generated in this way to their expectations. He found that the resulting linear system is the same as the linear system (8.1) associated with MIVQUE. Thus in the case of assumed normality, this approach is completely equivalent to the MIVQUE approach.

Rao (1970, 1971a, and 1972) proposed an intuitive estimation procedure that can be used in particular to estimate linear functions of the variance components associated with the ordinary ANOVA models. He observed in effect that, if $\boldsymbol{\beta}$ and $\boldsymbol{\varepsilon}$ (the realized or sample value of $\mathbf{e}$)

were known, a natural estimator for

$$\lambda'\theta = \sum_{i=1}^{c+1} \lambda_i \sigma_i^2$$

would be

$$(\lambda_{c+1}/n)\varepsilon'\varepsilon + \sum_{i=1}^{c} (\lambda_i/q_i)\beta_i'\beta_i = \omega'\Delta\omega \ , \qquad (8.2)$$

where $\omega' = (\beta', \varepsilon')$ and where Δ is a suitably defined matrix. Since β and ε are in fact unknown, Rao suggested estimating $\sum_i \lambda_i\sigma_i^2$ by the translation-invariant quadratic unbiased estimator that most closely resembles (8.2). More precisely, observing that $y'\Gamma y = \omega'U'\Gamma U\omega$, with $U = (Z, I)$, for any translation-invariant quadratic estimator $y'\Gamma y$, he proposed the estimator $y'\Gamma^* y$, where Γ^* minimizes $\|U'\Gamma U - \Delta\|$ for Γ such that $y'\Gamma y$ is a translation-invariant quadratic unbiased estimator of $\sum_i \lambda_i\sigma_i^2$. Here, $\|\cdot\|$ denotes a matrix norm. It can be shown that, when the Euclidean norm is used, $y'\Gamma^* y = \lambda'\hat{\theta}$, where, with $\theta^* = 1$, $\hat{\theta}$ is a solution to the linear system (8.1). Rao went on to observe that the difference between a translation-invariant quadratic estimator $y'\Gamma y$ and $\omega'\Delta\omega$ can be expressed as $\eta'\Lambda(U'\Gamma U - \Delta)\Lambda\eta$, where $\Lambda = \text{diag}\,(\sigma_1 I, \ldots, \sigma_{c+1} I)$ and η represents the standardized vector $\Lambda^{-1}\omega$. Taking Λ^* to be the value of Λ at $\theta = \theta^*$, where the value of θ^* can be based on prior information, we could also consider estimating $\sum_i \lambda_i\sigma_i^2$ by $y'\Gamma^* y$, where now Γ^* minimizes $\|\Lambda(U'\Gamma U - \Delta)\Lambda\|$ for translation-invariant quadratic unbiased estimators $y'\Gamma y$. Again, when the Euclidean norm is employed, it can be shown that $y'\Gamma^* y = \lambda'\hat{\theta}$ where $\hat{\theta}$ is any solution to the linear system (8.1). Rao called these estimators MINQUE's (minimum norm quadratic unbiased estimators). It is clear that a MINQUE of $\sum_i \lambda_i\sigma_i^2$ (based on a Euclidean norm) is the same as a MIVQUE (derived on the basis of the normality assumption).

Several observers (Harville 1969b; LaMotte 1970; and Rao 1972) have suggested an iterative MIVQUE procedure. The iterates could be defined in terms of the linear system (8.1). If the procedure converges to some point in the parameter space, that point is necessarily a stationary point of L_1 (see Section 6.1). Thus, if we disregard any complications that might be caused by constraints on the parameter space or by nonconvergence or convergence to a point that does not correspond to a maximum of L_1, then iterative MIVQUE is identical to REML. A similar observation was made by Hocking and Kutner (1975). Note that the iterates produced by the iterative MIVQUE procedure are the same as those defined by the REML analog of Anderson's iterative algorithm (again refer to Section 6.1), implying in particular that the initial iterate of the REML version of Anderson's procedure is a MIVQUE.

