ARL TR 75-0175

MAXIMUM LIKELIHOOD APPROACHES TO VARIANCE COMPONENT ESTIMATION AND TO RELATED PROBLEMS

APPLIED MATHEMATICS RESEARCH LABORATORY/ARL

JUNE 1975

FINAL REPORT MAY 1973 – MARCH 1975

AEROSPACE RESEARCH LABORATORIES/LB
Building 450 – Area B
Wright-Patterson Air Force Base, Ohio 45433

AIR FORCE SYSTEMS COMMAND
United States Air Force

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Organizations or individuals receiving reports via Aerospace Research Laboratories automatic mailing lists should refer to the ARL number of the report received when corresponding about change of address or cancellation. Such changes should be directed to the specific laboratory orginating the report. Do not return this copy; retain or destroy.

Reports are not stocked by the Aerospace Research Laboratories. Copies may be obtained from:

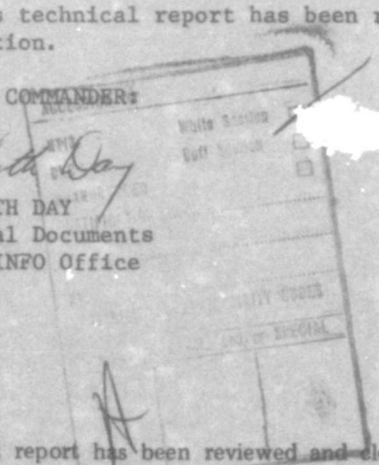
National Technical Information Services
Clearinghouse
Springfield, VA 22161

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER:

Elizabeth Day

ELIZABETH DAY
Technical Documents
and STINFO Office



This report has been reviewed and cleared for open publication and public release by the appropriate Office of Information in accordance with AFR 190-12 and DODD 5230.0. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service.

AIR FORCE/56780/22 July 1975 — 500

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: ARL-75-0175
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle): MAXIMUM LIKELIHOOD APPROACHES TO VARIANCE COMPONENT ESTIMATION AND TO RELATED PROBLEMS
5. TYPE OF REPORT & PERIOD COVERED: Technical Final rept. May 1973—March 1975
6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): David A. Harville
8. CONTRACT OR GRANT NUMBER(s): In-House Research
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Applied Mathematics Research Laboratory/LB, Aerospace Research Laboratories
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: DoD Element 61102F; 7071 02; AF-7071
11. CONTROLLING OFFICE NAME AND ADDRESS: Aerospace Research Laboratories (AFSC), Building 450 - Area B, Wright-Patterson AFB, Ohio 45433
12. REPORT DATE: June 1975
13. NUMBER OF PAGES: 107 (12) 111 p.
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
15. SECURITY CLASS. (of this report): Unclas
15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

DDC RECEIVED SEP 3 1975 D

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

mixed linear models
variance components
estimation
maximum likelihood
restricted maximum likelihood
asymptotic properties
constrained nonlinear optimization algorithms

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Several recent developments promise to increase greatly the popularity of maximum likelihood (ML) as a technique for estimating variance components. Patterson and Thompson [Biometrika, Vol. 58, December 1971, pp. 545-554] proposed a restricted maximum likelihood (REML) approach which takes into account the loss in degrees of freedom resulting from estimating fixed effects. Miller [Technical Report No. 12, Department of Statistics, Stanford University, 1973] developed a realistic asymptotic theory for ML estimators of

NEXT Page

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DN

009950

20.
variance components. There are many iterative algorithms that can be considered for computing ML or REML estimates of variance components. Some were developed specifically for the variance component problem and related problems. Others are general nonlinear optimization procedures. 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. MINQUE's of variance components can be computed from one iteration of the REML version of Anderson's [Annals of Statistics, Vol. 1, January 1973, pp. 135-141] iterative procedure.

PREFACE

This report constitutes the final report on Work Unit 10 "Design and Analysis of Experiments", included in Task 2 "Mathematical Statistics" of Project 7071 "Research in Applied Mathematics" of the Aerospace Research Laboratories, Wright-Patterson Air Force Base, Ohio 45433. It was compiled in-house from May 1973 - March 1975 by Dr. David A. Harville, who was the principal investigator on Work Unit 10. Some of the work was accomplished in June-July 1974, while the author was a visitor at the Biometrics Unit, Cornell University, Ithaca, New York. During that visit, he benefited substantially from many informal discussions with Shayle R. Searle and C. R. Henderson.

The author's present address is: Department of Statistics, Snedecor Hall, Iowa State University, Ames, Iowa 50010.

TABLE OF CONTENTS

SECTION 1

INTRODUCTION

The testing and estimation procedures associated with the analysis of variance (ANOVA) and with 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 [34] 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 have been proposed. Simultaneously, there has been renewed interest in maximum likelihood techniques for estimating variance components.

The maximum likelihood approach to estimation problems has a number of well-known features. Some of these are especially relevant to the variance component problem. Among the properties of maximum likelihood estimators of variance components are the following: (i) The maximum likelihood estimators are consistent and are asymptotically normal and efficient. (ii) The maximum likelihood estimators are functions of every sufficient set of statistics. (iii) The maximum likelihood approach is well-defined for the many

useful generalizations of the ordinary ANOVA models. (iv) The non-negativity constraints on the variance components or other constraints on the parameter space cause no conceptual difficulties. (v) The maximum likelihood estimates and information matrix for a given parameterization of the model can be computed readily from those for any other parameterization. (vi) For at least some unbalanced designs, there exist estimators in the class of locally best translation-invariant quadratic unbiased estimators that have uniformally smaller variance than the Henderson estimators [54]. Moreover, the locally best estimators are related closely to maximum likelihood estimators [40].

In spite of the above 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 that, 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 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. Some of these were derived by statisticians specifically for the variance component problem, while others pertain to the general problem of numerically solving constrained nonlinear optimization problems and tend to be less familiar to statisticians. Even in situations where the computation of the maximum likelihood estimates is still unfeasible, it may be possible, as will be described in Section 7, to compute estimates that approximate them.

Two other problems that have kept maximum likelihood from becoming a more popular technique for estimating variance components are the following: (i) 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. (ii) The maximum likelihood estimators are derived under the assumption of a particular parametric form, generally normal, for the distribution of the data vector. The first of these problems has in effect been eliminated by Patterson and Thompson [55] through their 'restricted maximum likelihood' approach. With regard to the second problem, we argue 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 what follows, we attempt a unified review of the maximum likelihood approach to variance component estimation, with emphasis on the computational problems. Some of the

results that are covered are available only in sources not ordinarily read by statisticians. A few of the results are new.

Among the topics included in our coverage are the following: (i) the current state of maximum likelihood theory as applied to the estimation of variance components, (ii) the relationship (shown to be intimate) between the maximum likelihood estimation of the variance components and the estimation or prediction of the model's fixed and random effects, (iii) the exploitation of that relationship for purposes of computation and approximation, (iv) numerical algorithms for computing maximum likelihood estimates of variance components, (v) the use of maximum likelihood as a vehicle for relating the various methods that have been proposed for estimating variance components, and (vi) directions for further 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 that is to be estimated. Throughout the paper, we attempt to promote that viewpoint. Many of the ideas that are discussed are applicable to the more general problem. We specialize to the variance component case only when it appears necessary or instructive to do so.

SECTION 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

$$\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\alpha} + \underset{\sim}{Z}\underset{\sim}{b} + \underset{\sim}{e}. \qquad (2.1)$$

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

In the ordinary fixed ANOVA or regression models, each observation is assumed to differ from its expectation by a simple random error. Usually, these random errors are assumed to be uncorrelated with common variance, so that, in terms of

the model (2.1), $q = 0$, $m = 1$, $\theta_1 \equiv \mathrm{var}(e_i)$, $\underset{\sim}{V} = \underset{\sim}{R} = \theta_1 \underset{\sim}{I}$, and $\Omega = \{\underset{\sim}{\theta} : \theta_1 > 0\}$. Sometimes, the error variances are taken to be heteroscedastic (see, e.g., [59]), in which case $q = 0$, m is some specified number between 1 and n, $\theta_j \equiv \mathrm{var}(e_i)$ for $i \in S_j$ where $\{S_1, \ldots, S_m\}$ is a specified partitioning of the first n integers, $\underset{\sim}{V} = \underset{\sim}{R} = \mathrm{diag}[\theta_{j(1)}, \ldots, \theta_{j(n)}]$ where $j(i) = j$ if $i \in S_j$, and $\Omega = \{\underset{\sim}{\theta} : \theta_j > 0 \ (j = 1, \ldots, m)\}$.

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, with the levels of the other factors, and with the 'residual effects.' Associated with the ith random factor is a parameter σ_i^2, representing 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 the model (2.1), $\underset{\sim}{b}' = (\underset{\sim}{b}_1', \ldots, \underset{\sim}{b}_c')$ where $\underset{\sim}{b}_i$ is a $q_i \times 1$ vector whose elements are the levels of the ith random factor, $m = c + 1$,

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

$$\underset{\sim}{R} = \theta_m \underset{\sim}{I}, \tag{2.3}$$

$$\underset{\sim}{D} = \mathrm{diag}[\theta_1 \underset{\sim}{I}, \ldots, \theta_{m-1} \underset{\sim}{I}], \tag{2.4}$$

$$\underset{\sim}{V} = \theta_m \underset{\sim}{I} + \textstyle\sum_{i=1}^{m-1} \theta_i \underset{\sim}{Z}_i \underset{\sim}{Z}_i' \tag{2.5}$$

where $\underset{\sim}{Z}_i$ is a $n \times q_i$ matrix defined by the partitioning

$\underset{\sim}{Z} = (\underset{\sim}{Z}_1, \ldots, \underset{\sim}{Z}_{m-1})$, and

$$\Omega = \{\underset{\sim}{\theta} : \theta_m > 0, \theta_i \geq 0 \ (i = 1, \ldots, m-1)\}. \qquad (2.6)$$

Generally, each row of $\underset{\sim}{Z}_i$ has a single element equal to one and its remaining elements 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 $\underset{\sim}{X}$ will usually also have only 0-1 entries. In particular, in the ordinary random ANOVA models, $\underset{\sim}{X} = \underset{\sim}{1}$. ($\underset{\sim}{1}$ denotes a column vector each of whose elements is one.)

The ordinary mixed and random ANOVA models are sometimes parameterized in terms of $\gamma_{c+1} = \sigma^2_{c+1}$, $\gamma_i = \sigma^2_i/\sigma^2_{c+1}$ $(i = 1, \ldots, c)$ rather than in terms of the c+1 variance components $\sigma^2_1, \ldots, \sigma^2_{c+1}$. If we had taken

$$\theta_i = \gamma_i \ (i = 1, \ldots, m) \qquad (2.7)$$

instead of taking $\underset{\sim}{\theta}$ to be as specified by (2.2), we would have had $D = \theta_m \operatorname{diag}[\theta_1 \underset{\sim}{I}, \ldots, \theta_{m-1} \underset{\sim}{I}]$ and $V = \theta_m(\underset{\sim}{I} + \sum_{i=1}^{m-1} \theta_i \underset{\sim}{Z}_i \underset{\sim}{Z}_i')$ in place of the representations (2.4) and (2.5).

There would seem to be no point in describing particular ANOVA models or in introducing specific examples of data sets to which they can be applied. Most readers will be familiar with these models and will be able to draw on their own experience for examples. In any case, the more common ANOVA models are described in many statistics books (see, e.g., [67]),

and a wide variety of specific applications can also be found in the literature. In particular, [42], [10], [71], [15], and [41] contain examples from genetics, biology, agriculture, the physical and engineering sciences, and the behavioral sciences, respectively.

It is a mistake to think of linear models only in terms of the ordinary regression and ANOVA models, even though the latter models are very useful. 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 $\underset{\sim}{y}$ or various subvectors of $\underset{\sim}{y}$ 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 the model (2.1) in which $q = 0$ and $\underset{\sim}{e}$ or subvectors of $\underset{\sim}{e}$ are generated by stochastic processes like autoregressive processes, moving average processes, or mixed autoregressive moving average processes (see [5] and [11]). Particularly useful generalizations of these special cases can be obtained by supposing that the elements of $\underset{\sim}{b}$ or of various of its subvectors are also ordered by time and may have been generated by similar stochastic processes. Models of this kind should be suitable for a wide variety of growth curve data. Also, many types of data that ordinarily are analyzed by the usual ANOVA models are better fit by these generalized time series models. As further evidence of the versatility of the

latter models, we note that they were used as the basis for a system for rating high school or college football teams [32]. The specification of $\underset{\sim}{Z}$, m, $\underset{\sim}{R}$, $\underset{\sim}{D}$, and Ω and the interpretation of $\underset{\sim}{b}$ and $\underset{\sim}{\theta}$ in these time series models 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 the formulation (2.1). While the particular models described above are essentially univariate models, i.e., models in which each component of $\underset{\sim}{y}$ 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 $\underset{\sim}{y}$ so that, 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 r-variate example where the model is likewise a special case of (2.1). In these r-variate analogues, we suppose that the n observations have been numbered so that $\underset{\sim}{y}$ is a vector of s (= n/r) multivariate vectors. It will generally be the case in applications of the r-variate models that $\underset{\sim}{X} = \underset{\sim}{I} \underset{\sim}{\times} \underset{\sim}{\chi}$ and $\underset{\sim}{Z} = \underset{\sim}{I} \underset{\sim}{\times} \underset{\sim}{\zeta}$ where $\underset{\sim}{\chi}$ and $\underset{\sim}{\zeta}$ are s × p and s × q matrices and $\underset{\sim}{\times}$ is the direct product operation as defined on p. 197 of [23], though there are exceptions as noted by Thompson [74].

The models that underlie the multivariate analysis of variance (MANOVA) are generalizations of the ordinary ANOVA

models. In conjunction with these models, take $\underset{\sim}{\Sigma}_i$ to be the $r \times r$ matrix whose jkth element is the common covariance between the levels of the ith factor of the jth variate and the corresponding levels of the ith factor of the kth variate ($i = 1, \ldots, c$). Similarly, take $\underset{\sim}{\Sigma}_{c+1}$ to be the $r \times r$ matrix whose jkth element is the common covariance between the residual effects associated with the jth variate and the corresponding residual effects associated with the kth variate. The diagonal elements of $\underset{\sim}{\Sigma}_1, \ldots, \underset{\sim}{\Sigma}_{c+1}$ are the variance components, while the off-diagonal elements are called covariance components. In terms of the model (2.1),

$$\underset{\sim}{b}' = (\underset{\sim}{b}_{11}', \ldots, \underset{\sim}{b}_{1q_1}', \ldots, \underset{\sim}{b}_{c1}', \ldots, \underset{\sim}{b}_{cq_c}')$$

where $\underset{\sim}{b}_{ij}$ is a $r \times 1$ vector whose kth element is the jth level of the ith factor for the kth variate; $m = r(r+1)(c+1)/2$; the components of $\underset{\sim}{\theta}$ consist of the elements of the upper (or lower) triangular portions of the symmetric matrices $\underset{\sim}{\Sigma}_1, \ldots, \underset{\sim}{\Sigma}_{c+1}$; $\underset{\sim}{R} = \underset{\sim}{\Sigma}_{c+1} \times \underset{\sim}{I}$; $\underset{\sim}{D} = \text{diag}[\underset{\sim}{\Sigma}_1 \times \underset{\sim}{I}, \ldots, \underset{\sim}{\Sigma}_c \times \underset{\sim}{I}]$; $\underset{\sim}{V} = [\underset{\sim}{\Sigma}_{c+1} \times \underset{\sim}{I}] + \sum_{i=1}^{c} [\underset{\sim}{\Sigma}_i \times (\underset{\sim}{\zeta}_i \underset{\sim}{\zeta}_i')]$ where $\underset{\sim}{\zeta}_i$ is a $s \times q_i$ matrix defined by $\underset{\sim}{\zeta} = (\underset{\sim}{\zeta}_1, \ldots, \underset{\sim}{\zeta}_c)$; and $\Omega = \{\underset{\sim}{\theta} : \underset{\sim}{\Sigma}_{c+1}$ is positive definite, $\underset{\sim}{\Sigma}_i$ is non-negative ($i = 1, \ldots, c$)$\}$. The matrices $\underset{\sim}{X}, \underset{\sim}{\zeta}_1, \ldots, \underset{\sim}{\zeta}_c$ will ordinarily exhibit simple structures analagous to those noted in conjunction with the corresponding entities for ANOVA models.

In analogy to the reparameterization of the ANOVA models from $\sigma_1^2, \ldots, \sigma_{c+1}^2$ to $\gamma_1, \ldots, \gamma_{c+1}$, we could parameterize the MANOVA models in terms of the elements of the upper triangular portions of the symmetric matrices $\underset{\sim}{\Gamma}_{c+1} = \underset{\sim}{\Sigma}_{c+1}$, $\underset{\sim}{\Gamma}_i = (\underset{\sim}{\Sigma}_{c+1}^{\frac{1}{2}})^{-1} \underset{\sim}{\Sigma}_i (\underset{\sim}{\Sigma}_{c+1}^{\frac{1}{2}})^{-1}$ $(i = 1, \ldots, c)$, where $\underset{\sim}{\Sigma}_{c+1}^{\frac{1}{2}}$ is the square root (as defined, e.g., in [50]) of $\underset{\sim}{\Sigma}_{c+1}$, rather than in terms of the elements of the upper triangular portions of $\underset{\sim}{\Sigma}_1, \ldots, \underset{\sim}{\Sigma}_{c+1}$.

There are also multivariate analogues for the generalized time series models. Multivariate models of this kind form the basis for Kalman filtering techniques, which are much used in engineering applications (see, e.g., [16]), and can be applied to multivariate growth curve data.

In general there will be more than one way to formulate a given linear model as a special case of (2.1). In particular, if the given model is written in the form (2.1) with $\underset{\sim}{Z}$ (and $\underset{\sim}{D}$) non-null, then an equivalent (in the sense that the same probability distribution is effected for $\underset{\sim}{y}$) representation of the form (2.1) is obtained by taking $q = 0$ and letting $\underset{\sim}{e}$ in the new formulation represent $(\underset{\sim}{Z}\underset{\sim}{b} + \underset{\sim}{e})$ from the original formulation. This observation may cause some readers to question whether it makes any difference which formulation is chosen for the model or, put another way, whether the inclusion of the term $\underset{\sim}{Z}\underset{\sim}{b}$ in (2.1) serves any useful purpose other than that of indicating the form of $\underset{\sim}{V}$. That the answer to this question is yes will become clear as we proceed. Basically, a good choice

for the formulation is one that leads as directly as possible to useful procedures for analyzing the data.

So far, the special cases of the general linear model (2.1) that have been discussed are ones in which $\underset{\sim}{Z}$ is a known matrix. There are also useful special cases of (2.1) in which at least some elements of $\underset{\sim}{Z}$ are nontrivial functions of unobservable parameters. These special cases include some alternative formulations for the generalized time series models. They also include useful formulations for certain linear models employed by psychologists and other behavioral scientists. In particular, these formulations could be applied to factor analysis models with non-null elements of $\underset{\sim}{Z}$ representing the 'factor loadings.' (Discussion of factor analysis models and related models and their applications in the behavioral sciences can be found in [41].)

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

$$\underset{\sim}{V} = \sum_{i=1}^{m} \theta_i \underset{\sim}{G}_i \qquad (2.8)$$

for some $n \times r$ symmetric matrices $\underset{\sim}{G}_1, \ldots, \underset{\sim}{G}_m$ whose elements are known. As Anderson [3] indicated, this class includes many useful special cases. In particular, $\underset{\sim}{V}$ has the form (2.8) for the usual ANOVA models when we take $\theta_i = \sigma_i^2$ $(i = 1, \ldots, c+1)$ though not when we take $\theta_i = \gamma_i$. Similarly, the ordinary MANOVA models are included in the class of linear models

characterized by (2.8) provided that we take the elements of $\underset{\sim}{\theta}$ to be the elements of the upper triangular portions of $\underset{\sim}{\Sigma}_1, \ldots, \underset{\sim}{\Sigma}_{c+1}$.

Even if $\underset{\sim}{V}$ does not exhibit the form (2.8), $\underset{\sim}{V}^{-1}$ may be linear in the parameters, i.e., $\underset{\sim}{V}^{-1}$ may have the representation

$$\underset{\sim}{V}^{-1} = \sum_{i=1}^{m} \theta_i \underset{\sim}{H}_i \tag{2.9}$$

where $\underset{\sim}{H}_1, \ldots, \underset{\sim}{H}_m$ are symmetric $n \times n$ matrices whose elements are known. (See Section 5.2 in [3] for an example.) Specialized results are also available in the literature for those linear models characterized by (2.9).

SECTION 3

ESTIMATION OF FIXED AND RANDOM EFFECTS

Corresponding to an actual data vector, i.e., an observed value of $\underset{\sim}{y}$, is a 'realized' or 'sample' value of $\underset{\sim}{b}$ and a realized or sample value of $\underset{\sim}{e}$. These values, which are of course unobservable (at least at the time the data must be analyzed), will subsequently be denoted by $\underset{\sim}{\beta}$ and $\underset{\sim}{\varepsilon}$ can be thought of as parameter vectors just as $\underset{\sim}{\alpha}$ is a parameter vector. The only distinction is that something is assumed to be known about the origin of $\underset{\sim}{\beta}$ and $\underset{\sim}{\varepsilon}$. The problem of estimating estimable functions of $\underset{\sim}{\alpha}$ is considered to be of great practical importance and has been dealt with in many articles.

In contrast, the problem of estimating $\underset{\sim}{\beta}$ or linear combinations of the components of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ has not received much attention (at least not from statisticians). Nevertheless, the latter problem is also of considerable practical importance. In animal breeding applications, where the observations consist of production records on some trait such as dairy cattle milk production, certain components of $\underset{\sim}{b}$ may represent deviations of the breeding values of individual animals or breeding lines from a population mean which is expressible as an estimable function of $\underset{\sim}{\alpha}$, the vector of fixed effects. These deviations are regarded as random effects because the particular animals or lines that are represented in the data (or their breeding values) are considered to be a random sample from an underlying population of animals, lines, or breeding values. The determination of the characteristics of the underlying population need not in itself be the primary objective in analyzing the data. Rather, the main objective may be to estimate the breeding values of the very same animals that are represented in the data. These values are linear combinations of the components of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$. The estimates ultimately are used to decide which of the animals or lines to retain for future breeding purposes. (See, e.g., [38].)

Under quite general circumstances, the problem of estimating or predicting a future data point from data to which (2.1) apply can be formulated as a problem of estimating a linear combination of the components of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$. Such a

formulation is possible and meaningful if the future data point can be viewed as the sum of an estimable function of $\underset{\sim}{\alpha}$ and the sample value of a random variable that has zero mean and is uncorrelated with the elements of $\underset{\sim}{e}$. To see that this is so, note that this random variable can be inserted as an element of $\underset{\sim}{b}$ by taking the corresponding column of $\underset{\sim}{Z}$ to be null. Moreover, the problem in factor analysis of estimating the values, for given individuals, of the 'common factors' is also a particular case of the problem of estimating linear combinations of the components of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$. Still other examples of practical settings where the latter problem is encountered are given in [69].

We now review some results on the estimation of linear combinations of the components of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and even $\underset{\sim}{\varepsilon}$, for the case where the 'true' value of $\underset{\sim}{\theta}$ and possibly the true value of $\underset{\sim}{\alpha}$ are known. These results provide some insight into how to estimate such linear combinations when, as is being assumed here, the true values of $\underset{\sim}{\theta}$ and $\underset{\sim}{\alpha}$ are unknown. Also, as will become evident in subsequent sections, these results are very relevant to the problem of estimating $\underset{\sim}{\theta}$ by maximum likelihood.

Subsequently, we take $\underset{\sim}{\tilde{\alpha}}$ to be any solution to the 'normal' equations.

$$(\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X})\underset{\sim}{\tilde{\alpha}} = \underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{y}, \tag{3.1}$$

and put

$$\hat{\underset{\sim}{\Psi}} = \underset{\sim}{Z}'\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\alpha),$$

$$\hat{\underset{\sim}{\Psi}}^* = \underset{\sim}{Z}'\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\alpha}),$$

$$\hat{\underset{\sim}{\beta}} = \underset{\sim}{D}\hat{\underset{\sim}{\Psi}}, \tag{3.2}$$

and

$$\hat{\underset{\sim}{\beta}}^* = \underset{\sim}{D}\hat{\underset{\sim}{\Psi}}^*. \tag{3.3}$$

Since the elements of $\underset{\sim}{D}$, $\underset{\sim}{R}$, and possibly $\underset{\sim}{Z}$ are functions of the parameter vector $\underset{\sim}{\theta}$, so are the elements of $\underset{\sim}{V}$, $\hat{\underset{\sim}{\alpha}}$, $\hat{\underset{\sim}{\Psi}}^*$, and $\hat{\underset{\sim}{\beta}}^*$. Moreover, the elements of $\hat{\underset{\sim}{\Psi}}$ and $\hat{\underset{\sim}{\beta}}$ are functions of both $\underset{\sim}{\theta}$ and $\underset{\sim}{\alpha}$. 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 that vector or matrix associated with particular values for the parameter vectors. Thus, e.g., $\hat{\underset{\sim}{\beta}}(\underset{\sim}{\theta},\underset{\sim}{\alpha})$ and $\hat{\underset{\sim}{\beta}}^*(\underset{\sim}{\theta})$ are used interchangeably with $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{\beta}}^*$, respectively, and, if $\underset{\sim}{\theta}^*$ and $\underset{\sim}{\alpha}^*$ are particular values of $\underset{\sim}{\theta}$ and $\underset{\sim}{\alpha}$, $\hat{\underset{\sim}{\beta}}(\underset{\sim}{\theta}^*,\underset{\sim}{\alpha}^*)$ and $\hat{\underset{\sim}{\beta}}^*(\underset{\sim}{\theta}^*)$ are the values of $\hat{\underset{\sim}{\beta}}$ and $\hat{\underset{\sim}{\beta}}^*$ for $\underset{\sim}{\theta} = \underset{\sim}{\theta}^*$ and $\underset{\sim}{\alpha} = \underset{\sim}{\alpha}^*$. Note that $\hat{\underset{\sim}{\Psi}}^*(\underset{\sim}{\theta}) = \hat{\underset{\sim}{\Psi}}[\underset{\sim}{\theta},\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta})]$ and that $\hat{\underset{\sim}{\beta}}^*(\underset{\sim}{\theta}) = \hat{\underset{\sim}{\beta}}[\underset{\sim}{\theta},\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta})]$. Subsequently, we denote the true values of $\underset{\sim}{\theta}$ and $\underset{\sim}{\alpha}$ by $\underset{\sim}{\theta}^+$ and $\underset{\sim}{\alpha}^+$.

By the expectation of a random variable, we shall, unless otherwise indicated, mean its unconditional expectation with respect to the joint distribution of $\underset{\sim}{b}$ and $\underset{\sim}{e}$, as opposed say to its conditional expectation given $\underset{\sim}{b} = \underset{\sim}{\beta}$. We refer to an

estimator $t(\underset{\sim}{y})$ of a linear combination of the elements of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\varepsilon}$, say $\underset{\sim}{\lambda}_1'\underset{\sim}{\alpha} + \underset{\sim}{\lambda}_2'\underset{\sim}{\beta} + \underset{\sim}{\lambda}_3'\underset{\sim}{\varepsilon}$, as unbiased if $E[t(\underset{\sim}{y})] \equiv \underset{\sim}{\lambda}_1'\underset{\sim}{\alpha}$, label it linear if $t(y) = a + \underset{\sim}{u}'\underset{\sim}{y}$ for some scalar a and some $n \times 1$ vector $\underset{\sim}{u}$, and call $E[t(\underset{\sim}{y}) - \underset{\sim}{\lambda}_1'\underset{\sim}{\alpha} - \underset{\sim}{\lambda}_2'\underset{\sim}{b} - \underset{\sim}{\lambda}_3'\underset{\sim}{e}]^2$ its mean squared error.

An answer to the question of how to estimate $\underset{\sim}{\lambda}_1'\underset{\sim}{\alpha} + \underset{\sim}{\lambda}_2'\underset{\sim}{\beta} + \underset{\sim}{\lambda}_3'\underset{\sim}{\varepsilon}$ when $\underset{\sim}{\theta}^+$ and possibly $\underset{\sim}{\alpha}^+$ are known is given by the following theorem.

Theorem 1: (i) In the case where the true values $\underset{\sim}{\theta}^+$ and $\underset{\sim}{\alpha}^+$ of $\underset{\sim}{\theta}$ and $\underset{\sim}{\alpha}$ are known, the best (uniformly smallest mean squared error) linear unbiased estimator (BLUE) of a linear combination $\underset{\sim}{\lambda}_1'\underset{\sim}{\alpha} + \underset{\sim}{\lambda}_2'\underset{\sim}{\beta} + \underset{\sim}{\lambda}_3'\underset{\sim}{\varepsilon}$ of the elements of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\varepsilon}$ is

$$\underset{\sim}{\lambda}_1'\underset{\sim}{\alpha}^+ + \underset{\sim}{\lambda}_2'\hat{\underset{\sim}{\beta}}(\underset{\sim}{\theta}^+,\underset{\sim}{\alpha}^+) + \underset{\sim}{\lambda}_3'[\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha}^+ - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}(\underset{\sim}{\theta}^+,\underset{\sim}{\alpha}^+)]; \qquad (3.4)$$

(ii) in the case where only $\underset{\sim}{\theta}^+$ is known, the BLUE of a linear combination $\underset{\sim}{\lambda}_1'\underset{\sim}{\alpha} + \underset{\sim}{\lambda}_2'\underset{\sim}{\beta} + \underset{\sim}{\lambda}_3'\underset{\sim}{\varepsilon}$, where $\underset{\sim}{\lambda}_1'\underset{\sim}{\alpha}$ is estimable, is

$$\underset{\sim}{\lambda}_1'\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta}^+) + \underset{\sim}{\lambda}_2'\hat{\underset{\sim}{\beta}}^*(\underset{\sim}{\theta}^+) + \underset{\sim}{\lambda}_3'[\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta}^+) - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*(\underset{\sim}{\theta}^+)]. \qquad (3.5)$$

Part (i) of Theorem 1 is essentially the result described by Rao in Section 4a.11 of [58]. When $\underset{\sim}{\lambda}_2 = \underset{\sim}{0}$ and $\underset{\sim}{\lambda}_3 = \underset{\sim}{0}$, part (ii) reduces to the ordinary Gauss-Markov theorem, and, when $\underset{\sim}{\lambda}_1 = \underset{\sim}{0}$ and $\underset{\sim}{\lambda}_3 = \underset{\sim}{0}$, it reduces to a result derived by Henderson [35] (see also [38]). Part (ii) is essentially the same as a generalization of these two special cases that was derived independently by Henderson [38] and Harville [30].

When the joint distribution of $\underset{\sim}{b}$ and $\underset{\sim}{e}$ is multivariate normal, stronger statements can be made about the 'goodness' of (3.4) and (3.5). Then, in the case where $\underset{\sim}{\theta}^+$ and $\underset{\sim}{\alpha}^+$ are both known, the estimator (3.4) has minimum mean squared error among all estimators of $\underset{\sim}{\lambda}_1'\underset{\sim}{\alpha} + \underset{\sim}{\lambda}_2'\underset{\sim}{\beta} + \underset{\sim}{\lambda}_3'\underset{\sim}{\varepsilon}$, and in the case where only $\underset{\sim}{\theta}^+$ is known, the estimator (3.5) has minimum mean squared error among all unbiased estimators.

If $\underset{\sim}{R} \equiv \theta_m \underset{\sim}{R}^*$ and $\underset{\sim}{D} \equiv \theta_m \underset{\sim}{D}^*$ where $\theta_m \neq 0$ for $\underset{\sim}{\theta} \in \Omega$ and where the elements of $\underset{\sim}{R}^*$ and $\underset{\sim}{D}^*$ depend on $\theta_1, \ldots, \theta_{m-1}$ but not on θ_m, then $\hat{\underset{\sim}{\beta}}$, $\hat{\underset{\sim}{\beta}}^*$, and the solution space to the linear system (3.1) are free from dependence on θ_m. For example, in the ordinary ANOVA models, $\hat{\underset{\sim}{\beta}}$, $\hat{\underset{\sim}{\beta}}^*$, and the solution space to (3.1) depend on the variance components $\sigma_1^2, \ldots, \sigma_{c+1}^2$ only through the ratios $\gamma_1, \ldots, \gamma_c$. If $\underset{\sim}{R}$ and $\underset{\sim}{D}$ have this form, then part (i) of Theorem 1 still applies when only $\underset{\sim}{\alpha}^+$ and the first m-1 elements of $\underset{\sim}{\theta}^+$ are known and part (ii) remains valid when only the first m-1 elements of $\underset{\sim}{\theta}^+$ are known.

Other properties pertaining to (3.4) and (3.5) are described in [30], [35], [38], and [69]. Results relevant to the computation of $\hat{\underset{\sim}{\Psi}}$ and $\hat{\underset{\sim}{\beta}}$ or $\hat{\underset{\sim}{\Psi}}^*$ and $\hat{\underset{\sim}{\beta}}^*$ and to the computation of a solution to (3.1) are also given in those references. We state two of these results as the two parts of the following theorem.

Theorem 2: (i) The linear system

$$\begin{bmatrix} \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{X} & \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D} \\ \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{X} & \underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D} \end{bmatrix} \begin{bmatrix} \tilde{\underset{\sim}{\alpha}} \\ \tilde{\underset{\sim}{\psi}} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{y} \\ \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{y} \end{bmatrix}, \tag{3.6}$$

where $\tilde{\underset{\sim}{\alpha}}$ is $p \times 1$ and $\tilde{\underset{\sim}{\psi}}$ is $q \times 1$, is consistent. All solutions to this system are obtained by allowing $\tilde{\underset{\sim}{\alpha}}$ to range over the solution space to the normal equations (3.1) and by taking $\tilde{\underset{\sim}{\psi}} = \tilde{\underset{\sim}{\psi}}^{*}$. (ii) The linear system

$$\begin{bmatrix} \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{X} & \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{Z} \\ \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{X} & \underset{\sim}{D}^{-1} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z} \end{bmatrix} \begin{bmatrix} \tilde{\underset{\sim}{\alpha}} \\ \tilde{\underset{\sim}{\beta}} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{y} \\ \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{y} \end{bmatrix}, \tag{3.7}$$

where $\tilde{\underset{\sim}{\alpha}}$ is $p \times 1$ and $\tilde{\underset{\sim}{\beta}}$ is $q \times 1$ and where $\theta \in \Omega$ is such that $\underset{\sim}{D}$ is nonsingular, is consistent. All solutions to this system are obtained by allowing $\tilde{\underset{\sim}{\alpha}}$ to range over the solution space to (3.1) and by taking $\tilde{\underset{\sim}{\beta}} = \hat{\underset{\sim}{\beta}}^{*}$.

Results closely related to part (i) of Theorem 2 include:

$$\underset{\sim}{V}^{-1} \equiv \underset{\sim}{R}^{-1} - \underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D}(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{R}^{-1}, \tag{3.8}$$

leading to the form

$$[\underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{X} - \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D}(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{X}]\tilde{\underset{\sim}{\alpha}}$$

$$= [\underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{y} - \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D}(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{y}] \tag{3.9}$$

for the linear system (3.1); and

$$\underset{\sim}{Z}'\underset{\sim}{V}^{-1} \equiv (\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{R}^{-1}, \tag{3.10}$$

producing the identities

$$\hat{\underset{\sim}{\Psi}} \equiv (\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{R}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha}) \tag{3.11}$$

and

$$\hat{\underset{\sim}{\Psi}}^{*} \equiv (\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{R}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}). \tag{3.12}$$

Moreover, putting

$$\underset{\sim}{P} \equiv \underset{\sim}{V}^{-1} - \underset{\sim}{V}^{-1}\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X})^{-}\underset{\sim}{X}'\underset{\sim}{V}^{-1}$$

and

$$\underset{\sim}{S} \equiv \underset{\sim}{R}^{-1} - \underset{\sim}{R}^{-1}\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{X})^{-}\underset{\sim}{X}'\underset{\sim}{R}^{-1}$$

(for any matrix $\underset{\sim}{B}$, $\underset{\sim}{B}^{-}$ will denote an arbitrary generalized inverse of $\underset{\sim}{B}$, i.e., any solution to $\underset{\sim}{B}\underset{\sim}{B}^{-}\underset{\sim}{B} = \underset{\sim}{B}$),

$$\underset{\sim}{P} \equiv \underset{\sim}{S} - \underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D}(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{S}; \tag{3.13}$$

and

$$\underset{\sim}{Z}'\underset{\sim}{P} \equiv (\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{S}, \tag{3.14}$$

yielding the identity

$$\hat{\underset{\sim}{\Psi}} \equiv (\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{y}. \tag{3.15}$$

Note the analogy between the identities (3.8), (3.10), and (3.11) and the identities (3.13) - (3.15). Also,

$$(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D})^{-1} \equiv (\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1}[\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{X}$$

$$\cdot\{\underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{X} - \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D}(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{X}\}^{-}$$

$$\cdot\underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D}(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1}]. \qquad (3.16)$$

The expressions (3.8) - (3.16) are valid even for $\underset{\sim}{\theta} \in \Omega$ for which $\underset{\sim}{D}$ is singular.