Suppose that, assuming normality, there exists a UMIVQUE of $\lambda'\theta$, i.e., an estimator which, among all translation-invariant quadratic unbiased estimators of $\lambda'\theta$, has uniformly (for all $\theta \in \Omega$) minimum variance. Then, every MIVQUE of $\lambda'\theta$ is a UMIVQUE, implying that $\lambda'B^{-}d$ is functionally independent of θ. Taking $\hat{\theta}$ to be any REML estimate of θ, it follows that $\lambda'\hat{\theta}$ agrees with the UMIVQUE of $\lambda'\theta$, provided that $\hat{\theta}$ satisfies $\partial L_1/\partial\theta = 0$, as would necessarily be the case if $\hat{\theta}$ were an interior point of Ω. Moreover, if there is a UMIVQUE of every component of θ, then $B^{-1}d$ is functionally independent of θ, so that there is a $\theta \in \Omega$ satisfying the REML equations $\partial L_1/\partial\theta = 0$ if and only if $B^{-1}d \in \Omega$, in which case the REML equations admit the explicit solution $\theta = B^{-1}d$, and the REML version of Anderson's iterative procedure converges (to the UMIVQUE of θ) in a single iteration. (See Gautschi (1959) and Graybill and Hultquist (1961) for some discussion of the existence of uniformly minimum variance quadratic unbiased estimators of variance components.)

Rao's MINQUE approach does not require any normality assumptions, nor is its intuitive appeal diminished by nonnormality. The fact that MIVQUE's derived on the basis of normality turn out to be MINQUE's in important instances may, because of the relationships between MIVQUE and REML noted above, indicate that ML or REML estimators of θ derived under normality assumptions are reasonable estimators even when the form of the distribution of b and e is unspecified.

8.2 ANOVA-Like Methods

The most commonly used methods for estimating variance components are the Methods 1, 2, and 3 set forth by Henderson (1953). In these methods, mean squares associated with various ANOVA tables are set equal to their expectations, and estimates are obtained by solving the resulting linear equations. (In Method 2, the data vector is corrected for fixed effects before forming the ANOVA table.) Searle (1968, 1971a, and 1971b) gave excellent descriptions of Henderson's methods and indicated various generalizations. Henderson's methods yield translation-invariant quadratic unbiased estimators. In balanced-data cases, these estimators coincide with the normality-derived REML estimators, provided the nonnegativity constraints on the variance components do not come into play (Patterson and Thompson 1974). In general, however, the only parallel between Henderson's methods and REML would seem to be that both are based on equating translation-invariant quadratic forms to their expectations. In REML, the quadratic forms are functions of the variance components, the expectations are nonlinear, and modifications are incorporated to account for the nonnegativity constraints; while, in Henderson's methods, the quadratic forms are functionally independent of the variance components, the expectations are linear, and negative estimates of variance components can be realized. Cunningham and Henderson (1968) proposed a modified version (subsequently corrected by R. Thompson (1969)) of Henderson's Method 3 which seems more akin to REML. It uses equations of the form (3.8) in place of the normal equations ordinarily used in Method 3 to form reductions in sums of squares, with the consequence that the quadratic forms are no longer free of the variance components and an iterative process is necessary.

For those ANOVA models that can be parameterized so that $\mathbf{V}$ has the form (2.3) for some mutually orthogonal idempotent matrices $\mathbf{G}_1, \ldots, \mathbf{G}_{c+1}$, Nelder (1968) proposed an iterative ANOVA-like method for estimating variance components that is essentially equivalent to REML (see Patterson and Thompson 1974).

One problem with Henderson's methods for estimating variance and covariance components is that the methods are not necessarily well defined. That is, it is not always clear which mean squares from what ANOVA tables should be used (see Searle 1971a or 1971b). How these methods should be extended to the general problem of estimating $\boldsymbol{\theta}$ is even less clear. In contrast, ML and REML estimators are always well defined (at least conceptually). Moreover, except for balanced-data cases, little is known about the goodness of the Henderson estimators, other than they are unbiased and translation invariant. It is well-known that, at least in particular cases, there are biased estimators that have uniformly (assuming normality) smaller MSE's than the Henderson estimators (see Klotz, Milton, and Zacks 1969). What is more surprising is the recent discovery by Seely (1975) and by Olsen, Seely, and Birkes (1976) that, at least in the case of most unbalanced mixed- or random-effects models having one random factor (i.e., $c = 1$), there are translation-invariant quadratic unbiased estimators of σ_1^2 that have uniformly smaller variance than the Henderson Method-3 estimator. In contrast, MIVQUE estimators, which (as noted in Section 8.1) are closely related to REML estimators, are admissible in the class of translation-invariant quadratic unbiased estimators. Moreover, Olsen, Seely, and Birkes (1976) constructed, for a particular case, a MIVQUE estimator that is uniformly better than the corresponding Henderson Method-3 estimator. These revelations would seem to constitute a strong argument for using REML in preference to Henderson's methods when REML is feasible computationally and possibly for using an approximate REML approach similar to the one outlined in Section 7 when it is not.