For $\underset{\sim}{\theta} \in \Omega$ such that $\underset{\sim}{D}$ is nonsingular,

$$\underset{\sim}{D}(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D})^{-1} = (\underset{\sim}{D}^{-1} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z})^{-1} \qquad (3.17)$$

and

$$\underset{\sim}{D}(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D})^{-1} = (\underset{\sim}{D}^{-1} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z})^{-1}. \qquad (3.18)$$

By applying (3.17) and (3.18) together with (3.2) and (3.3) to the expressions (3.8) - (3.16), we can obtain expressions that bear the same relationships to part (ii) of Theorem 2 that (3.8) - (3.16) bear to part (i).

In our formulation of the ordinary ANOVA models as special cases of the general linear model (2.1), the matrices $\underset{\sim}{R}$, $\underset{\sim}{D}$, and and possibly $\underset{\sim}{Z}$ and $\underset{\sim}{X}$ exhibit relatively simply structures. Some or all of these matrices also have simple structures in other useful special cases of (2.1). The significance of Theorem 2 and the related results is that they provide us with the means for exploiting these structures; i.e., they enable us to compute $\hat{\underset{\sim}{\psi}}$ and/or $\hat{\underset{\sim}{\beta}}$ or $\hat{\underset{\sim}{\psi}}^*$, $\hat{\underset{\sim}{\beta}}^*$, and/or a solution to (3.1) much more efficiently when these structures are present than

is possible when they are not. In particular, if we compute $\hat{\underset{\sim}{\psi}}^*$ and a solution to (3.1) by solving the linear system (3.6), then these structures can be used to obvious advantage in computing the entries in the coefficient matrix and right hand side of (3.6) and again in the actual solution of the system. Equivalently, we could exploit these structures by using (3.9) to form and solve (3.1) and by computing $\hat{\underset{\sim}{\psi}}^*$ on the basis of (3.12). Or, again equivalently, we could compute $\hat{\underset{\sim}{\psi}}^*$ from (3.15), possibly using (3.16) also, and then compute a solution to (3.1) from the first p equations in the system (3.6). In carrying out the computations, advantage should be taken of the well-known fact (see, e.g., [78]) that $\underset{\sim}{F}^{-1}\underset{\sim}{C}$, where $\underset{\sim}{F}$ and $\underset{\sim}{C}$ are arbitrary except for obvious restrictions, is computed most efficiently by numerical techniques that solve the linear system $\underset{\sim}{F}\underset{\sim}{B} = \underset{\sim}{C}$ without explicitly forming $\underset{\sim}{F}^{-1}$.

The result stated as part (ii) of Theorem 2 is due to Henderson. Part (i) is one of several modified versions presented in [30] of Henderson's result. This particular version combines two features that make it especially well-suited for describing the relationship between the computation of maximum likelihood estimates of $\underset{\sim}{\theta}$, and the computation of 'estimates' of linear combinations of the elements of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\varepsilon}$: (i) it is applicable for all $\underset{\sim}{\theta} \in \Omega$; and (ii) the vector $\hat{\underset{\sim}{\psi}}^*$ is imbedded as a subvector in any solution to the linear system (3.6). The relevance of the latter feature is that in general the vector $\hat{\underset{\sim}{\psi}}^*$ is 'more fundamentally' related

to the maximum likelihood estimation of $\underset{\sim}{\theta}$ than is $\hat{\underset{\sim}{\beta}}^*$. Another feature of the system (3.6) that can be attractive is that its use does not require the inversion of $\underset{\sim}{D}$. On the other hand, the coefficient matrix of the system (3.7) is symmetric positive definite or symmetric positive semi-definite, which can also be a useful feature from a computational standpoint (see, e.g., [78]). Proofs for Theorem 2 and the results (3.8) - (3.16) can be produced readily from results given and cited in [30].

If $\underset{\sim}{\theta}^+$ and $\underset{\sim}{\alpha}^+$ are both unknown as is being assumed here and is almost always the case in practice, then in general (3.5) can no longer be regarded as an estimator of the linear combination $\underset{\sim}{\lambda}_1'\underset{\sim}{\alpha} + \underset{\sim}{\lambda}_2'\underset{\sim}{\beta} + \underset{\sim}{\lambda}_3'\underset{\sim}{\varepsilon}$. One way to proceed when both $\underset{\sim}{\theta}^+$ and $\underset{\sim}{\alpha}^+$ are unknown is to use as an estimator the expression (3.5) with $\underset{\sim}{\theta}^+$ replaced by an estimator of $\underset{\sim}{\theta}$. In particular, a maximum likelihood estimator of $\underset{\sim}{\theta}$ could be substituted. Thus, in some applications, a maximum likelihood estimate of $\underset{\sim}{\theta}$ may be sought for purposes of producing estimates of certain linear combinations of the elements of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\varepsilon}$ that are of interest. Of course, a maximum likelihood estimate of $\underset{\sim}{\theta}$ can also be of direct interest as in various genetic applications where certain functions of $\underset{\sim}{\theta}$ may be identified with particular characteristics of an underlying genetic population.

SECTION 4

THE MAXIMUM LIKELIHOOD APPROACH TO THE ESTIMATION OF $\underset{\sim}{\theta}$

4.1 Definition

In our discussion of the maximum likelihood estimation of $\underset{\sim}{\theta}$, we take the distribution of $\underset{\sim}{y}$ to be of the multivariate normal form, so that the logarithm of the likelihood function differs by only an additive constant from the function

$$L(\underset{\sim}{\theta},\underset{\sim}{\alpha};\underset{\sim}{y}) = -(\tfrac{1}{2})\log[\det(\underset{\sim}{V})] - (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha})'\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha}), \quad (4.1)$$

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

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

i.e., values at which L assumes a maximum for those $\underset{\sim}{\theta}$, $\underset{\sim}{\alpha}$ such that $\underset{\sim}{\theta} \in \Omega$. It is well-known that, for fixed $\underset{\sim}{\theta}$, L is maximized with respect to $\underset{\sim}{\alpha}$ by taking $\underset{\sim}{\alpha} = \hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta})$. Thus, putting

$$L_1(\underset{\sim}{\theta};\underset{\sim}{y}) = L[\underset{\sim}{\theta},\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta});\underset{\sim}{y}], \quad (4.2)$$

$\tilde{\underset{\sim}{\theta}}$ is a ML estimate of $\underset{\sim}{\theta}$ if and only if $\tilde{\underset{\sim}{\theta}} \in \underset{\sim}{\Omega}$ and $L_1(\tilde{\underset{\sim}{\theta}};\underset{\sim}{y}) = L_{sup}(\underset{\sim}{y})$, i.e., if and only if L_1 assumes a maximum at $\tilde{\underset{\sim}{\theta}}$ for $\underset{\sim}{\theta} \in \Omega$, in which case a ML estimate of $\underset{\sim}{\alpha}$ is $\hat{\underset{\sim}{\alpha}}(\tilde{\underset{\sim}{\theta}})$. Similarly, for fixed values of some number $\underline{a}$ of components of $\underset{\sim}{\theta}$, which without loss of generality we take to be the first $\underline{a}$ 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 $\underset{\sim}{\theta} \in \Omega$. Then, putting $L_2(\theta_1, \ldots, \theta_a; y) = L_1[\{\theta_1, \ldots, \theta_a, \tilde{\theta}_{a+1}(\theta_1, \ldots, \theta_a), \ldots, \tilde{\theta}_m(\theta_1, \ldots, \theta_a)\}; \underset{\sim}{y}]$, $\tilde{\theta}_1, \ldots, \tilde{\theta}_a$ 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 $\underset{\sim}{\theta} \in \Omega$ 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), $\underset{\sim}{\theta}$ is as indicated by (2.7), and, for fixed values of $\theta_1, \ldots, \theta_{m-1}$, L_1 is maximized for $\theta_m > 0$ by taking

$$\theta_m = (1/n)[\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta})]'[\underset{\sim}{I} + \textstyle\sum_{i=1}^{m-1}\theta_i \underset{\sim}{z}_i \underset{\sim}{z}_i']^{-1}[\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta})] \quad (4.3)$$

unless $\underset{\sim}{y}$ lies in the column space of $\underset{\sim}{X}$ (an event of probability zero when $n > p^*$). (The right hand side of (4.3) does not depend on θ_m since, as noted in Section 3, $\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta})$ does not depend on θ_m in this setting.) Except in certain fairly simple situations, it will be the case that $a \geq 1$; 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 the final solution.

Under what conditions does a ML estimate of $\underset{\sim}{\theta}$ exist; i.e., under what conditions is there a value of $\underset{\sim}{\theta}$ satisfying $\underset{\sim}{\theta} \in \Omega$ and $L_1(\tilde{\underset{\sim}{\theta}}; \underset{\sim}{y}) = L_{sup}(\underset{\sim}{y})$? Except for [3], [4], and [26], this

question does not seem to have received much attention. In [3] and [4], Anderson considered the case where $\underset{\sim}{y}$ is made up of s (= n/r) r-variate vectors that are independently and identically distributed, where the elements of the common covariance matrix of these vectors are linear combinations of the components of $\underset{\sim}{\theta}$, and where Ω consists of all $\underset{\sim}{\theta}$ such that the common covariance matrix is positive definite. He indicated that, for any fixed value of the common mean vector $\underset{\sim}{\mu}$ of these vectors, a sufficient (though not necessary) condition for the existence of a value of $\underset{\sim}{\theta}$ that maximizes L for $\underset{\sim}{\theta} \in \Omega$ is the nonsingularity of the matrix $(1/s) \sum_{i=1}^{s} (\underset{\sim}{y}_i - \underset{\sim}{\mu})(\underset{\sim}{y}_i - \underset{\sim}{\mu})'$, where $\underset{\sim}{y}_i$ represents the ith of the vectors. This result implies in particular that in the case where no linear structure is assumed for $\underset{\sim}{\mu}$, i.e., where $\underset{\sim}{\mu}$ is completely unspecified, a sufficient condition for the existence of a ML estimate of $\underset{\sim}{\theta}$ is the nonsingularity of the matrix $(1/s) \sum_{i=1}^{s} (\underset{\sim}{y}_i - \overline{\underset{\sim}{y}})(\underset{\sim}{y}_i - \overline{\underset{\sim}{y}})'$, where $\overline{\underset{\sim}{y}} = (1/s) \sum_{i=1}^{s} \underset{\sim}{y}_i$. These results leave unanswered the question of the existence of a ML estimate of $\underset{\sim}{\theta}$ for many cases of practical interest, including those associated with the ANOVA, because, for any partitioning of $\underset{\sim}{y}$ into independently and identically distributed subvectors, $\underset{\sim}{\mu}$ is ordinarily not known nor without linear structure and the sufficient condition may not hold anyhow. Hartley and Rao (Section 2 of [26]) gave conditions which were claimed to insure the existence of ML estimates for the variance components associated with the ordinary ANOVA models. Their conditions are quite unrestrictive.

(See Appendix D in [52] for some discussion of the Hartley-Rao conditions.)

4.2 Asymptotic Properties

Suppose again that $\underset{\sim}{y}$ is made up of s (=n/r) r-variate vectors $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_s$ that are independently and identically distributed, so that $\underset{\sim}{X}' = (\underset{\sim}{\chi}', \ldots, \underset{\sim}{\chi}')$ for some r × p matrix $\underset{\sim}{\chi}$ and $\underset{\sim}{V} = \text{diag}(\underset{\sim}{\Sigma}, \ldots, \underset{\sim}{\Sigma})$ for some r × r matrix $\underset{\sim}{\Sigma}$. Take $\underset{\sim}{\Lambda}'\underset{\sim}{\alpha}$ to be any p^* linearly independent estimable functions of $\underset{\sim}{\alpha}$. If $\underset{\sim}{\theta}^+$ is an interior point of Ω and if the functions of $\underset{\sim}{\theta}$ that comprise the elements of $\underset{\sim}{\Sigma}$ satisfy certain regularity conditions, then, with probability one, the likelihood equations for $\underset{\sim}{\Lambda}'\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ have a root, and that root is consistent and asymptotically efficient as $s \to \infty$ (with r fixed). Moreover, letting $\hat{\underset{\sim}{\theta}}$ represent the $\underset{\sim}{\theta}$-component of the root (implying that the $(\underset{\sim}{\Lambda}'\underset{\sim}{\alpha})$-component is $\underset{\sim}{\Lambda}'\hat{\underset{\sim}{\alpha}}(\hat{\underset{\sim}{\theta}})$), the limiting (again as $s \to \infty$ with r fixed) distribution of $\sqrt{s}\,\underset{\sim}{\Lambda}'[\hat{\underset{\sim}{\alpha}}(\hat{\underset{\sim}{\theta}}) - \underset{\sim}{\alpha}^+]$ and $\sqrt{s}(\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}^+)$ is normal with mean vector $\underset{\sim}{0}$ and covariance matrix $\text{diag}(\underset{\sim}{\Lambda}'\underset{\sim}{J}_1^{-}\underset{\sim}{\Lambda}, \underset{\sim}{J}_2^{-1})$, where $\underset{\sim}{J}_1 = \underset{\sim}{\chi}'[\underset{\sim}{\Sigma}(\underset{\sim}{\theta}^+)]^{-1}\underset{\sim}{\chi}$ and $\underset{\sim}{J}_2$ is the m × m matrix with ijth element

$$(\tfrac{1}{2})\ \text{tr}[\{\underset{\sim}{\Sigma}(\underset{\sim}{\theta}^+)\}^{-1}\{\partial\underset{\sim}{\Sigma}/\partial\theta_i\}\{\underset{\sim}{\Sigma}(\underset{\sim}{\theta}^+)\}^{-1}\{\partial\underset{\sim}{\Sigma}/\partial\theta_j\}],$$

where the partial derivatives are evaluated at $\underset{\sim}{\theta} = \underset{\sim}{\theta}^+$. (See, e.g., [4].)

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. The above asymptotic results can be applied with confidence if s is 'sufficiently large.' However, for many useful models of the form (2.1), $\underset{\sim}{y}$ 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 setup 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 [26] were the first to attempt an asymptotic theory that would be truly appropriate for the more complicated of the ANOVA models. They derived the limiting properties of the ML estimators as $n \to \infty$ and $q_i \to \infty$ $(i = 1, \ldots, c)$ simultaneously 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 [52] 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 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 $\underset{\sim}{Z}_i$ consist only of zeros 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 $\eta_{c+1} = n - \text{rank}(\underset{\sim}{Z})$, and assumed the existence of a function η_0 of n such that the matrix

$$\lim_{n \to \infty} \eta_0^{-1} \underset{\sim}{X}' [\underset{\sim}{V}(\underset{\sim}{\theta}^+)]^{-1} \underset{\sim}{X} \tag{4.4}$$

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 $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta} = (\sigma_1^2, \ldots, \sigma_{c+1}^2)$ 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 that 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 $\underset{\sim}{\alpha}$-component is $\hat{\underset{\sim}{\alpha}}(\hat{\underset{\sim}{\theta}})$ where $\hat{\underset{\sim}{\theta}}' = (\hat{\sigma}_1^2, \ldots, \hat{\sigma}_{c+1}^2)$), the limiting distribution of $\sqrt{\eta_0}[\hat{\underset{\sim}{\alpha}}(\hat{\underset{\sim}{\theta}}) - \underset{\sim}{\alpha}^+]$, $\sqrt{\eta_1}[\hat{\sigma}_1^2 - (\sigma_1^2)^+]$, $\ldots$, $\sqrt{\eta_{c+1}}[\hat{\sigma}_{c+1}^2 - (\sigma_{c+1}^2)^+]$ is normal with mean vector $\underset{\sim}{0}$ and covariance matrix $\text{diag}(\underset{\sim}{J}_1^{-1}, \underset{\sim}{J}_2^{-1})$, where $\underset{\sim}{J}_1$ is the matrix given by expression (4.4) and $\underset{\sim}{J}_2$ is the $(c+1) \times (c+1)$ matrix with ijth element

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

4.3 Restricted Maximum Likelihood

One criticism of the ML approach to the estimation of variance components is that the ML estimators of these parameters take no account of the loss in degrees of freedom resulting from estimating $\underset{\sim}{\alpha}$, the vector of fixed effects. In particular, in the ordinary fixed ANOVA or regression models where $q = 0$, $m = 1$, $\underset{\sim}{V} = \theta_1 \underset{\sim}{I}$, and $\Omega = \{\underset{\sim}{\theta} : \theta_1 > 0\}$ so that there is only a single 'variance component' θ_1, the ML estimator of θ_1 is

$$(1/n)(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}). \tag{4.5}$$

This estimator 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. In contrast, the ANOVA estimator

$$[1/(n-p^*)](\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}) \tag{4.6}$$

is unbiased. In many instances, the ANOVA estimator of θ_1 also compares favorably with the ML estimator on the basis of mean squared error (MSE). It is well-known that, if $p = 1$ and $\underset{\sim}{X} = \underset{\sim}{1}$, the ML estimator of θ_1 has uniformly smaller MSE than the ANOVA estimator. In fact, it is easy to show that the ML estimator has uniformly smaller MSE whenever $p^* \leq 4$. What is less well-known is that, for $p^* \geq 5$, the ANOVA estimator of θ_1 has uniformly smaller MSE than the ML estimator provided that

$n > p^*(p^*-2)/(p^*-4)$ (if $n = p^*(p^*-2)/(p^*-4)$, they have the same MSE, and, if $n < p^*(p^*-2)/(p^*-4)$, it is the ML estimator that has uniformly smaller MSE). In particular, for $p^* \geq 13$, the ANOVA estimator has the smaller MSE whenever there are more than two degrees of freedom. In conjunction with the above comparisons, it should be noted that the estimator

$$[1/(n-p^*+2)](\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}), \tag{4.7}$$

whose downward bias is 'only' $2\theta_1/(n-p^*+2)$, has uniformly smaller MSE than both the ML and ANOVA 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 $(1/k)(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})$.

Patterson and Thompson [55] proposed a restricted maximum likelihood approach to the problem of making inferences about the vector $\underset{\sim}{\theta}$ in the general linear model (2.1). The estimator of $\underset{\sim}{\theta}$ produced by their approach has the virtue that, for the ordinary fixed ANOVA or regression models, the estimator of θ_1 simplifies to the ANOVA estimator (4.6) rather than to the ML estimator (4.5). By an error contrast, we shall mean a linear combination $\underset{\sim}{u}'\underset{\sim}{y}$ of the observations such that $E(\underset{\sim}{u}'\underset{\sim}{y}) \equiv 0$, i.e., such that $\underset{\sim}{u}'\underset{\sim}{X} = \underset{\sim}{0}$ (where $\underset{\sim}{u}$ does not depend on $\underset{\sim}{\theta}$ or $\underset{\sim}{\alpha}$). 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 $\underset{\sim}{A}\underset{\sim}{y}$ where $\underset{\sim}{A}$ is a $(n-p^*) \times n$ matrix whose rows are any $n - p^*$

linearly independent rows of the matrix $\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-}\underset{\sim}{X}'$. Patterson and Thompson suggested that $\underset{\sim}{\theta}$ be estimated by applying the ML approach to $n - p^*$ linearly independent error contrasts rather than to the full data vector $\underset{\sim}{y}$. 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 contrasts are included but does not depend on $\underset{\sim}{\theta}$ or $\underset{\sim}{\alpha}$) from the function

$$L_1^*(\underset{\sim}{\theta};\underset{\sim}{y}) = -(\tfrac{1}{2})\log[\det(\underset{\sim}{V})] - (\tfrac{1}{2})\log[\det(\underset{\sim}{X}^{*\prime}\underset{\sim}{V}^{-1}\underset{\sim}{X}^{*})] - (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}), \quad (4.8)$$

defined for $\underset{\sim}{\theta} \in \Omega$ (see [31]). Thus, by definition, a restricted maximum likelihood (REML) estimator is a value of $\underset{\sim}{\theta}$ that maximizes L_1^* for $\underset{\sim}{\theta} \in \Omega$. As in full maximum likelihood, numerical techniques must crdinarily be employed to determine the estimate, though sometimes analytical techniques can be used to reduce the dimensions of the numerical problem. Suppose that, for fixed values of the first a components of $\underset{\sim}{\theta}$, it is 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 $\underset{\sim}{\theta} \in \Omega$. Then, putting

$L_2^*(\theta_1, \ldots, \theta_a; y)$
$= L_1^*[\{\theta_1, \ldots, \theta_a, \tilde{\theta}_{a+1}(\theta_1, \ldots, \theta_a), \ldots, \tilde{\theta}_m(\theta_1, \ldots, \theta_a)\}; \underset{\sim}{y}]$, $\tilde{\theta}_1, \ldots, \tilde{\theta}_a$ are REML estimates of $\theta_1, \ldots, \theta_a$ if and only if they

maximize L_2^* for those $\theta_1, \ldots, \theta_a$ that satisfy $\underset{\sim}{\theta} \in \Omega$ for some $(\theta_{a+1}, \ldots, \theta_m)$-value, in which case REML 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 the case of the ordinary ANOVA models with $\underset{\sim}{\theta}' = (\gamma_1, \ldots, \gamma_{c+1})$, L_1^* is maximized for fixed $\theta_1, \ldots, \theta_{m-1}$ and for $\theta_m > 0$ by taking

$$\theta_m = [1/(n-p^*)][\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta})]'[\underset{\sim}{I} + \textstyle\sum_{i=1}^{m-1}\theta_i \underset{\sim}{Z}_i \underset{\sim}{Z}_i']^{-1}[\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta})].$$

The arguments given above for favoring the REML approach over the full ML approach, which are based on bias and MSE considerations, differ somewhat from the argument advanced by Patterson and Thompson. Suppose that $\underset{\sim}{A}_1\underset{\sim}{y}$ is a vector of $n - p^*$ linearly independent error contrasts and that $\underset{\sim}{A}_2$ is any $p^* \times n$ matrix of constants such that $\underset{\sim}{A}' = (\underset{\sim}{A}_1', \underset{\sim}{A}_2')$ is nonsingular. The likelihood function for the transformed data vector $\underset{\sim}{A}\underset{\sim}{y}$ is proportional to that for $\underset{\sim}{y}$ so that we can just as well base our inferences on $\underset{\sim}{A}\underset{\sim}{y}$ as on $\underset{\sim}{y}$. Since $E(\underset{\sim}{A}_2\underset{\sim}{y})$ consists of estimable functions of the unknown vector $\underset{\sim}{\alpha}$, Patterson and Thompson argue that $\underset{\sim}{A}_2\underset{\sim}{y}$ contains 'no information' about $\underset{\sim}{\theta}$ and that therefore inferences for $\underset{\sim}{\theta}$ should be based only on the vector $\underset{\sim}{A}_1\underset{\sim}{y}$ of error contrasts. A similar, though perhaps more precise, argument is that the full ML estimator of $\underset{\sim}{\theta}$ necessarily depends on $\underset{\sim}{y}$ only through a set of $n - p^*$ linearly independent error contrasts, i.e., as a function of $\underset{\sim}{A}\underset{\sim}{y}$ it depends on $\underset{\sim}{A}_1\underset{\sim}{y}$ but not on $\underset{\sim}{A}_2\underset{\sim}{y}$, so that the REML approach does not ignore any information that is actually <u>used</u> by the full approach. (That the

full ML estimator of θ depends only on error contrasts follows immediately upon observing that it depends on $\underset{\sim}{y}$ only through the function $L_1(\underset{\sim}{\theta};\underset{\sim}{y})$ which in turn depends on $\underset{\sim}{y}$ only through $\underset{\sim}{A}_1\underset{\sim}{y}$.) A related observation has to do with translation invariance. (An estimator $\underset{\sim}{T}(\underset{\sim}{y})$ of a scalar- or vector-valued function of $\underset{\sim}{\theta}$ is called translation invariant if $\underset{\sim}{T}(\underset{\sim}{y}+\underset{\sim}{X}\underset{\sim}{a}) = \underset{\sim}{T}(\underset{\sim}{y})$ for all $\underset{\sim}{y}$ and all $p \times 1$ vectors $\underset{\sim}{a}$.) It is well-known that the REML estimator of $\underset{\sim}{\theta}$ 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.

SECTION 5

DERIVATION AND COMPUTATION OF DERIVATIVES AND OTHER RELEVANT ITEMS

In Section 6, we shall discuss various procedures for computing ML or REML estimates of $\underset{\sim}{\theta}$. All of these procedures are iterative procedures, requiring the repeated evaluation of items like the likelihood function, the first and second order partial derivatives of the likelihood function, and/or the expected values of the second order partial derivatives. To implement any one of these methods in the best possible way, we need to know how to evaluate the required items efficiently.

This information is also needed for purposes of choosing from among the methods in a given application. An important consideration in making such a choice is the amount of computation required by each method on a given iteration, which depends in turn on which of the above items are utilized by the method and on how they are to be computed.

It should be noted that, if we obtain the first and second order partial derivations of L, L_1, or L_1^* and the expected values of the partial derivatives for a particular parameterization of the model (2.1), then we can obtain the corresponding items for a second parameterization by making use of the chain rule of calculus (see, e.g., Section 5 in Chapter 6 of [53]). For example, in the case of the ordinary ANOVA models, partial derivatives taken with respect to $\sigma_1^2, \ldots, \sigma_{c+1}^2$ can be obtained from those taken with respect $\gamma_1, \ldots, \gamma_{c+1}$ and vice versa. In particular, general formulas for going from one information matrix to the other are given on p. 227 of [79]. 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^*.

For any function h and any row or column vectors $\underset{\sim}{u}$ and $\underset{\sim}{v}$, we use $\partial h/\partial \underset{\sim}{u}$ to denote the column vector whose ith element is the partial derivative of h with respect to the ith element of $\underset{\sim}{u}$ and $\partial^2 h/\partial \underset{\sim}{u} \partial \underset{\sim}{v}$ to denote the matrix whose ijth element is the second order partial derivative of h with respect to the ith element of $\underset{\sim}{u}$ and the jth element of $\underset{\sim}{v}$.

Using well-known results on matrix differentiation (see, e.g., Section 10.8 in [23] and Section 7 in Chapter 6 of [53]), we find

$$\partial L/\partial \underset{\sim}{\alpha} = \underset{\sim}{X}'\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha}), \tag{5.1}$$

$$\partial L/\partial \theta_i = -(\tfrac{1}{2})\ \mathrm{tr}[\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial \theta_i)] + (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha})'\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial \theta_i)\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha}), \tag{5.2}$$

$$E(\partial L/\partial \underset{\sim}{\alpha}) = \underset{\sim}{0}, \tag{5.3}$$

$$E(\partial L/\partial \theta_i) = 0, \tag{5.4}$$

$$\partial^2 L/\partial \underset{\sim}{\alpha}\partial \underset{\sim}{\alpha} = E(\partial^2 L/\partial \underset{\sim}{\alpha}\partial \underset{\sim}{\alpha}) = -\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X}, \tag{5.5}$$

$$\partial^2 L/\partial \underset{\sim}{\alpha}\partial \theta_i = -\underset{\sim}{X}'\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial \theta_i)\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha}), \tag{5.6}$$

$$\partial^2 L/\partial \theta_i \partial \theta_k = -(\tfrac{1}{2})\ \mathrm{tr}[V^{-1}\{(\partial^2 V/\partial \theta_i \partial \theta_k) - (\partial V/\partial \theta_i)V^{-1}(\partial V/\partial \theta_k)\}] + (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha})'\underset{\sim}{V}^{-1}[(\partial^2 \underset{\sim}{V}/\partial \theta_i \partial \theta_k) - 2(\partial \underset{\sim}{V}/\partial \theta_i)\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial \theta_k)]\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha}), \tag{5.7}$$

$$E(\partial^2 L/\partial \underset{\sim}{\alpha}\partial \theta_i) = \underset{\sim}{0}, \tag{5.8}$$

$$E(\partial^2 L/\partial \theta_i \partial \theta_k) = -(\tfrac{1}{2})\ \mathrm{tr}[\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial \theta_i)\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial \theta_k)], \tag{5.9}$$

$$\partial L_1/\partial \theta_i = -(\tfrac{1}{2})\ \mathrm{tr}[\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial \theta_i)] + (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial \theta_i)\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}), \tag{5.10}$$

$$\partial^2 L_1/\partial\theta_i\partial\theta_k = -(\tfrac{1}{2})\ \mathrm{tr}[\underset{\sim}{V}^{-1}\{(\partial^2\underset{\sim}{V}/\partial\theta_i\partial\theta_k) - (\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{V}^{-1}(\partial\underset{\sim}{V}/\partial\theta_k)\}] + (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'\underset{\sim}{V}^{-1}[(\partial^2\underset{\sim}{V}/\partial\theta_i\partial\theta_k) - 2(\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{P}(\partial\underset{\sim}{V}/\partial\theta_k)]\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}), \quad (5.11)$$

$$E(\partial^2 L_1/\partial\theta_i\partial\theta_k) = -(\tfrac{1}{2})\ \mathrm{tr}[\underset{\sim}{V}^{-1}\underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X})^{-}\underset{\sim}{X}'\underset{\sim}{V}^{-1}(\partial^2 V/\partial\theta_i\partial\theta_k)] - \underset{\sim}{V}^{-1}(\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{V}^{-1}(\partial\underset{\sim}{V}/\partial\theta_k) + 2\underset{\sim}{P}(\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{P}(\partial\underset{\sim}{V}/\partial\theta_k)], \quad (5.12)$$

$$\partial L_1^*/\partial\theta_i = -(\tfrac{1}{2})\ \mathrm{tr}[\underset{\sim}{P}(\partial V/\partial\theta_i)] + (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'\underset{\sim}{V}^{-1}(\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}), \quad (5.13)$$

$$E(\partial L_1^*/\partial\theta_i) = 0, \quad (5.14)$$

$$\partial^2 L_1^*/\partial\theta_i\partial\theta_k = -(\tfrac{1}{2})\ \mathrm{tr}[\underset{\sim}{P}\{(\partial^2\underset{\sim}{V}/\partial\theta_i\partial\theta_k) - (\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{P}(\partial\underset{\sim}{V}/\partial\theta_k)] + (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'\underset{\sim}{V}^{-1}[(\partial^2\underset{\sim}{V}/\partial\theta_i\partial\theta_k) - 2(\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{P}(\partial\underset{\sim}{V}/\partial\theta_k)]\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}), \quad (5.15)$$

and

$$E(\partial^2 L_1^*/\partial\theta_i\partial\theta_k) = -(\tfrac{1}{2})\ \mathrm{tr}[\underset{\sim}{P}(\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{P}(\partial\underset{\sim}{V}/\partial\theta_k)]. \quad (5.16)$$

The above expressions give the first and second order partial derivatives of L, L_1, and L_1^* and certain of their expected values in terms of $\partial \underset{\sim}{V}/\partial\theta_i$ and $\partial^2 \underset{\sim}{V}/\partial\theta_i\partial\theta_k$. These items can also be expressed in terms of $\partial \underset{\sim}{V}^{-1}/\partial\theta_i$ and $\partial^2 \underset{\sim}{V}^{-1}/\partial\theta_i\partial\theta_k$, but the latter expressions will not be given here.

Denote by $\underset{\sim}{B}$ the $m \times m$ matrix with ikth element $(\frac{1}{2})\ \mathrm{tr}[\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial\theta_i)\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial\theta_k)]$ and by $\underset{\sim}{B}^*$ the $m \times m$ matrix with ikth element $(\frac{1}{2})\ \mathrm{tr}[\underset{\sim}{P}(\partial \underset{\sim}{V}/\partial\theta_i)\underset{\sim}{P}(\partial \underset{\sim}{V}/\partial\theta_k)]$. From (5.5), (5.8), (5.9), and (5.16), we have that the matrix $\mathrm{diag}[\underset{\sim}{B}, (\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X})]$ is the information matrix associated with the full likelihood function (as observed by Searle [66]) and that $\underset{\sim}{B}^*$ is the information matrix associated with the likelihood function determined by any $n - p^*$ linearly independent error contrasts.

In practice, it is generally inefficient and possibly unfeasible computationally to evaluate L, L_1, and L_1^*, their partial derivatives, or the expected values of their second order partial derivatives directly from the right hand sides of (4.1), (4.2), (4.8), and (5.1) - (5.16). However, as noted in Section 3, the matrices $\underset{\sim}{R}$, $\underset{\sim}{D}$, and possibly $\underset{\sim}{Z}$ and $\underset{\sim}{X}$ often have relatively simple structures. We can derive formulas that make it clear how these structures can be exploited for purposes of computing the above items, just as we were able to develop formulas for exploiting them in the computation of BLUE's of linear combinations of the elements of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\varepsilon}$. To do so, we require the representations

$$\partial \underset{\sim}{V}/\partial\theta_i = \partial \underset{\sim}{R}/\partial\theta_i + \underset{\sim}{Z}(\partial \underset{\sim}{D}/\partial\theta_i)\underset{\sim}{Z}' + (\partial \underset{\sim}{Z}/\partial\theta_i)\underset{\sim}{D}\underset{\sim}{Z}' + \underset{\sim}{Z}\underset{\sim}{D}(\partial \underset{\sim}{Z}/\partial\theta_i)' \tag{5.17}$$

and

$$\begin{aligned}\partial^2 \underset{\sim}{V}/\partial\theta_i\partial\theta_k = {} & \partial^2 \underset{\sim}{R}/\partial\theta_i\partial\theta_k + \underset{\sim}{Z}(\partial^2 \underset{\sim}{D}/\partial\theta_i\partial\theta_k)\underset{\sim}{Z}' + (\partial \underset{\sim}{Z}/\partial\theta_k)(\partial \underset{\sim}{D}/\partial\theta_i)\underset{\sim}{Z}' \\ & + \underset{\sim}{Z}(\partial \underset{\sim}{D}/\partial\theta_i)(\partial \underset{\sim}{Z}/\partial\theta_k)' + (\partial^2 \underset{\sim}{Z}/\partial\theta_i\partial\theta_k)\underset{\sim}{D}\underset{\sim}{Z}' \\ & + (\partial \underset{\sim}{Z}/\partial\theta_i)(\partial \underset{\sim}{D}/\partial\theta_k)\underset{\sim}{Z}' + (\partial \underset{\sim}{Z}/\partial\theta_i)\underset{\sim}{D}(\partial \underset{\sim}{Z}/\partial\theta_k)' \\ & + (\partial \underset{\sim}{Z}/\partial\theta_k)\underset{\sim}{D}(\partial \underset{\sim}{Z}/\partial\theta_i)' + \underset{\sim}{Z}(\partial \underset{\sim}{D}/\partial\theta_k)(\partial \underset{\sim}{Z}/\partial\theta_i)' \\ & + \underset{\sim}{Z}\underset{\sim}{D}(\partial^2 \underset{\sim}{Z}/\partial\theta_i\partial\theta_k)',\end{aligned}$$

as well as certain of the well-known properties of the trace operation that are described in Section 9.1 of [23]. We also need the results in our Section 3 and the related identities

$$\det(\underset{\sim}{V}) \equiv \det(\underset{\sim}{R}) \cdot \det(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z}\underset{\sim}{D}), \tag{5.18}$$

$$\det(\underset{\sim}{V}) \cdot \det(\underset{\sim}{X}^{*\prime}\underset{\sim}{V}^{-1}\underset{\sim}{X}^{*}) \equiv \det(\underset{\sim}{R}) \cdot \det(\underset{\sim}{C}) \tag{5.19}$$

$$\equiv \det(\underset{\sim}{R}) \cdot \det(\underset{\sim}{X}^{*\prime}\underset{\sim}{R}^{-1}\underset{\sim}{X}^{*}) \cdot \det(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D}), \tag{5.20}$$

$$\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha}) \equiv \underset{\sim}{\cdot}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}),$$

$$\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}) \equiv \underset{\sim}{R}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^{*}) \equiv \underset{\sim}{S}(\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^{*}), \tag{5.21}$$

$$\underset{\sim}{P} \equiv \underset{\sim}{R}^{-1} - \underset{\sim}{R}^{-1}(\underset{\sim}{X}, \underset{\sim}{Z}\underset{\sim}{D})\underset{\sim}{C}^{-}(\underset{\sim}{X}, \underset{\sim}{Z})'\underset{\sim}{R}^{-1}, \tag{5.22}$$

$$\underset{\sim}{Z}'\underset{\sim}{P} \equiv (\underset{\sim}{0},\ \underset{\sim}{I}_{q\times q})\underset{\sim}{C}^{-}(\underset{\sim}{X},\ \underset{\sim}{Z})'\underset{\sim}{R}^{-1}, \tag{5.23}$$

where $\underset{\sim}{C}$ represents the coefficient matrix of the linear system (3.6). The identities (5.18) - (5.20) follow from the well-known results on the determinants of partitioned matrices described on page 165 in [23]. The identity (5.21) can be established by applying Theorem 2(i). Theorem 3 from [30] enters in the proofs of (5.22) and (5.23). Still another relevant result is that, for $\underset{\sim}{\theta} \in \Omega$ such that $\underset{\sim}{D}$ is nonsingular, a generalized inverse for $\underset{\sim}{C}$ is

$$\begin{bmatrix} \underset{\sim}{I} & \underset{\sim}{0} \\ \underset{\sim}{0} & \underset{\sim}{D}^{-1} \end{bmatrix} \begin{bmatrix} \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{X} & \underset{\sim}{X}'\underset{\sim}{R}^{-1}\underset{\sim}{Z} \\ \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{X} & \underset{\sim}{D}^{-1} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim}{Z} \end{bmatrix}^{-}.$$