8.3 Bayesian Methods

A review of some pre-1970 Bayesian results on inference for variance components was given by Harville (1969b). When loss is proportional to squared error, the estimator of a variance component (or of any other parameter) that minimizes Bayes risk is the parameter's posterior mean. However, in all but fairly simple cases, the computation of the posterior mean of a variance component or, more generally, of $\boldsymbol{\theta}$ is found to be unfeasible even when numerical integration techniques are used. Moreover, if an improper prior is employed in place of the "true" prior, the posterior mean may, because of its sensitivity to the tails of the posterior density, represent a rather unsatisfactory condensation of the data. Because of these difficulties with the posterior mean, posterior modes are sometimes proposed as estimators. We can use either the mode of the marginal posterior density of a parameter or the relevant component of the mode of the joint posterior density of that parameter and various other parameters. It would seem preferable to use the posterior density that has the maximum possible number of "nuisance" parameters integrated out (at least among those that have proper priors). A posterior mode can be computed numerically by techniques like those outlined in Sections 6.2 and 6.3. Moreover, a posterior mode is insensitive to the tails of the posterior density.

Suppose that $\mathbf{y}$ has the representation (2.1). If we wish to analyze the data by Bayesian techniques, we need to specify a prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$. Lindley and Smith (1972) suggested in effect that, in many cases, it is possible to redefine the terms of the model (2.1) so as to arrive at a second model of the form (2.1) in which it is reasonable, a priori, to take the components of $\boldsymbol{\alpha}$ to be independently and uniformly distributed over the real line and to be independent of $\boldsymbol{\theta}$, even though this assumption might not be realistic for the original model. The Lindley-Smith technique amounts to expressing various of the fixed effects in the original model as deviations from hyperparameters, expressing the hyperparameters as deviations from hyper-hyperparameters or "second-order" hyperparameters, expressing second-order hyperparameters as deviations from third-order hyperparameters, etc. In the redefined model, the highest-order hyperparameters comprise the components of $\boldsymbol{\alpha}$; the components of $\boldsymbol{\beta}$ include the deviations of various orders or, possibly, appropriate linear combinations of those deviations, together with the components of the original $\boldsymbol{\beta}$ vector; and additional "parameters" may be inserted into the original $\boldsymbol{\theta}$ vector to accommodate the new entries in the vector $\boldsymbol{\beta}$. Of course, in the new model, some components of $\mathbf{b}$ are random variables only in a subjective sense, but this is not objectionable if a Bayesian analysis is anticipated.

Lindley and Smith proposed, as an estimate for $\boldsymbol{\theta}$, the $\boldsymbol{\theta}$ component of the mode of the joint posterior density of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$. They acknowledged that this estimate may be unsatisfactory if vague priors are assumed for certain components of $\boldsymbol{\theta}$. In fact, their approach can lead to estimators of variance components that are identically equal to zero when used with vague priors. Lacking evidence to the contrary, it must be assumed that their approach can also lead to unsatisfactory estimators when used with informative priors, though they may be less obviously unsatisfactory. The problem with their approach may stem from the severe "dependencies" that undoubtedly exist between components of $\boldsymbol{\theta}$ and components of $\boldsymbol{\beta}$ in the joint posterior density of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$, which may lead to the $\boldsymbol{\theta}$ component of the mode of the joint posterior density being far removed from, say, $E(\boldsymbol{\theta}|\mathbf{y})$.