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

$$L_1^* = -(\tfrac{1}{2})\log[\det(\underset{\sim}{R})] - (\tfrac{1}{2})\log[\det(\underset{\sim}{C})]$$
$$- (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'\underset{\sim}{R}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*) \tag{5.24}$$

$$= -(\tfrac{1}{2})\log[\det(\underset{\sim}{P})] - (\tfrac{1}{2})\log[\det(\underset{\sim}{X}^{*\prime}\underset{\sim}{R}^{-1}\underset{\sim}{X}^*)]$$
$$- (\tfrac{1}{2})\log[\det(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D})] - (\tfrac{1}{2})\underset{\sim}{y}'\underset{\sim}{S}(\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*) \tag{5.25}$$

follow immediately from (5.19) - (5.21). Next, consider $\partial L_1^*/\partial\theta_i$. Combining (5.13) with (5.21) and (3.3), we find

$$\partial L_1^*/\partial\theta_i = -(\tfrac{1}{2})\ \mathrm{tr}[\underset{\sim}{P}(\partial\underset{\sim}{R}/\partial\theta_i)]$$

$$-(\tfrac{1}{2})\ \mathrm{tr}[(\underset{\sim}{Z}'\underset{\sim}{P})\{\underset{\sim}{Z}(\partial\underset{\sim}{D}/\partial\theta_i)\underset{\sim}{Z} + 2(\partial\underset{\sim}{Z}/\partial\theta_i)\underset{\sim}{D}\}]$$

$$+(\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*)'\underset{\sim}{S}(\partial\underset{\sim}{R}/\partial\theta_i)\underset{\sim}{S}(\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*)$$

$$+(\tfrac{1}{2})\hat{\underset{\sim}{\Psi}}^{*\prime}(\partial\underset{\sim}{D}/\partial\theta_i)\hat{\underset{\sim}{\Psi}}^* + (\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*)'\underset{\sim}{S}(\partial\underset{\sim}{Z}/\partial\theta_i)\hat{\underset{\sim}{\beta}}^*,$$

where (3.13) or (5.22) and (3.14) or (5.23) can be substituted for $\underset{\sim}{P}$ and $\underset{\sim}{Z}'\underset{\sim}{P}$, respectively, and $\underset{\sim}{S}(\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*)$ could be replaced by $\underset{\sim}{R}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*)$. If $\underset{\sim}{D}$ depends on θ_i but $\underset{\sim}{R}$ and $\underset{\sim}{Z}$ do not (as is the case for $i = 1, \ldots, m-1$ in our formulations of the ordinary ANOVA models), then the above development simplifies immediately to

$$\partial L_1^*/\partial\theta_i = -(\tfrac{1}{2})\ \mathrm{tr}[(\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D})^{-1}\underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}(\partial\underset{\sim}{D}/\partial\theta_i)] + (\tfrac{1}{2})\hat{\underset{\sim}{\Psi}}^{*\prime}(\partial\underset{\sim}{D}/\partial\theta_i)\hat{\underset{\sim}{\Psi}}^*.$$

As a final illustration, we obtain suitable expressions for $\partial L_1^*/\partial\theta_i$, $E(\partial^2 L_1^*/\partial\theta_i\partial\theta_k)$, and $\partial^2 L_1^*/\partial\theta_i\partial\theta_k$ in the case of the ordinary ANOVA models, taking $\theta_i = \gamma_i$ $(i = 1, \ldots, m)$. Define the $q_i \times q_j$ matrix $\underset{\sim}{T}_{ij}^*$ by

$$\begin{bmatrix} \underset{\sim}{T}_{11}^* & \cdots & \underset{\sim}{T}_{1c}^* \\ \vdots & & \vdots \\ \underset{\sim}{T}_{c1}^* & \cdots & \underset{\sim}{T}_{c}^* \end{bmatrix} = (\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{S}\underset{\sim}{Z}\underset{\sim}{D})^{-1}.$$

and observe that (for the ordinary ANOVA models)

$$\underset{\sim}{T}^{*}_{ik} + \gamma_{c+1}\gamma_k \sum_{j=1}^{c} \underset{\sim}{T}^{*}_{ij}\underset{\sim}{Z}_j'\underset{\sim}{S}\underset{\sim}{Z}_k = \underset{\sim}{I}, \text{ if } k = i,$$

$$= \underset{\sim}{0}, \text{ if } k \neq i. \tag{5.26}$$

We find

$$\partial L_1^{*}/\partial\gamma_i = -(\tfrac{1}{2})\gamma_i^{-1}[q_i - \mathrm{tr}(\underset{\sim}{T}^{*}_{ii})] + (\tfrac{1}{2})\gamma_{c+1}\hat{\underset{\sim}{\Psi}}^{*\prime}_i\hat{\underset{\sim}{\Psi}}^{*}_i, \tag{5.27}$$

$$\partial L_1^{*}/\partial\gamma_{c+1} = -(\tfrac{1}{2})\gamma_{c+1}^{-1}[(n - p^{*}) - \underset{\sim}{y}'\underset{\sim}{S}(\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^{*})], \tag{5.28}$$

$$(-2)\ E(\partial^2 L_1^{*}/\partial\gamma_i\partial\gamma_i) = \gamma_i^{-2}\ \mathrm{tr}[(\underset{\sim}{I} - \underset{\sim}{T}^{*}_{ii})^2], \tag{5.29}$$

$$(-2)\ E(\partial^2 L_1^{*}/\partial\gamma_i\partial\gamma_k) = \gamma_i^{-1}\gamma_k^{-1}\ \mathrm{tr}[\underset{\sim}{T}^{*}_{ik}\underset{\sim}{T}^{*}_{ki}], \tag{5.30}$$

$$(-2)\ E(\partial^2 L_1^{*}/\partial\gamma_i\partial\gamma_{c+1}) = \gamma_i^{-1}\gamma_{c+1}^{-1}[q_i - \mathrm{tr}(\underset{\sim}{T}^{*}_{ii})], \tag{5.31}$$

$$(-2)\ E(\partial^2 L_1^{*}/\partial\gamma_{c+1}\partial\gamma_{c+1}) = \gamma_{c+1}^{-2}(n - p^{*}), \tag{5.32}$$

$$(-2)\ \partial^2 L_1^{*}/\partial\gamma_i\partial\gamma_i = -\gamma_i^{-2}\mathrm{tr}[(\underset{\sim}{I} - \underset{\sim}{T}^{*}_{ii})^2] + 2\gamma_{c+1}\gamma_i^{-1}\hat{\underset{\sim}{\Psi}}^{*}_i(\underset{\sim}{I} - \underset{\sim}{T}^{*}_{ii})\hat{\underset{\sim}{\Psi}}^{*}_i, \tag{5.33}$$

$$(-2)\ \partial^2 L_1^{*}/\partial\gamma_i\partial\gamma_k = -\gamma_i^{-1}\gamma_k^{-1}\mathrm{tr}[\underset{\sim}{T}^{*}_{ik}\underset{\sim}{T}^{*}_{ki}] - 2\gamma_{c+1}\gamma_k^{-1}\hat{\underset{\sim}{\Psi}}^{*\prime}_i\underset{\sim}{T}^{*}_{ik}\hat{\underset{\sim}{\Psi}}^{*}_k, \tag{5.34}$$

$$(-2)\ \partial^2 L_1^{*}/\partial\gamma_i\partial\gamma_{c+1} = \hat{\underset{\sim}{\Psi}}^{*\prime}_i\hat{\underset{\sim}{\Psi}}^{*}_i, \tag{5.35}$$

and

$$(-2)\ \partial^2 L_1^{*}/\partial\gamma_{c+1}\partial\gamma_{c+1} = -\gamma_{c+1}^{-2}[(n - p^{*}) - 2\underset{\sim}{y}'\underset{\sim}{S}(\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^{*})], \tag{5.36}$$

for $k \neq i = 1, \ldots, c$, where $\hat{\underset{\sim}{\Psi}}^{*}_i$ is the $q_i \times 1$ vector defined by $\hat{\underset{\sim}{\Psi}}^{*\prime} = (\hat{\underset{\sim}{\Psi}}^{*\prime}_1, \ldots, \hat{\underset{\sim}{\Psi}}^{*\prime}_c)$. In (5.27), (5.29) - (5.31), and (5.33) - (5.35), we require $\gamma_i > 0$ and, in (5.30) and (5.34),

we also require $\gamma_k > 0$. The representations (5.24), (5.25), and (5.27 - (5.36) are closely related to various of the representations given by Patterson and Thompson [55], by Henderson ([36] and [37]), and by Corbeil and Searle [13].

In certain special cases, $\underset{\sim}{V}$ can be inverted algebraically, leading to even 'simpler' expressions for L, L_1, and L_1^*, their partial derivatives, and the expected values of their second order partial derivatives. See, for example, [66] for an illustration of this approach in the case of the random two-way nested ANOVA model.

The above results completely link the problem described in Section 3 of estimating linear combinations of the elements of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\varepsilon}$ when $\underset{\sim}{\theta}^+$ (and possibly $\underset{\sim}{\alpha}^+$) are known to the problem of evaluating L, L_1, and L_1^*, their first and second order partial derivatives, and the expected values of their 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, that there is a consideration in the second problem that is not ordinarily present in the first. In the estimation of linear combinations of the elements of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\varepsilon}$ when $\underset{\sim}{\theta}^+$ is known, there is only a single set of computations, while, in the iterative procedures for the ML or REML estimation of $\underset{\sim}{\theta}$, similar computations must be performed for each of a sequence of $\underset{\sim}{\theta}$-values. When the computations must be carried out for more than one $\underset{\sim}{\theta}$-value, they should be accomplished in such a way

that, to the greatest extent possible, those operations that depend on $\underset{\sim}{\theta}$ are segregated from those that do not, so that the latter operations need be performed only once. Hemmerle and Hartley [33] discuss this point in the context of the ML estimation of variance components, and Corbeil and Searle [13] 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 (5.27) - (5.36), it would appear that, 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 like that provided in [78] can be invaluable.

SECTION 6

NUMERICAL PROCEDURES FOR MAXIMUM LIKELIHOOD ESTIMATION

Ordinarily, we must resort to an iterative numerical procedure in order to obtain a ML or REML estimate of $\underset{\sim}{\theta}$. However, there are simple cases where a ML or REML estimate of $\underset{\sim}{\theta}$ can be found by analytical means. Anderson [2] considered the question of the existence of explicit solutions to the likelihood equations (the equations $\partial L/\partial \underset{\sim}{\alpha} = \underset{\sim}{0}$ and $\partial L/\partial \theta_i = 0$, $i = 1, \ldots, m$) for those linear models where $\underset{\sim}{V}$ has the representation (2.8). He determined sufficient conditions for their existence. In fact, when Anderson's conditions are met, the likelihood equations are linear in the unknown parameters. Explicit solutions exist for the ordinary balanced two-way nested ANOVA models, though not for so simple a model as the ordinary balanced two-way crossed random-effects ANOVA model with interaction ([26], [52], and [39]).

There are many iterative numerical algorithms that can be regarded as candidates for computing ML or REML estimates of $\underset{\sim}{\theta}$. Some of them were developed specifically for special cases of that problem; e.g., some were developed for computing ML estimates of variance components. Others are general procedures for the numerical solution of broad classes of constrained nonlinear optimization problems. Certain of the latter procedures have long histories and are familiar to most statisticians, however there are also general procedures that are

relatively new, including some that have proved to be superior to the more traditional procedures in many applications.

Anyone familiar with the literature on nonlinear mathematical programming will realize that there is no real hope for finding a single iterative numerical algorithm for the ML or REML estimation of $\underset{\sim}{\theta}$ that will be best, or perhaps even satisfactory, for every application. An algorithm that converges quickly to a ML estimate in one setting may converge slowly or even fail to converge in another. The total computing time for a particular algorithm depends both on the amount of computation per iteration and the number of iterations to convergence. In general, algorithms that utilize a relatively small amount of information on each iteration will tend to require more iterations. In deciding on an algorithm for a particular application, we must make some judgments as to the computations per iteration and the convergence properties of the various procedures as applied to that setting. Any past experience in similar settings will of course be useful. When past experience is lacking or inconclusive, the judgments must be based on the characteristics of the various algorithms and, in the case of the amount of computation per iteration, on results like those given in Section 5. An obvious, but nevertheless very important, consideration is that, no matter how attractive a particular algorithm may appear to be for a given application, its usefulness is diminished if it is not readily available to the practitioner in a convenient form.

The present section is devoted to describing the various algorithms and their characteristics. The initial descriptions, given in Sub-Sections 6.1 and 6.2, ignore any complications brought about by constraints on the parameter space, i.e., by $\underset{\sim}{\theta}$ being confined to Ω when Ω is a proper subset of Euclidean m-space. In Sub-Section 6.3, several techniques are considered for modifying the various algorithms so as to cope with constraints.

6.1 Specialized Algorithms

On the kth iteration of an iterative algorithm for producing a ML or REML estimate of $\underset{\sim}{\theta}$, the current value for the estimate is converted into a new value. In what follows, we denote by $\underset{\sim}{\tilde{\theta}}^{(k)}$ the value produced by the algorithm on its kth iteration. Thus, $\underset{\sim}{\tilde{\theta}}^{(0)}$ represents the value used to start the algorithm. This value must of course be supplied by the user. Further, for any quantity f which is a function of $\underset{\sim}{\theta}$, we use $f^{(k)}$ to denote the value of f at $\underset{\sim}{\theta} = \underset{\sim}{\tilde{\theta}}^{(k)}$. For example, $\underset{\sim}{V}^{(k)} = \underset{\sim}{V}\{\underset{\sim}{\tilde{\theta}}^{(k)}\}$.

Anderson [6] and Henderson [36] have proposed iterative algorithms designed specifically to handle certain special cases of the problem of computing ML estimates of $\underset{\sim}{\theta}$. In both instances, the approach is in effect based on manipulating the equation $\partial L_1/\partial \underset{\sim}{\theta} = \underset{\sim}{0}$ into the form

$$\underset{\sim}{\theta} = \underset{\sim}{g}(\underset{\sim}{\theta}; \underset{\sim}{y}) \tag{6.1}$$

for some $m \times 1$ vector $\underset{\sim}{g}$ of functions of $\underset{\sim}{\theta}$. Nonlinear equations of this form can be solved by the method of successive approximations (see, e.g., Section 1.2 in [9]). As applied to (6.1), the method of successive approximations consists of taking $\tilde{\underset{\sim}{\theta}}^{(k+1)} = \underset{\sim}{g}\{\tilde{\underset{\sim}{\theta}}^{(k)}; \underset{\sim}{y}\}$. A general discussion of the convergence properties of the method of successive approximations can be found in [9].

Anderson's iterative algorithm for computing a ML estimate of $\tilde{\underset{\sim}{\theta}}$ is for the special case where $\underset{\sim}{V}$ has the representation (2.8). Anderson in effect found that $\partial L_1/\partial\theta_i$ has the representation

$$\partial L_1/\partial\theta_i = -(\tfrac{1}{2}) \sum_{j=1}^{m} \theta_j \operatorname{tr}[\underset{\sim}{V}^{-1}(\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{V}^{-1}(\partial\underset{\sim}{V}/\partial\theta_j)]$$

$$+(\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'\underset{\sim}{V}^{-1}(\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}).$$

Thus, setting $\partial L_1/\partial\underset{\sim}{\theta} = \underset{\sim}{0}$ yields the equation

$$\underset{\sim}{B}\underset{\sim}{\theta} = \underset{\sim}{d}, \tag{6.2}$$

where $\underset{\sim}{B}$ is as defined in Section 5 and $\underset{\sim}{d}$ is the $m \times 1$ vector whose ith element is

$$(\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'\underset{\sim}{V}^{-1}(\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}).$$

For fixed $\underset{\sim}{\theta}$ such that $\underset{\sim}{V}$ is nonsingular, $\underset{\sim}{B}$ is necessarily nonnegative and the linear system $\underset{\sim}{B}\tilde{\underset{\sim}{\theta}} = \underset{\sim}{d}$ is consistent for $\tilde{\underset{\sim}{\theta}}$ (follows from results in [46]). The matrix $\underset{\sim}{B}$ is nonsingular

(and thus positive definite) if and only if $G_1, \ldots, G_m$ are linearly independent matrices if and only if, for α^+ known, θ_i is estimable in the class of estimators of the form $(y - X\alpha^+)'A_i(y - X\alpha^+)$ $(i = 1, \ldots, m)$. When B is nonsingular, (6.2) is equivalent to the equation

$$\theta = B^{-1}d. \tag{6.3}$$

The method of successive approximations based on (6.3) is to take the (k+1)st iterate to be

$$\tilde{\theta}^{(k+1)} = [B^{(k)}]^{-1}d^{(k)}. \tag{6.4}$$

In the event that Anderson's sufficient conditions for the existence of an explicit solution to the likelihood equations are met, the iterative procedure converges in one iteration from any starting value $\theta^{(0)}$ [52].

A similar iterative algorithm can be constructed for computing a REML estimate of θ for the case where V has the representation (2.8). In analogy to (6.2), the likelihood equations for REML estimation yield

$$B^*\theta = d. \tag{6.5}$$

(B^* is the information matrix--see Section 5.) For fixed θ with V nonsingular, B^* is non-negative and the linear system $B^*\tilde{\theta} = d$ is consistent for $\tilde{\theta}$ (see pp. 316 and 327-8 in [46]). The matrix B^* is nonsingular if and only if θ_i is estimable in the class of quadratic translation-

invariant estimators for i = 1, ..., m (again see [46]), in which case (6.5) is equivalent to

$$\underset{\sim}{\theta} = \underset{\sim}{B}^{*-1}\underset{\sim}{d}. \tag{6.6}$$

Applying the method of successive approximations to (6.6) yields an iterative algorithm for computing a REML estimate of $\underset{\sim}{\theta}$ whose (k+1)st iterate is

$$\underset{\sim}{\tilde{\theta}}^{(k+1)} = [\underset{\sim}{B}^{*(k)}]^{-1}\underset{\sim}{d}^{(k)}. \tag{6.7}$$

If we alter the definition of $\underset{\sim}{d}$ by taking $\underset{\sim}{d}$ to be the $m \times 1$ vector with ith element

$$(\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\hat{\alpha}})'(\partial \underset{\sim}{V}^{-1}/\partial\theta_i)(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\hat{\alpha}}),$$

then (6.4) and (6.7) define suitable algorithms for computing ML estimates and REML estimates, respectively, of $\underset{\sim}{\theta}$ for those linear models where $\underset{\sim}{V}^{-1}$ has the representation (2.9).

Anderson's iterative algorithm (6.4) and its analogue (6.7) can of course be used to compute ML and REML estimates of the variance components σ_1^2, ..., σ_{c+1}^2 associated with the ordinary ANOVA models. However, Anderson's algorithm differs from the iterative algorithm proposed by Henderson [36]. Henderson's algorithm, which is the same in principle as the procedure proposed by Hartley and Rao in Section 5 of [26], is designed specifically for computing ML estimates of variance components. By using representations for $\partial L_1/\partial\gamma_i$ (i = 1, ..., c) and $\partial L_1/\partial\gamma_{c+1}$ analogous to the representations for $\partial L_1^*/\partial\gamma_i$ and

$\partial L_1^*/\partial \gamma_{c+1}$ given by (5.27) and (5.28), the equations $\partial L_1/\partial \gamma_i = 0$ $(i = 1, \ldots, c+1)$ and $\partial L_1/\partial \gamma_{c+1} = 0$ can be put in the form

$$\sigma_i^2 = [\hat{\underset{\sim}{\beta}}^{*\prime} \hat{\underset{\sim}{\beta}}_i^* + \sigma_i^2 \operatorname{tr}(\underset{\sim}{T}_{ii})]/q_i \qquad (6.8)$$

and

$$\sigma_{c+1}^2 = \underset{\sim}{y}'(\underset{\sim}{y} - \underset{\sim\sim}{X}\hat{\alpha} - \underset{\sim\sim}{Z}\hat{\underset{\sim}{\beta}}^*)/n, \qquad (6.9)$$

where $\hat{\underset{\sim}{\beta}}_i^*$ is the $q_1 \times 1$ vector defined by $\hat{\underset{\sim}{\beta}}^{*\prime} = (\hat{\underset{\sim}{\beta}}_1^{*\prime}, \ldots, \hat{\underset{\sim}{\beta}}_c^{*\prime})$ and the $q_i \times q_j$ matrix $\underset{\sim}{T}_{ij}$ is defined by

$$\begin{bmatrix} \underset{\sim}{T}_{11} & \cdots & \underset{\sim}{T}_{1c} \\ \vdots & & \vdots \\ \underset{\sim}{T}_{c1} & \cdots & \underset{\sim}{T}_{cc} \end{bmatrix} = (\underset{\sim}{I} + \underset{\sim}{Z}'\underset{\sim}{R}^{-1}\underset{\sim\sim}{ZD})^{-1}.$$

The representations (6.8) and (6.9) suggest the iterative procedure whose (k+1)st iteration produces

$$\{\tilde{\sigma}_i^2\}^{(k+1)} = [\hat{\underset{\sim}{\beta}}_i^{*(k)\prime} \hat{\underset{\sim}{\beta}}_i^{*(k)} + \{\tilde{\sigma}_i^2\}^{(k)} \operatorname{tr}\{\underset{\sim}{T}_{ii}^{(k)}\}]/q_i \qquad (6.10)$$

$(i = 1, \ldots, c)$ and

$$\{\tilde{\sigma}_{c+1}^2\}^{(k+1)} = \underset{\sim}{y}'[\underset{\sim}{y} - \underset{\sim\sim}{X}\hat{\alpha}^{(k)} - \underset{\sim\sim}{Z}\hat{\underset{\sim}{\beta}}^{*(k)}]/n$$

as values for the variance components. Note that the equation (6.8) can be rewritten as

$$\sigma_i^2 = [\hat{\underset{\sim}{\beta}}_i^{*\prime} \hat{\underset{\sim}{\beta}}_i^*]/[q_i - \operatorname{tr}(\underset{\sim}{T}_{ii})].$$

This representation suggests that an interesting modification of the Henderson procedure might be to replace (6.10) with

$$\{\tilde{\sigma}_i^2\}^{(k+1)} = [\hat{\underset{\sim}{\beta}}_i^{*(k)\prime}\hat{\underset{\sim}{\beta}}_i^{*(k)}]/[q_i - \mathrm{tr}\{\underset{\sim}{T}_{ii}^{(k)}\}]. \qquad (6.11)$$

The analogous algorithm for computing REML estimates of $\sigma_1^2, \ldots, \sigma_{c+1}^2$ is the procedure whose kth iterate is defined by

$$\{\tilde{\sigma}_i^2\}^{(k+1)} = [\hat{\underset{\sim}{\beta}}_i^{*(k)\prime}\hat{\underset{\sim}{\beta}}_i^{*(k)} + \{\tilde{\sigma}_i^2\}^{(k)}\ \mathrm{tr}\{\underset{\sim}{T}_{ii}^{*(k)}\}]/q_i \qquad (6.12)$$

or

$$\{\tilde{\sigma}_i^2\}^{(k+1)} = [\hat{\underset{\sim}{\beta}}_i^{*(k)\prime}\hat{\underset{\sim}{\beta}}_i^{*(k)}]/[q_i - \mathrm{tr}\{\underset{\sim}{T}_{ii}^{*(k)}\}] \qquad (6.13)$$

$(i = 1, \ldots, c)$ and

$$\{\tilde{\sigma}_{c+1}^2\}^{(k+1)} = \underset{\sim}{y}'[\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-}\underset{\sim}{X}'][\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^{*(k)}]/(n - p^*).$$

The following lemma is pertinent.

Lemma 1: For the ordinary ANOVA models (those for which $\underset{\sim}{R}$, $\underset{\sim}{D}$, and Ω have representations of the form (2.3), (2.4), and (2.6)); (i) $\mathrm{tr}(\underset{\sim}{T}_{ii}) > 0$; (ii) $\mathrm{tr}(\underset{\sim}{T}_{ii}^*) > 0$; (iii) $q_i \geq \mathrm{tr}(\underset{\sim}{T}_{ii})$ provided $\sigma_i^2 > 0$, with strict inequality holding if $\underset{\sim}{Z}_i \neq 0$; and (iv) $q_i \geq \mathrm{tr}(\underset{\sim}{T}_{ii}^*)$ provided $\sigma_i^2 > 0$, with strict inequality holding if $\mathrm{rank}(\underset{\sim}{X}, \underset{\sim}{Z}_i) > p^*$.

Parts (i) and (ii) of the lemma can be proved by using (3.17) and (3.18), respectively, together with the well-known result that is Theorem 12.2.1 in [23]. To prove part (iv), we observe that

$$(\sigma_i^2)^{-1}[q_i - \mathrm{tr}(\underset{\sim}{T}_{ii}^*)] = \mathrm{tr}[\mathrm{var}(\hat{\Psi}^*)] \geq 0$$

with equality holding if and only if $\sum_{j=1}^{c} \underset{\sim}{T}_{ij}^*\underset{\sim}{Z}_j'\underset{\sim}{S} = 0$. Using

(5.26), we find that $\sum_{j=1}^{c} \underset{\sim}{T}^{*}_{ij} \underset{\sim}{Z}_j' \underset{\sim}{S} = \underset{\sim}{0}$ only if, for all k, either $\underset{\sim}{Z}_i' \underset{\sim}{S} \underset{\sim}{Z}_k = \underset{\sim}{0}$ or $\sigma_k^2 = 0$. But $\underset{\sim}{Z}_i' \underset{\sim}{S} \underset{\sim}{Z}_i = \underset{\sim}{0}$ only when $\text{rank}(\underset{\sim}{X}, \underset{\sim}{Z}_i) = p^*$, and part (iv) of the lemma follows. Part (iii) of the lemma can be proved in similar fashion.

Lemma 1 implies that Henderson's iterative procedure for computing ML estimates of variance components and its analogue for computing REML estimates have an apparently pleasing property (which is not shared by Anderson's algorithm). Suppose that $\underset{\sim}{y}$ does not lie in the column space of $\underset{\sim}{X}$, which is the case with probability one when, e.g., $\underset{\sim}{y}$ 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 Lemma 1 is that the modified Henderson procedure in which (6.11) or (6.13) is used in place of (6.10) or (6.12) is well-defined, unless at some point a zero value is attained for some variance component. I.e., the denominators of (6.11) and (6.13) are strictly positive unless $\{\tilde{\sigma}_i^2\}^{(k)} = 0$ in which case the denominators are zero. The latter phenomenon causes no difficulty if we agree to take $\{\tilde{\sigma}_i^2\}^{(k+1)} = 0$ whenever

$\{\sigma_i^2\}^{(k)} = 0$. 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 defined by (6.11) or (6.13) 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 $\underset{\sim}{\beta}_i$ defined by $\underset{\sim}{\beta}' = (\underset{\sim}{\beta}_1', \ldots, \underset{\sim}{\beta}_c')$ and by then dividing by a number between q_i and zero.

6.2 General Algorithms

To locate a ML or REML estimate of $\underset{\sim}{\theta}$, 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 sub-section, some general algorithms and their properties are described, and references are indicated where more detailed information can be found. The discussion will be in terms of the problem of computing a REML estimate of $\underset{\sim}{\theta}$, i.e., the problem of computing a value of $\underset{\sim}{\theta}$ that maximizes L_1^*. This causes no real loss of generality, since the extensions to the problem of maximizing L with respect to $\underset{\sim}{\theta}$ and $\underset{\sim}{\alpha}$, to the problem of maximizing L_1 with respect to $\underset{\sim}{\theta}$, and more generally

to the problem of maximizing an arbitrary function will be obvious.

Many of the general iterative algorithms are gradient methods. A gradient method, as applied to the maximization of L_1^*, is one whose (k+1)st iterate takes the form

$$\tilde{\underset{\sim}{\theta}}^{(k+1)} = \tilde{\underset{\sim}{\theta}}^{(k)} + \rho_k \underset{\sim}{N}_k \{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)}, \tag{6.14}$$

where $\underset{\sim}{N}_k$ is a $m \times m$ matrix which, together with $\{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)}$, identifies the direction of search and ρ_k is a positive constant that serves to determine the distance traversed in the indicated search direction. The various gradient methods are characterized by different choices for $\underset{\sim}{N}_k$ and ρ_k. If $\underset{\sim}{N}_k$ is chosen to be positive definite, then necessarily there exists a ρ_k ($\rho_k > 0$) such that

$$L_1^*\{\tilde{\underset{\sim}{\theta}}^{(k+1)}; \underset{\sim}{y}\} > L_1^*\{\tilde{\underset{\sim}{\theta}}^{(k)}; \underset{\sim}{y}\}, \tag{6.15}$$

unless of course $\{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)} = \underset{\sim}{0}$. In fact, the inequality (6.15) holds for positive definite $\underset{\sim}{N}_k$ if ρ_k is taken to be sufficiently close to zero (see, e.g., [9]). With regard to the choice of ρ_k, the gradient methods fall into three categories: (i) $\rho_k = 1$; (ii) ρ_k equals (to some degree of approximation) μ_k, where μ_k is the value of the scalar ρ that maximizes

$$f_k(\rho) = L_1^*[\tilde{\underset{\sim}{\theta}}^{(k)} + \rho \underset{\sim}{N}_k \{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)}; \underset{\sim}{y}] \tag{6.16}$$

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 determining ρ_k in these cases can be found, e.g., in [17], [56], and [8]. The added computations are those associated with evaluating $f_k(\rho)$ at various trial values of ρ. Thus, these one-dimensional searches can be very time-consuming ((ii) more so than (iii)) in instances where the evaluation of L_1^* requires extensive computations.

When $\underset{\sim}{N}_k = \underset{\sim}{I}$ for all k, the iterative algorithm whose (k+1)st iterate is defined by (6.14) is the method of steepest ascent. Customary practice with this method would be to take $\rho_k = \mu_k$. The steepest ascent algorithm is one of the few that is supported by convergence theorems (see, e.g., [9] and [57]). Unfortunately, its rate of convergence is often found to be intolerably slow [57]. Bard [8] states that "the method is not recommended for practical applications." Hartley and Vaughn [27] 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.

Steepest ascent is an old optimization method. So is the Newton-Raphson procedure, which as applied to the maximization

of L_1^* is the gradient method whose (k+1)st iterate is defined by putting $N_k = \{J^{(k)}\}^{-1}$ and $\rho_k = 1$, where J is the m × m matrix whose ijth element is $-\partial^2 L_1^*/\partial\theta_i\partial\theta_j$. (It is assumed here that $J^{(k)}$ 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^* that 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., [8]). However, if the starting value is poor, it may converge to a stationary point that is not a local or global maximum and often does not converge at all [57]. 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 would be determined so that $L_1^*\{\tilde{\theta}^{(k+1)}; y\}$ is at least somewhat larger than $L_1^*\{\tilde{\theta}^{(k)}; y\}$. (If the directional derivative is negative in the direction $\{J^{(k)}\}^{-1}\{\partial L_1^*/\partial\theta\}^{(k)}$, the search direction $-\{J^{(k)}\}^{-1}\{\partial L_1^*/\partial\theta\}^{(k)}$, 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 $N_k = \{B^{*(k)}\}^{-1}$ and $\rho_k = 1$. Note that 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^* (refer to Section 4). 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 that, when applied to the maximization of L_1^*, the method

of scoring defines a search direction in which at least some increase in L_1^* can be achieved (provided $\{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)} \neq \underset{\sim}{0}$) since ordinarily $\underset{\sim}{B}^{*(k)}$ will be positive definite (see Section 6.1), which 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 $\underset{\sim}{\theta}$ and by then substituting $\hat{\underset{\sim}{\alpha}}\{\hat{\underset{\sim}{\theta}}^{(k)}\}$ for $\underset{\sim}{\alpha}$ is, in the case where $\underset{\sim}{V}$ has the linear representation (2.8), 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 [52]). Moreover, the iterates produced by the REML analogue of Anderson's algorithm are identical to those defined by applying the method of scoring to the maximization of L_1^* with respect to $\underset{\sim}{\theta}$ [40]. 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. Of course, the improvement is often achieved at the expense of more computations per iteration. A similar philosophy underlies the gradient method described in Section 5-8 of [8], which is based on the work of Levenberg [48], Marquardt [51], and Goldfeld, Quandt, and Trotter [22]. When applied to the maximization of L_1^*, the (k+1)st iterate

of the latter method is given by (6.14) with $\rho_k = 1$ and $\underset{\sim}{N}_k = [\underset{\sim}{A}_k + \lambda_k \underset{\sim}{M}_k]^{-1}$. Here, $\underset{\sim}{M}_k$ is a positive definite matrix, λ_k is a scalar that ordinarily is taken to be positive, and $\underset{\sim}{A}_k$ either equals $J^{(k)}$ (the negative of the Hessian matrix at $\underset{\sim}{\theta} = \tilde{\underset{\sim}{\theta}}^{(k)}$) or is some approximation to it. The matrix $\underset{\sim}{A}_k + \lambda_k \underset{\sim}{M}_k$ will be positive definite provided λ_k is taken to be sufficiently large (even if $\underset{\sim}{A}_k$ is indefinite). When $\underset{\sim}{A}_k = \underset{\sim}{J}^{(k)}$ and $\underset{\sim}{M}_k = \underset{\sim}{I}$, 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 $\underset{\sim}{M}_k$ is to take it to be the diagonal matrix whose diagonal elements are the absolute values of the diagonal elements of $\underset{\sim}{A}_k$, except that zeros are replaced by ones (see Section 5-8 in [8]). The scalar λ_k should be chosen so that $L_1^*\{\tilde{\underset{\sim}{\theta}}^{(k+1)}; \underset{\sim}{y}\} > L_1^*\{\tilde{\underset{\sim}{\theta}}^{(k)}; \underset{\sim}{y}\}$. Algorithms for determining a suitable λ_k are discussed in [8], [22], [51], [57], and [70]. 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 $\underset{\sim}{A}_k = \underset{\sim}{B}^{*(k)}$ rather than $\underset{\sim}{A}_k = \underset{\sim}{J}^{(k)}$. When $\underset{\sim}{A}_k = \underset{\sim}{B}^{*(k)}$, 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 has led to a relatively new class of gradient methods called variable metric methods. Like other of the algorithms, they are designed to simulate the performance of the Newton-Raphson algorithm in the vicinity of a maximizing value while improving on its very limited ability to converge to a maximizing value from a poor initial estimate. In terms of the maximization of L_1^*, variable metric methods are characterized by their use of an $\underset{\sim}{N}_k$ whose construction does not require second order partial derivatives (or their expected values) but which nevertheless approximates $\underset{\sim}{J}^{(k)}$ for sufficiently large k. That second order partial derivates need not be computed can be a valuable feature in many instances. The Davidon-Fletcher-Powell variable metric algorithm takes

$$\underset{\sim}{N}_{k+1} = \underset{\sim}{N}_k - (\underset{\sim}{\delta}_k'\underset{\sim}{\pi}_k)^{-1}\underset{\sim}{\delta}_k\underset{\sim}{\delta}_k' - (\underset{\sim}{\pi}_k'\underset{\sim}{N}_k\underset{\sim}{\pi}_k)^{-1}\underset{\sim}{N}_k\underset{\sim}{\pi}_k\underset{\sim}{\pi}_k'\underset{\sim}{N}_k,$$

where $\underset{\sim}{\delta}_k = \underset{\sim}{\tilde{\theta}}^{(k+1)} - \underset{\sim}{\tilde{\theta}}^{(k)}$ and $\underset{\sim}{\pi}_k = \{\partial L_1^*/\partial\underset{\sim}{\theta}\}^{(k+1)} - \{\partial L_1^*/\partial\underset{\sim}{\theta}\}^{(k)}$, and takes $\rho_k = \mu_k$. If $\underset{\sim}{N}_k$ in this method is positive definite, then necessarily so is $\underset{\sim}{N}_{k+1}$. Ordinarily, $\underset{\sim}{N}_0 = \underset{\sim}{I}$ so that the first step is a steepest ascent step, though some other positive definite matrix could be used for $\underset{\sim}{N}_0$. The method locates the maximum of a quadratic function having a negative definite Hessian matrix in a number of steps equal to the dimension of the search space. The Davidon-Fletcher-Powell algorithm has been widely used and has been very successful. Other variable metric algorithms are discussed in [8] and [57]. Some of

these methods take $\rho_k = 1$, eliminating the need for a one-dimensional search to locate μ_k.