A seemingly superior approach would be to take the estimate of $\boldsymbol{\theta}$ to be the $\boldsymbol{\theta}$ component of the mode of the marginal posterior density of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ or, better yet, the mode of the marginal posterior density of $\boldsymbol{\theta}$. Suppose the distribution of $\mathbf{b}$, $\mathbf{e}$ is multivariate normal, $p^* = p$, and

a priori the components of $\boldsymbol{\alpha}$ are independently and uniformly distributed over the real line and are independent of $\boldsymbol{\theta}$ so that the joint prior density of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ is proportional to $h(\boldsymbol{\theta})$ for some function h. For purposes of estimating $\boldsymbol{\theta}$ alone, it can, for reasons noted by Harville (1974), make sense to adopt such a prior density even if the model is such that prior information on $\boldsymbol{\alpha}$ is actually available. It follows from Harville (1974) that the marginal posterior density of $\boldsymbol{\theta}$ (the density obtained by formally integrating out $\boldsymbol{\alpha}$) is proportional to the product of $h(\boldsymbol{\theta})$ and the likelihood function of an arbitrary set of $(n - p^*)$ linearly independent error contrasts. For $h(\boldsymbol{\theta}) \equiv 1$, the $\boldsymbol{\theta}$ component of the mode of the marginal posterior density of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ is simply the ML estimate, and the mode of the marginal posterior density of $\boldsymbol{\theta}$ is the REML estimate. For $h(\boldsymbol{\theta}) = [\det \{\mathbf{B}(\boldsymbol{\theta})\}]^{\frac{1}{2}}$, which is the Jeffreys' prior for $\boldsymbol{\theta}$ derived from L_1 alone, we have, in the case of the ordinary fixed ANOVA or regression models as defined by (4.3), that the $\boldsymbol{\theta}$ component of the mode of the marginal posterior density of $\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ is the estimator $[1/(n + 2)](\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})$, and the mode of the marginal posterior density of $\boldsymbol{\theta}$ is the estimator $[1/(n - p^* + 2)](\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})$. The latter estimator has a downward bias of "only" $2\theta_1/(n - p^* + 2)$ and has uniformly smaller MSE than both the ML and REML estimators of θ_1 (except in the case $p^* = 2$, where it coincides with the ML estimator) and in fact has uniformly smaller MSE than any other estimator of the form $(1/k)(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\alpha}})$, so that it may have appeal for frequentists who care about MSE but not about small biases. It is an intriguing possibility that the pseudo-Bayesian procedure that estimates $\boldsymbol{\theta}$ by maximizing

$$L_1(\boldsymbol{\theta}; \mathbf{y}) + (\tfrac{1}{2}) \log [\det \{\mathbf{B}(\boldsymbol{\theta})\}] , \qquad (8.3)$$

for $\boldsymbol{\theta} \in \Omega$, might be an equally satisfying procedure in more complicated settings.

9. FURTHER RESEARCH

There are still many aspects of the problem of estimating variance components, and more generally the problem of estimating $\boldsymbol{\theta}$, that need to be investigated. In this, the final section, an attempt is made to identify some of these areas.

The "realistic" asymptotic results developed by Miller (1973) for ML estimators of variance components for the ordinary ANOVA models need to be extended to various other models of the form (2.1) and to REML estimators. The results of Weiss (1971 and 1973) should prove useful here. Also, for particular models, such as the ordinary ANOVA models, it would be nice to know what parameterizations produce the "fastest convergence" to asymptotic normality.

In Sections 3 and 5, results were described which can be used to exploit structure in the $\mathbf{R}$, $\mathbf{D}$, $\mathbf{Z}$, and $\mathbf{X}$ matrices for purposes of computing L, L_1^*, or L_1, their first- and second-order partial derivatives, and expected values of their second-order derivatives. Some explicit simplifications were given for the ordinary ANOVA models. It may be worthwhile to work out detailed procedures for other commonly used models. Thompson (1973) did essentially this for MANOVA models.