The above compilation of optimization algorithms is by no means exhaustive. Other algorithms are discussed in references like [8] and [57]. Moreover, for most any method, it is easy to propose variations. Finite difference techniques can sometimes be used to approximate required derivatives (see Section 5-18 in [8]). In fact, Stewart [72] developed a successful modification of the Davidon-Fletcher-Powell variable metric algorithm in which finite difference approximations replace first order partial derivatives. Algorithms that automatically take $\rho_k = 1$ can be tried with $\rho_k = \mu_k$ or with ρ_k determined so that at least some increase in the objective function is achieved (as in going from the Newton-Raphson algorithm to the extended Newton-Raphson algorithm). Henderson's iterative algorithm for computing ML estimates of variance components could be modified in this way by using it to determine the search direction but not the search distance. Conversely, an algorithm like the Davidon-Fletcher-Powell algorithm which takes $\rho_k = \mu_k$ could be tried with $\rho_k = 1$. Consideration can also be given to switching from one algorithm to some other algorithm at various points during the iterative procedure.

6.3 Modifications to Accommodate Constraints on $\underset{\sim}{\theta}$

Ordinarily, the space Ω to which $\underset{\sim}{\theta}$ is constrained has a representation of the form

$$\Omega = \{\underset{\sim}{\theta} : g_1(\underset{\sim}{\theta})[>, \geq] 0, \ldots, g_d(\underset{\sim}{\theta})[>, \geq]0\} \qquad (6.17)$$

for some functions $g_1, \ldots, g_d$. 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 $g_i(\underset{\sim}{\theta}) = \theta_i$ $(i = 1, \ldots, c+1)$, and take the last inequality in (6.17) to be a strict inequality. The space Ω in our formulation of the ordinary MANOVA models can also be put in the form (6.17), though in a less obvious way. To do so, we put $d = r(c+1)$, let $g_1(\underset{\sim}{\theta}), \ldots, g_d(\underset{\sim}{\theta})$ represent $\lambda_{11}, \ldots, \lambda_{1r}, \ldots, \lambda_{c+1,1}, \ldots, \lambda_{c+1,r}$ where λ_{ij} denotes the jth largest eigenvalue of $\underset{\sim}{\Sigma}_i$, and take the last r of the inequalities in (6.17) to be strict inequalities: In what follows, we suppose that Ω has a representation of the form (6.17).

As noted in Section 6.2, Henderson's iterative algorithm for computing ML estimates of variance components and its REML analogue are not affected by the constraints on the parameter space; i.e., by the non-negativity constraints on the variance components. With it, negative components are simply never encountered. Unfortunately, none of the gradient algorithms discussed in Section 6.3 share this 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 [52].) Hemmerle and Hartley [33] encountered this difficulty in applying the Newton-Raphson method to the problem of computing ML estimates for σ^2_{c+1} 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 local maximum of L_1, let alone a ML estimate. (See the discussion in Section 6-3 of [8].) The same criticism applies, though to a lesser extent, to the approach taken by Miller [52] in using Anderson's procedure to compute ML estimates of variance components. He initially applies Anderson's procedure, disregarding the non-negativity constraints. If the procedure converges to a vector having one or more negative components, he restarts the algorithm with those components subsequently constrained to remain at zero. He continues to fix any zero components and to restart the algorithm until no negative values are 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 $\underset{\sim}{\theta}$ 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 available in references like [8] and [9]. At least three of these more or less meet our needs: (i) the penalty technique, (ii) the gradient projection technique, and (iii) the transformation technique.

There are actually several penalty techniques. The one proposed by Carroll [12] is well-suited for the ML or REML estimation of $\underset{\sim}{\theta}$ because it is an interior technique, i.e., it causes each iterate to lie in the interior of the constraint space. In terms of the maximization of L_1^*, Carroll's technique is to apply the unconstrained maximization algorithm to the function

$$L_1^*(\underset{\sim}{\theta}; \underset{\sim}{y}) - \Sigma_{j=1}^{d} \phi_j / g_j(\underset{\sim}{\theta}), \tag{6.18}$$

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 (6.18) 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 (6.18) 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 al thm is applied to (6.18) 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. In applying the algorithms described in Section 6.2 to the maximization of (6.18), we generally require first order, and sometimes even second order, partial derivatives for (6.18) and thus for $g_1, \ldots, g_d$, though Bard (Section 6-1 in [8]) suggests an approximation which eliminates the need for second order partial derivatives for $g_1, \ldots, g_d$. In the formulation of Ω for MANOVA models that was given earlier in this sub-section, $g_1, \ldots, g_d$ represent the totality of the eigenvalues of the matrices $\underset{\sim}{\Sigma}_1, \ldots, \underset{\sim}{\Sigma}_{c+1}$. Note that convenient expressions for the partial derivatives of eigenvalues can be obtained from [18].

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

$$g_i(\underset{\sim}{\theta}) = \underset{\sim i}{u}'\underset{\sim}{\theta} - c_i$$

for some $m \times 1$ vector $\underset{\sim i}{u}$ 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 $\underset{\sim i}{u}'\underset{\sim}{\theta} \geq c_i$ $(i = 1, \ldots, d)$, and that $\underset{\sim i}{u}$ has been normalized so that

$\underset{\sim}{u}_i'\underset{\sim}{u}_i = 1$. (A strict inequality $\underset{\sim}{u}_i'\underset{\sim}{\theta} > c_i$ can be approximated by the constraint $\underset{\sim}{u}_i'\underset{\sim}{\theta} \geq c_i + \varepsilon_i$, 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 $\underset{\sim}{u}_i'\tilde{\underset{\sim}{\theta}}^{(k)} > c_i$ for some values of i, and $\underset{\sim}{u}_i'\tilde{\underset{\sim}{\theta}}^{(k)} = c_i$ 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 in [20], [21], and [64]. Put $\underset{\sim}{U}_k = (\underset{\sim}{u}_{j_1}, \ldots, \underset{\sim}{u}_{j_\nu})$ where constraints $j_1, \ldots, j_\nu$ are the active constraints, and denote by $\underset{\sim}{\Lambda}_k$ the choice for $\underset{\sim}{N}_k$ in the unconstrained gradient algorithm. The gradient projection technique is to take the (k+1)st iterate to be (6.14) with

$$\underset{\sim}{N}_k = [\underset{\sim}{I} - \underset{\sim}{\Lambda}_k\underset{\sim}{U}_k(\underset{\sim}{U}_k'\underset{\sim}{\Lambda}_k\underset{\sim}{U}_k)^{-1}\underset{\sim}{U}_k']\underset{\sim}{\Lambda}_k$$

and with ρ_k restricted so that no constraint is violated but otherwise determined as in the unconstrained case. The gradient projection technique thus modifies the original search direction $\underset{\sim}{\Lambda}_k\{\partial L_1^*/\partial\underset{\sim}{\theta}\}^{(k)}$ by projecting it into the space determined by those vectors $\underset{\sim}{\theta}$ satisfying $\underset{\sim}{U}_k'\underset{\sim}{\theta} = 0$. Bard [8] has suggested a somewhat different approach to gradient projection based on quadratic programming techniques. Goldfarb ([20] and, together with Lapidus, [21]) has developed an algorithm for maximizing a function, subject to linear constraints, that bears the same relationship to the Newton-Raphson procedure,

when the latter procedure is modified by gradient projection, as the Davidon-Fletcher-Powell algorithm bears to the original Newton-Raphson procedure. The gradient projection technique can be extended to handle nonlinear constraints by piecewise linearization of such constraints [20]. The gradient projection 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 [8]. For example, the ordinary ANOVA models can be parameterized so that $\underset{\sim}{R} = \theta_m^2 \underset{\sim}{I}$, $\underset{\sim}{D} = \text{diag}[\theta_1^2 \underset{\sim}{I}, \ldots, \theta_{m-1}^2 \underset{\sim}{I}]$, and $\Omega = \{\underset{\sim}{\theta} : \theta_m \neq 0\}$. 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^*, 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 Vaugh [27]. One possible drawback in using this technique, say to compute REML estimates of variance components, 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. This rules out the Newton-Raphson algorithm and the method of scoring, though not the extended versions of those procedures.

6.4 Discussion

In a given application, it may be possible to improve the performance of the iterative optimization algorithms discussed in Section 6.2 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 [33] 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 $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ by applying the iterative optimization algorithms discussed in Section 6.2 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 $\underset{\sim}{\theta}$ by applying the iterative

algorithms to the reduced functions. Analytical reductions have proved to be useful in nonlinear least squares problems (see, e.g., [47]).

There is ordinarily no assurance that a value of $\underset{\sim}{\theta}$ 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 $\underset{\sim}{\theta}$. Even if such a $\underset{\sim}{\theta}$-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 $\underset{\sim}{\theta}$-value, we can be more confident that we have located a ML or REML estimate.

Certain of the algorithms that do not require the evaluation of second order partial derivatives or the expected values of second order partial derivatives may yield, as a by-product, approximations to those partial derivatives. When these algorithms are applied directly to the maximization of L or L_1^*, this output can be used to approximate the relevant information matrix. This approximation might be useful in an instance where the evaluation of the information matrix is very difficult relative to the evaluation of L or L_1^* and/or their first order partial derivatives.

Actual numerical experience in using the various iterative algorithms to compute ML or REML estimates of va iance components seems to be very limited, being largely confined to a

variation of the method of steepest ascent [27], to the Newton-Raphson procedure ([33] and [13]), and to Anderson's method [52].

SECTION 7
APPROXIMATING THE RESTRICTED MAXIMUM LIKELIHOOD APPROACH

Efficient computer programs for computing ML or REML estimates of $\underset{\sim}{\theta}$ can be devised by making use of the results outlined in Sections 3-6. Given the speed of modern computers, it is now possible to produce ML and REML estimates in many settings where their computation would once have been unthinkable. Nevertheless, there remain numerous situations where their computation is not possible. The latter situations are essentially those where the computations necessary to form and solve the linear system (3.6) are too extensive. In this section, we outline an approach to the estimation of $\underset{\sim}{\theta}$ 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 (4.8) 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 $\underset{\sim}{\theta}$ is essentially that of forming

the system of nonlinear equations $\partial L_1^*/\partial \underset{\sim}{\theta} = \underset{\sim}{0}$ and solving it for $\underset{\sim}{\theta}$. If the computations required to evaluate $\partial L_1^*/\partial \underset{\sim}{\theta}$ are too extensive, REML estimation cannot be undertaken. The equations $\partial L_1^*/\partial \underset{\sim}{\theta} = \underset{\sim}{0}$ consist in effect of m translation invariant quadratic forms set equal to their expectations. This observation suggests that, when the REML approach is unfeasible computationally, we proceed by equating some other set of m translation invariant quadratic forms to their expectations and solving for $\underset{\sim}{\theta}$, employing quadratic forms that resemble those used in REML estimation but which are easier to evaluate. The quadratic forms used in REML are

$$(\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'\underset{\sim}{V}^{-1}(\partial \underset{\sim}{V}/\partial\theta_i)\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}) \tag{7.1}$$

$$= (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*)'\underset{\sim}{R}^{-1}(\partial \underset{\sim}{V}/\partial\hat{\theta}_i)\underset{\sim}{R}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*) \tag{7.2}$$

$$= (\tfrac{1}{2})(\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*)'\underset{\sim}{S}(\partial \underset{\sim}{V}/\partial\theta_i)\underset{\sim}{S}(\underset{\sim}{y} - \underset{\sim}{Z}\hat{\underset{\sim}{\beta}}^*), \tag{7.3}$$

$i = 1, \ldots, m$. If $\underset{\sim}{D}$ depends on θ_i but $\underset{\sim}{R}$ and $\underset{\sim}{Z}$ do not, then the quadratic form (7.1) has the additional representation

$$-(\tfrac{1}{2})\hat{\underset{\sim}{\beta}}^{*\prime}(\partial \underset{\sim}{D}^{-1}/\partial\theta_i)\hat{\underset{\sim}{\beta}}^*, \tag{7.4}$$

provided $\underset{\sim}{\theta}$ is such that $\underset{\sim}{D}$ is nonsingular. One technique for 'approximating' the quadratic forms used in REML is to replace $\hat{\underset{\sim}{\alpha}}$ and/or $\hat{\underset{\sim}{\beta}}^*$ in (7.2), (7.3), or (7.4) with

$$\tilde{\underset{\sim}{\alpha}} = (\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{X})^{-}\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{y}$$

and

$$\tilde{\beta}^* = A(y - X\tilde{\alpha}) = A[I - X(X'HX)^{-}X'H]y,$$

where A is a $q \times n$ matrix and H is a $n \times n$ symmetric positive definite matrix, which must be specified. The elements of A and H may be functions of θ. Note that $E(\tilde{\beta}^*) \equiv 0$ and that, for any estimable function $\lambda'\alpha$, $E(\lambda'\tilde{\alpha}) \equiv \lambda'\alpha$. The matrices A and H should be chosen so that, for the case where θ^+ is known, $X\tilde{\alpha}$ and $\tilde{\beta}^*$ with $\theta = \theta^+$ are 'good' estimators of $X\alpha$ and β, but, at the same time, they must be such that $\tilde{\alpha}$ and $\tilde{\beta}^*$ are computable for any given θ-value. The representation (7.4) should not be used as a basis for approximating the quadratic form (7.1), unless D is nonsingular for all θ in Ω (which it is not in the case of the ordinary ANOVA models). In cases where R is hard to invert, we could replace R^{-1} in (7.2) or (7.3) with some positive definite matrix that is an 'approximation,' as well as substituting $\tilde{\alpha}$ and $\tilde{\beta}^*$ for $\hat{\alpha}$ and $\hat{\beta}^*$.

Suppose that we are going to take our estimate of θ to be the solution to the system of equations formed by setting a vector $Q(\theta; y) = (Q_1, \ldots, Q_m)'$ of translation invariant quadratic forms equal to its expectation. In particular, $Q_1, \ldots, Q_m$ can be the quadratic forms formed by substituting $\tilde{\alpha}$ and/or $\tilde{\beta}^*$ for $\hat{\alpha}$ and $\hat{\beta}^*$ in (7.2), (7.3), or (7.4). Put $G(\theta; y) = Q(\theta; y) - E(Q)$, let $\hat{\theta}$ represent the solution to $G(\theta; y) = 0$, and take $K(\theta; y)$ to be the $m \times m$ matrix whose jth column is $-\partial G/\partial\theta_j$. We have

$$\underset{\sim}{G}(\hat{\underset{\sim}{\theta}};\ \underset{\sim}{y}) = \underset{\sim}{G}(\underset{\sim}{\theta}^{+};\ \underset{\sim}{y}) - [\underset{\sim}{K}(\underset{\sim}{\theta}^{*};\ \underset{\sim}{y})](\hat{\underset{\sim}{\theta}} - \underset{\sim}{\theta}^{+}),$$

where $\underset{\sim}{\theta}^{*}$ is some point on the line segment between $\underset{\sim}{\theta}^{+}$ and $\hat{\underset{\sim}{\theta}}$. Using reasoning similar to that employed in deriving the asymptotic distribution of ML estimators (see, e.g., Section 5.5 in [79]),

$$[E(\underset{\sim}{K})]^{-1}[\mathrm{var}(\underset{\sim}{Q})][E(\underset{\sim}{K}')]^{-1}, \tag{7.5}$$

with $\underset{\sim}{\theta} = \underset{\sim}{\theta}^{+}$, may be a useful approximation to $\mathrm{var}(\hat{\underset{\sim}{\theta}})$ for 'large' samples, provided $E(\underset{\sim}{K})$ is nonsingular and $\underset{\sim}{\theta}^{+}$ is an interior point of Ω. When $Q_1, \ldots, Q_m$ are the quadratic forms used in REML rather than their approximations,

$$\mathrm{var}(\underset{\sim}{Q}) = E(\underset{\sim}{K}) = \underset{\sim}{B}^{*} \tag{7.6}$$

as is easily verified, and (7.5) simplifies to $\underset{\sim}{B}^{*-1}$.