While it is unlikely that any one of the iterative procedures for computing ML or REML estimates of $\boldsymbol{\theta}$ will be best or even satisfactory in every instance, useful guidelines for choosing a procedure may be possible for particular classes of models such as ANOVA models. Also, it would be nice to know how the various models should be parameterized in order to effect convergence in the fewest possible number of iterations. Analytical results like those discussed in Section 5 can be very useful in deciding on an algorithm, but well-planned numerical studies, like those carried out by Bard (1970) for nonlinear least-squares problems, will ultimately be needed. If Henderson's iterative algorithm for computing ML estimates of variance components and its REML analog are demonstrated to be superior computational procedures, it would be worthwhile to attempt extensions, e.g., to the problem of computing ML or REML estimates of covariance components.

The approximate REML scheme outlined in Section 7 needs to be further developed and evaluated. A good start would be to determine, for particular models such as ANOVA models, good choices for $\tilde{\boldsymbol{\alpha}}$ and $\tilde{\boldsymbol{\beta}}$ or, equivalently, for $\mathbf{H}$ and $\mathbf{A}$. It would be nice to know how the approximate REML scheme compares with, say, Henderson's Methods 1 and 2 as a procedure for estimating variance and covariance components.

The pseudo-Bayesian procedure that estimates $\boldsymbol{\theta}$ by maximizing the expression (8.3) would seem to be worth investigating. This might be done first for balanced ANOVA models. If the procedure looks good there, its performance in more complicated settings could be evaluated.

In cases where the estimate of $\boldsymbol{\theta}$ is not of interest in itself, but rather is to be used indirectly in estimating $\boldsymbol{\lambda}_1'\boldsymbol{\alpha} + \boldsymbol{\lambda}_2'\boldsymbol{\beta}$ (by substituting it for $\boldsymbol{\theta}^+$ in (3.2)), the possible estimators of $\boldsymbol{\theta}$ should be compared on the basis of a loss function reflecting those intentions. Efron and Morris (1976) made some comparisons of this kind for a particular case. Further work along these lines is needed.

[*Received December 1976.*]

REFERENCES

Anderson, R.L., and Bancroft, T.A. (1952), *Statistical Theory in Research*, New York: McGraw-Hill Book Co.

Anderson, T.W. (1969), "Statistical Inference for Covariance Matrices with Linear Structure," in *Multivariate Analysis-II*, ed. P.R. Krishnaiah, New York: Academic Press, 55–66.

——— (1970), "Estimation of Covariance Matrices Which Are Linear Combinations or Whose Inverses Are Linear Combinations of Given Matrices," in *Essays in Probability and Statistics*, eds. R.C. Bose, I.M. Chakravarti, P.C. Mahalanobis, C.R. Rao, and K.J.C. Smith, Chapel Hill, North Carolina: University of North Carolina Press, 1–24.

——— (1971), "Estimation of Covariance Matrices with Linear Structure and Moving Average Processes of Finite Order," Technical Report No. 6, Department of Statistics, Stanford University, Stanford, California.

——— (1973), "Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure," *Annals of Statistics*, 1, 135–41.
Bard, Yonathan (1970), "Comparison of Gradient Methods for the Solution of Nonlinear Parameter Estimation Problems," *SIAM Journal on Numerical Analysis*, 7, 157–86.
——— (1974), *Nonlinear Parameter Estimation*, New York: Academic Press.
Beltrami, Edward J. (1970), *An Algorithmic Approach to Nonlinear Analysis and Optimization*, New York: Academic Press.
Box, George E.P., and Jenkins, Gwilym M. (1970), *Time Series Analysis*, San Francisco: Holden-Day.
Carroll, C.W. (1961), "The Created Response Surface Technique for Optimizing Nonlinear, Restrained Systems," *Operations Research*, 9, 169–84.
Corbeil, Robert R., and Searle, Shayle R. (1976a), "Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model," *Technometrics*, 18, 31–38.
———, and Searle, Shayle R. (1976b), "A Comparison of Variance Component Estimators," *Biometrics*, 32, 779–91.
Cunningham, E.P., and Henderson, Charles R. (1968), "An Iterative Procedure for Estimating Fixed Effects and Variance Components in Mixed Model Situations," *Biometrics*, 24, 13–25.
Duncan, David B., and Horn, Susan D. (1972), "Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis," *Journal of the American Statistical Association*, 67, 815–21.
Efron, Bradley, and Morris, Carl (1973), "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–30.
———, and Morris, Carl (1976), "Multivariate Empirical Bayes and Estimation of Covariance Matrices," *Annals of Statistics*, 4, 22–32.
Gautschi, Werner (1959), "Some Remarks on Herbach's Paper, 'Optimum Nature of the *F*-Test for Model II in the Balanced Case,'" *Annals of Mathematical Statistics*, 30, 960–3.
Gill, P.E., and Murray, William A., eds. (1974), *Numerical Methods for Constrained Optimization*, New York: Academic Press.
Goldfeld, S.M., Quandt, R.E., and Trotter, H.F. (1966), "Maximization by Quadratic Hill Climbing," *Econometrica*, 34, 541–51.
Graybill, Franklin A. (1969), *Introduction to Matrices with Applications in Statistics*, Belmont, California: Wadsworth Publishing Company, Inc.
———, and Hultquist, Robert A. (1961), "Theorems Concerning Eisenhart's Model II," *Annals of Mathematical Statistics*, 32, 261–9.
Hartigan, J.A. (1969), "Linear Bayesian Methods," *Journal of the Royal Statistical Society*, Series B, 31, 446–54.
Hartley, H.O., and Rao, J.N.K. (1967), "Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model," *Biometrika*, 54, 93–108.
———, and Vaughn, William 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, Ames, Iowa: Iowa State University Press.
Harville, David A. (1969a), "Quadratic Unbiased Estimation of Variance Components for the One-Way Classification," *Biometrika*, 56, 313–26. (Correction, (1970), *Biometrika*, 57, 226.)
——— (1969b), "Variance-Component Estimation for the Unbalanced One-Way Random Classification-A Critique," Technical Report No. 69-0180, Aerospace Research Laboratories, Wright-Patterson AFB, Ohio.
——— (1974), "Bayesian Inference for Variance Components Using Only Error Contrasts," *Biometrika*, 61, 383–5.
——— (1975), "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems," Technical Report No. 75-0175, Aerospace Research Laboratories, Wright-Patterson AFB, Ohio.
——— (1976), "Extension of the Gauss-Markov Theorem to Include the Estimation of Random Effects," *Annals of Statistics*, 4, 384–95.
Hemmerle, William J., and Hartley, H.O. (1973), "Computing Maximum Likelihood Estimates for the Mixed A.O.V. Model Using the *W* Transformation," *Technometrics*, 15, 819–31.
Henderson, Charles R. (1953), "Estimation of Variance and Covariance Components," *Biometrics*, 9, 226–52.
——— (1963), "Selection Index and Expected Genetic Advance," in *Statistical Genetics and Plant Breeding*, National Academy of Sciences–National Research Council Publication No. 982, 141–63.
——— (1973a), "Maximum Likelihood Estimation of Variance Components," unpublished manuscript.
——— (1973b), "MINQUE of Variance Components," unpublished manuscript.
——— (1973c), "Sire Evaluation and Genetic Trends," in *Proceedings of the Animal Breeding and Genetics Symposium in Honor of Dr. Jay L. Lush*, Champaign, Illinois: American Society of Animal Science, 10–41.
——— (1975), "Best Linear Unbiased Estimation and Prediction Under a Selection Model," *Biometrics*, 31, 423–47.
Herbach, Leon H. (1959), "Properties of Model II-Type Analysis of Variance Tests, A: Optimum Nature of the *F*-Test for Model II in the Balanced Case," *Annals of Mathematical Statistics*, 30, 939–59.
Hocking, Ronald R., and Kutner, Michael H. (1975), "Some Analytical and Numerical Comparisons of Estimators for the Mixed A.O.V. Model," *Biometrics*, 31, 19–28.
Jennrich, R.I., and Sampson, P.F. (1976), "Newton-Raphson and Related Algorithms for Maximum Likelihood Variance Component Estimation," *Technometrics*, 18, 11–17.
Klotz, Jerome H., Milton, Roy C., and Zacks, S. (1969), "Mean Square Efficiency of Estimators of Variance Components," *Journal of the American Statistical Association*, 64, 1383–402.
LaMotte, Lynn R. (1970), "A Class of Estimators of Variance Components," Technical Report No. 10, Department of Statistics, University of Kentucky, Lexington, Kentucky.
——— (1971), "Locally Best Quadratic Estimators of Variance Components," Technical Report No. 22, Department of Statistics, University of Kentucky, Lexington, Kentucky.
——— (1973), "Quadratic Estimation of Variance Components," *Biometrics*, 29, 311–30.
Lawton, William H., and Sylvestre, Edward A. (1971), "Elimination of Linear Parameters in Nonlinear Regression," *Technometrics*, 13, 461–7.
Levenberg, K. (1944), "A Method for the Solution of Certain Non-Linear Problems in Least Squares," *Quarterly of Applied Mathematics*, 2, 164–8.
Lindley, Dennis V., and Smith, A.F.M. (1972), "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society*, Series B, 1–18.
Marquardt, Donald W. (1963), "An Algorithm for Least Squares Estimation of Nonlinear Parameters," *SIAM Journal*, 11, 431–41.
Miller, John J. (1973), "Asymptotic Properties and Computation of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance," Technical Report No. 12, Department of Statistics, Stanford University, Stanford, California.
Murray, William A., ed. (1972), *Numerical Methods for Unconstrained Optimization*, New York: Academic Press.
Nelder, J.A. (1968), "The Combination of Information in Generally Balanced Designs," *Journal of the Royal Statistical Society*, Ser. B, 30, 303–11.
Nering, Evar D. (1970), *Linear Algebra and Matrix Theory*, 2nd ed., New York: John Wiley & Sons.
Olsen, Anthony, Seely, Justus, and Birkes, David (1976), "Invariant Quadratic Unbiased Estimation for Two Variance Components," *Annals of Statistics*, 4, 878–90.
Patterson, H.D., and Thompson, Robin (1971), "Recovery of Inter-Block Information when Block Sizes Are Unequal," *Biometrika*, 58, 545–54.
———, and Thompson, Robin (1974), "Maximum Likelihood Estimation of Components of Variance," *Proceedings of the 8th International Biometric Conference*, 197–207.
Powell, M.J.D. (1970), "A Survey of Numerical Methods for Unconstrained Optimization," *SIAM Review*, 12, 79–97.
Rao, C.R. (1965), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons.
——— (1970), "Estimation of Heteroscedastic Variances in Linear Models," *Journal of the American Statistical Association*, 65, 161–72.
——— (1971a), "Estimation of Variance and Covariance Components—MINQUE Theory," *Journal of Multivariate Analysis*, 1, 257–75.
——— (1971b), "Minimum Variance Quadratic Unbiased Estimation of Variance Components," *Journal of Multivariate Analysis*, 1, 445–56.
——— (1972), "Estimation of Variance and Covariance Components in Linear Models," *Journal of the American Statistical Association*, 67, 112–15.

Rao, J.N.K. (1968), "On Expectations, Variances, and Covariances of ANOVA Mean Squares by 'Synthesis'," *Biometrics*, 24, 963–78.
Russell, Thomas S., and Bradley, Ralph A. (1958), "One-Way Variances in a Two-Way Classification," *Biometrika*, 45, 111–29.
Searle, Shayle R., (1968), "Another Look at Henderson's Methods of Estimating Variance Components" (with discussion), *Biometrics*, 24, 749–87.
——— (1970), "Large Sample Variances of Maximum Likelihood Estimators of Variance Components Using Unbalanced Data," *Biometrics*, 26, 505–24.
——— (1971a), *Linear Models*, New York: John Wiley & Sons.
——— (1971b), "Topics in Variance Component Estimation," *Biometrics*, 27, 1–76.
——— (1974), "Prediction, Mixed Models, and Variance Components," in *Reliability and Biometry*, eds. F. Proschan and R.J. Serfling, Philadelphia, Pennsylvania: Society of Industrial and Applied Mathematics, 229–66.
——— (1976), "Detailed Derivation of Results and Relationships in the ML, REML and MINQUE Methods of Estimating Variance Components," Paper No. BU-586-M, Biometrics Unit, Cornell University, Ithaca, New York.
Seely, Justus, (1975), "An Example of an Inquadmissible Analysis of Variance Estimator for a Variance Component, *Biometrika*, 62, 689–90.
Sprott, D.A. (1975), "Marginal and Conditional Sufficiency," *Biometrika*, 62, 599–605.
Thompson, Robin (1969), "Iterative Estimation of Variance Components for Nonorthogonal Data," *Biometrics*, 25, 767–73.
——— (1973), "The Estimation of Variance and Covariance Components with an Application when Records Are Subject to Culling," *Biometrics*, 29, 527–50.
——— (1975), "A Note on the *W* Transformation," *Technometrics*, 17, 511–2.
Thompson, W.A., Jr. (1962), "The Problem of Negative Estimates of Variance Components," *Annals of Mathematical Statistics*, 33, 273–89.
Townsend, Edwin C. (1968), "Unbiased Estimators of Variance Components in Simple Unbalanced Designs," Ph.D. dissertation, Cornell University, Ithaca, New York.
———, and Searle, Shayle, R. (1971), "Best Quadratic Unbiased Estimation of Variance Components from Unbalanced Data in the 1-Way Classification," *Biometrics*, 27, 643–57.
Weiss, Lionel (1971), "Asymptotic Properties of Maximum Likelihood Estimators in Some Nonstandard Cases," *Journal of the American Statistical Association*, 66, 345–50.
——— (1973), "Asymptotic Properties of Maximum Likelihood Estimators in Some Nonstandard Cases, II," *Journal of the American Statistical Association*, 68, 428–30.
Westlake, Joan R. (1968), *A Handbook of Numerical Matrix Inversion and Solution of Linear Equations*, New York: John Wiley & Sons.
Zacks, S. (1971), *The Theory of Statistical Inference*, New York: John Wiley & Sons.

Comment

J. N. K. RAO*

1. Harville has covered a lot of ground in this excellent review paper. It contains several important features: (1) a treatment of the variance component models as special cases of the general linear model (2.1) which creates a unified presentation; (2) a review of the recent work of Henderson and Harville on the estimation of random effects in the model and its relationship to variance components estimation, especially from the viewpoint of computations; (3) a thorough discussion of specialized methods as well as general algorithms for computing maximum likelihood (ML) estimates; and (4) an examination of the relationships of ML estimates to ANOVA-type estimates of Henderson, MINQUE of C.R. Rao and others, and Bayesian estimates. However, the scope of this paper is somewhat narrow since it is mainly concerned with point estimation of the variance components.

2. A brief account of Miller's (1973) recent work on asymptotic properties of ML estimates is provided in Section 4.2. Miller removed a restriction of Hartley and Rao (1967), viz., the number of observations falling into any particular level of any random factor stays below a universal constant. However, he was able to establish only a Cramér-type consistent result (viz., *some* root of the likelihood equation is consistent) which is not so gratifying as the Wald-type result: an estimate of the parameter vector which makes the likelihood function an absolute maximum is consistent, i.e., a ML estimate is consistent. Using sufficient conditions similar to those of Wald, a corrected version of Hartley-Rao's proof establishes the Wald-type consistency result under the above-mentioned restriction, but it will not work if this restriction is removed (see Miller (1973), p. 257). Perhaps an alternative set of sufficient conditions might give the desired result without the restriction of Hartley-Rao.

3. Recent work on the computational aspects of ML estimates (Section 6) is certainly impressive, and we need further research in this direction. However, none of the proposed algorithms guarantees a solution which indeed is a ML estimate. The behavior of the likelihood (as a function of the variance components) appears complex; even for the simple unbalanced one-way layout, the likelihood equation may have multiple roots or the ML estimate is a boundary value rather than any of these roots (see Klotz and Putter (1970)). We therefore need to exercise more caution when resorting to numerical algorithms, especially with Monte Carlo studies involving the computing of ML estimates for several thousand samples.

Henderson's iterative algorithm (which is a simplification of Hartley-Rao's equation 51 (1967)) appears

* J.N.K. Rao is Professor, Department of Mathematics, Carleton University, Ottawa K1S 5B6, Canada.