There are several hazards in estimating $\underset{\sim}{\theta}$ by the approximate REML approach described above, i.e., by solving the equations $\underset{\sim}{G}(\underset{\sim}{\theta};\ \underset{\sim}{y}) = \underset{\sim}{0}$, where Q_i is given by (7.2), (7.3), or (7.4) with $\tilde{\underset{\sim}{\alpha}}$ and/or $\tilde{\underset{\sim}{\beta}}^{*}$ substituted for $\hat{\underset{\sim}{\alpha}}$ and $\hat{\underset{\sim}{\beta}}^{*}$. 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 $\underset{\sim}{G}(\underset{\sim}{\theta};\ \underset{\sim}{y}) = \underset{\sim}{0}$ 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 $\underset{\sim}{G}(\underset{\sim}{\theta}; \underset{\sim}{y}) = \underset{\sim}{0}$, we could proceed just as though we were maximizing a function whose gradient vector is $G(\underset{\sim}{\theta}; \underset{\sim}{y})$. 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. The final iterate would comprise our estimate of $\underset{\sim}{\theta}$. In implementing this approach, $\underset{\sim}{G}\{\underset{\sim}{\theta}^{(k)}; \underset{\sim}{y}\}$ would replace $\{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)}$, $\underset{\sim}{K}(\underset{\sim}{\theta}; \underset{\sim}{y})$ would play the role of $\underset{\sim}{J}$, and $E(\underset{\sim}{K})$ would be substituted for $\underset{\sim}{B}^*$ (in light of (7.6), an interesting variation would be to use $\text{var}(\underset{\sim}{Q})$ rather than $E(\underset{\sim}{K})$ in place of $\underset{\sim}{B}^*$). The function $f_k(\rho)$, defined in terms of L_1^* by (6.16), has as its derivative

$$[\{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)\prime} \underset{\sim}{N}_k' \underset{\sim}{N}_k \{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)}]^{-\frac{1}{2}} \cdot \partial L_1^*/\partial \underset{\sim}{\theta} \Big|_{\underset{\sim}{\theta}=\underset{\sim}{\theta}^{(k)}+\rho \underset{\sim}{N}_k \{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)}} \underset{\sim}{N}_k \{\partial L_1^*/\partial \underset{\sim}{\theta}\}^{(k)} \qquad (7.7)$$

(see, e.g., Section 1.3 in [9]), and

$$f_k(\rho) = f_k(0) + \int_0^\rho (\partial f_k/\partial \rho)\, \partial \rho. \qquad (7.8)$$

The relevance of the representations (7.7) and (7.8) is that the function to be maximized in the approximate REML approach is defined only in terms of its partial derivatives, so that the representation (6.16) does not extend directly. Rather, we must use the right hand side of (7.8), together with (7.7), to achieve the extension. The value of $f_k(0)$ is irrelevant for purposes of determining the distance to be traversed in the indicated search direction, so that $f_k(0)$ can be set equal to an arbitrary constant. The integral on the right hand side of (7.8) can be approximated by numerical integration techniques.

The translation invariant quadratic forms $Q_1, \ldots, Q_m$ have representations $Q_i = \underset{\sim}{y}'\underset{\sim}{\Gamma}_i\underset{\sim}{y}$ $(i = 1, \ldots, m)$ for some $n \times n$ symmetric matrices $\underset{\sim}{\Gamma}_1, \ldots, \underset{\sim}{\Gamma}_m$. For example, when Q_i is taken to be (7.2) with $\tilde{\underset{\sim}{\alpha}}$ and $\tilde{\underset{\sim}{\beta}}^*$ substituted for $\hat{\underset{\sim}{\alpha}}$ and $\hat{\underset{\sim}{\beta}}^*$,

$$\underset{\sim}{\Gamma}_i = [\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{X})^{-}\underset{\sim}{X}'\underset{\sim}{H}]'[\underset{\sim}{I} - \underset{\sim}{A}'\underset{\sim}{Z}']\underset{\sim}{R}^{-1}$$

$$\cdot (\partial\underset{\sim}{V}/\partial\theta_i)\underset{\sim}{R}^{-1}[\underset{\sim}{I} - \underset{\sim}{Z}\underset{\sim}{A}][\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{H}\underset{\sim}{X})^{-}\underset{\sim}{X}'\underset{\sim}{H}]. \qquad (7.9)$$

In applying various of the gradient algorithms in our approximate REML approach, we must, on each iteration, evaluate Q_i, $E(Q_i)$, $\mathrm{cov}(Q_i, Q_j)$, $E(\partial Q_i/\partial\theta_j)$, $\partial[E(Q_i)]/\partial\theta_j$, and/or $\partial Q_i/\partial\theta_j$ $(i, j = 1, \ldots, m)$. In evaluating these items, we can make use of the representations

$$E(Q_i) = \mathrm{tr}(\underset{\sim}{\Gamma}_i\underset{\sim}{V}) = \mathrm{tr}(\underset{\sim}{\Gamma}_i\underset{\sim}{R}) + \mathrm{tr}(\underset{\sim}{Z}'\underset{\sim}{\Gamma}_i\underset{\sim}{Z}\underset{\sim}{D}), \qquad (7.10)$$

$$\operatorname{cov}(Q_i, Q_j) = 2\operatorname{tr}(\underset{\sim}{\Gamma}_i\underset{\sim}{V}\underset{\sim}{\Gamma}_j\underset{\sim}{V})$$

$$= 2\operatorname{tr}(\underset{\sim}{\Gamma}_i\underset{\sim}{R}\underset{\sim}{\Gamma}_j\underset{\sim}{R}) + 2\operatorname{tr}(\underset{\sim}{Z}'\underset{\sim}{\Gamma}_i\underset{\sim}{Z}\underset{\sim}{D}\underset{\sim}{Z}'\underset{\sim}{\Gamma}_j\underset{\sim}{Z}\underset{\sim}{D})$$

$$+ 4\operatorname{tr}(\underset{\sim}{Z}'\underset{\sim}{\Gamma}_i\underset{\sim}{R}\underset{\sim}{\Gamma}_j\underset{\sim}{Z}\underset{\sim}{D}), \qquad (7.11)$$

$$E(\partial Q_i/\partial\theta_j) = \operatorname{tr}[(\partial\underset{\sim}{\Gamma}_i/\partial\theta_j)\underset{\sim}{V}]$$

$$= \operatorname{tr}[(\partial\underset{\sim}{\Gamma}_i/\partial\theta_j)\underset{\sim}{R}] + \operatorname{tr}[\underset{\sim}{Z}'(\partial\underset{\sim}{\Gamma}_i/\partial\theta_j)\underset{\sim}{Z}\underset{\sim}{D}], \qquad (7.12)$$

and

$$\partial[E(Q_i)]/\partial\theta_j = \operatorname{tr}[(\partial\underset{\sim}{\Gamma}_i/\partial\theta_j)\underset{\sim}{R}] + \operatorname{tr}[\underset{\sim}{\Gamma}_i(\partial\underset{\sim}{R}/\partial\theta_j)]$$

$$+ \operatorname{tr}[\underset{\sim}{Z}'(\partial\underset{\sim}{\Gamma}_i/\partial\theta_j)\underset{\sim}{Z}\underset{\sim}{D}] + \operatorname{tr}[\underset{\sim}{Z}'\underset{\sim}{\Gamma}_i\underset{\sim}{Z}(\partial\underset{\sim}{D}/\partial\theta_j)] \qquad (7.13)$$

(in the last representation, it is assumed that $\underset{\sim}{Z}$ does not depend on θ_j). The matrices $\underset{\sim}{A}$ and $\underset{\sim}{H}$ should be chosen so that, after exploiting any simple structures inherent in the $\underset{\sim}{R}$, $\underset{\sim}{D}$, $\underset{\sim}{Z}$ and $\underset{\sim}{X}$ matrices, the evaluation of the necessary items is manageable. In particular, when $\underset{\sim}{\Gamma}_i$ is given by (7.9), some simplification can be achieved (e.g., in the evaluation of $\operatorname{tr}(\underset{\sim}{\Gamma}_i\underset{\sim}{R})$) by taking $\underset{\sim}{H}$ to be $\underset{\sim}{R}^{-1}$. The evaluation of $E(\underset{\sim}{Q})$, $\operatorname{var}(\underset{\sim}{Q})$, $E(\partial\underset{\sim}{Q}/\partial\theta_j)$, and/or $\partial[E(Q_i)]/\partial\theta_j$ on the basis of the right hand sides of (7.10) - (7.13) closely parallels Hartley's method of synthesis (see [63]).

In instances where the computations required by a single iteration of an iterative procedure for computing a REML estimate of $\underset{\sim}{\theta}$ are found to be extensive but manageable, we might consider a different approach to approximating the REML

procedure. We could simply stop the iterative process after some reasonable number of iterations, even though convergence may not have been completed. The resulting estimation procedure might be nearly as satisfactory as the complete REML procedure. We could also consider stoping short of convergence in the iterative procedure that uses approximations to the REML quadratic forms.

SECTION 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 $\underset{\sim}{\theta}$ when $\underset{\sim}{V}$ has the representation (2.8), 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 ([76] and, together with Searle, [77]). He 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 $\underset{\sim}{y}$ is normal and the mean vector is $\underset{\sim}{0}$. Harville [28] 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 are left in very inconvenient form. These early efforts were generalized and greatly improved upon by LaMotte ([44], [45], and [46]). LaMotte's results apply to all linear models for which $\underset{\sim}{V}$ has the representation (2.8), though he did assume normality. He considered several classes of estimators for a linear function $\underset{\sim}{\lambda}'\underset{\sim}{\theta}$ 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 $\underset{\sim}{\theta} = \underset{\sim}{\theta}^*$ is $\underset{\sim}{\lambda}'\hat{\underset{\sim}{\theta}}$ where $\hat{\underset{\sim}{\theta}}$ is any solution to the linear system

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

(provided that $\underset{\sim}{\lambda}'\underset{\sim}{\theta}$ is estimable in the class of translation-invariant quadratic estimators which is the case if and only if the equations $[\underset{\sim}{B}^*(\underset{\sim}{\theta}^*)]\underset{\sim}{\tau} = \underset{\sim}{\lambda}$ have a solution for $\underset{\sim}{\tau}$). Rao ([61] and [62]) 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 $\underset{\sim}{\lambda}'\underset{\sim}{\theta}$ (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 (p. 783 in [65]), 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 in Section 3.3.5 of [29], such comparisons are potentially misleading.

The traditional approach to the estimation of $\underset{\sim}{\theta}$, when $\underset{\sim}{V}$ has the representation (2.8) as in the case of the ordinary ANOVA models, is to equate m translation-invariant quadratic forms (that are not functionally dependent on $\underset{\sim}{\theta}$) to their expectations and to solve the resulting linear system for $\underset{\sim}{\theta}$. The ith of the likelihood equations $\partial L/\partial\underset{\sim}{\theta} = \underset{\sim}{0}$ depends on the data only through the quadratic form $(\frac{1}{2})(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha})'\underset{\sim}{V}^{-1}\underset{\sim}{G}_i\underset{\sim}{V}^{-1}(\underset{\sim}{y} - \underset{\sim}{X}\underset{\sim}{\alpha})$. Suppose that, in this quadratic form, we substitute $\hat{\underset{\sim}{\alpha}}(\underset{\sim}{\theta})$ for $\underset{\sim}{\alpha}$ and then replace $\underset{\sim}{\theta}$ with a fixed value $\underset{\sim}{\theta}^*$. The result is a translation-invariant quadratic form that is functionally independent of $\underset{\sim}{\theta}$. LaMotte [44] considered estimating $\underset{\sim}{\theta}$ by equating the m translation-invariant quadratic forms generated in this way to their expectation. 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 ([59], [60], and [62]) 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, i.e., those models for which $\underset{\sim}{\theta}$, $\underset{\sim}{R}$, and $\underset{\sim}{D}$ are given by (2.2) - (2.4). Rao observed in effect that, if $\underset{\sim}{\beta}$ and $\underset{\sim}{\varepsilon}$ were known, a natural estimator for $\underset{\sim}{\lambda}'\underset{\sim}{\theta} = \sum_{i=1}^{c+1} \lambda_i \sigma_i^2$ would be

$$(\lambda_{c+1}/n)\,\underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} + \sum_{i=1}^{c}(\lambda_i/q_i)\underset{\sim}{\beta}_i'\underset{\sim}{\beta}_i = \underset{\sim}{\omega}'\underset{\sim}{\Delta}\underset{\sim}{\omega}, \tag{8.2}$$

where $\underset{\sim}{\beta}_i$ is the $q_i \times 1$ vector defined by $\underset{\sim}{\beta}' = (\underset{\sim}{\beta}_1', \ldots, \underset{\sim}{\beta}_c')$, $\omega' = (\underset{\sim}{\beta}', \underset{\sim}{\varepsilon}')$, and $\underset{\sim}{\Delta}$ is a suitably defined matrix. Since $\underset{\sim}{\beta}$ and $\underset{\sim}{\varepsilon}$ are in fact unknown, Rao suggested estimating $\Sigma_i \lambda_i \sigma_i^2$ by the translation-invariant quadratic unbiased estimator that most closely resembles (8.2). More precisely, observing that $\underset{\sim}{y}'\underset{\sim}{\Gamma}\underset{\sim}{y} = \underset{\sim}{\omega}'\underset{\sim}{U}'\underset{\sim}{\Gamma}\underset{\sim}{U}\underset{\sim}{\omega}$, with $\underset{\sim}{U} = (\underset{\sim}{Z}, \underset{\sim}{i})$, for any translation-invariant quadratic estimator $\underset{\sim}{y}'\underset{\sim}{\Gamma}\underset{\sim}{y}$, he proposed the estimator $\underset{\sim}{y}'\underset{\sim}{\Gamma}^*\underset{\sim}{y}$ where $\underset{\sim}{\Gamma}^*$ minimizes $\|\underset{\sim}{U}'\underset{\sim}{\Gamma}\underset{\sim}{U} - \underset{\sim}{\Delta}\|$ for $\underset{\sim}{\Gamma}$ such that $\underset{\sim}{y}'\underset{\sim}{\Gamma}\underset{\sim}{y}$ is a translation-invariant quadratic unbiased estimator of $\Sigma_i \lambda_i \sigma_i^2$. Here, $\|\cdot\|$ denotes a matrix norm. It can be shown that, when the Euclidean norm is used, $\underset{\sim}{y}'\underset{\sim}{\Gamma}^*\underset{\sim}{y} = \underset{\sim}{\lambda}'\hat{\underset{\sim}{\theta}}$ where, with $\underset{\sim}{\theta}^* = \underset{\sim}{1}$, $\hat{\theta}$ is a solution to the linear system (8.1). Rao went on to observe that the difference between a translation-invariant quadratic estimator $\underset{\sim}{y}'\underset{\sim}{\Gamma}\underset{\sim}{y}$ and $\underset{\sim}{\omega}'\underset{\sim}{\Delta}\underset{\sim}{\omega}$ can be expressed as $\underset{\sim}{\eta}'\underset{\sim}{\Lambda}(\underset{\sim}{U}'\underset{\sim}{\Gamma}\underset{\sim}{U} - \underset{\sim}{\Delta})\underset{\sim}{\Lambda}\underset{\sim}{\eta}$, where $\underset{\sim}{\Lambda} = \mathrm{diag}(\sigma_1\underset{\sim}{I}, \ldots, \sigma_{c+1}\underset{\sim}{I})$ and $\underset{\sim}{\eta}$ represents the standardized vector $\underset{\sim}{\Lambda}^{-1}\underset{\sim}{\omega}$. Taking $\underset{\sim}{\Lambda}^*$ to be the value of $\underset{\sim}{\Lambda}$ at $\underset{\sim}{\theta} = \underset{\sim}{\theta}^*$ where the of $\underset{\sim}{\theta}^*$ can be based on prior information, we could also consider

estimating $\Sigma_i \lambda_i \sigma_i^2$ by $\underset{\sim}{y}'\underset{\sim}{\Gamma}^*\underset{\sim}{y}$, where now $\underset{\sim}{\Gamma}^*$ minimizes $\| \underset{\sim}{A}(\underset{\sim}{U}'\underset{\sim}{\Gamma}\underset{\sim}{U} - \underset{\sim}{\Delta})\underset{\sim}{A}\|$ for translation-invariant quadratic unbiased estimators $\underset{\sim}{y}'\underset{\sim}{\Gamma}\underset{\sim}{y}$. Again, when the Euclidean norm is employed, it can be shown that $\underset{\sim}{y}'\underset{\sim}{\Gamma}^*\underset{\sim}{y} = \underset{\sim}{\lambda}'\hat{\underset{\sim}{\theta}}$ where $\hat{\underset{\sim}{\theta}}$ 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 $\Sigma_i \lambda_i \sigma_i^2$ (based on a Euclidean-norm) is the same as a MIVQUE (derived on the basis of the normality assumption).

Several observers (see, e.g., [29], [44], and [62]) 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 non-convergence 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 [40]. Note that the iterates produced by the iterative MIVQUE procedure are the same as those defined by the REML analogue 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 $\underset{\sim}{\lambda}'\underset{\sim}{\theta}$, i.e., an estimator which, among all translation-invariant quadratic unbiased estimators of $\underset{\sim}{\lambda}'\underset{\sim}{\theta}$, has uniformly

(for all $\theta \in \Omega$) minimum variance. Then, every MIVQUE of $\underset{\sim}{\lambda}'\underset{\sim}{\theta}$ is a UMIVQUE, implying that $\underset{\sim}{\lambda}'\underset{\sim}{B}^{*-}\underset{\sim}{d}$ is functionally independent of $\underset{\sim}{\theta}$. Taking $\hat{\underset{\sim}{\theta}}$ to be any REML estimate of $\underset{\sim}{\theta}$, it follows that $\underset{\sim}{\lambda}'\hat{\underset{\sim}{\theta}}$ agrees with the UMIVQUE of $\underset{\sim}{\lambda}'\underset{\sim}{\theta}$, provided that $\hat{\underset{\sim}{\theta}}$ satisfies $\partial L_1^*/\partial\underset{\sim}{\theta} = \underset{\sim}{0}$ as would necessarily be the case if $\hat{\underset{\sim}{\theta}}$ were an interior point of Ω. Moreover, if there is a UMIVQUE of every component of $\underset{\sim}{\theta}$, then $\underset{\sim}{B}^{*-1}\underset{\sim}{d}$ is functionally independent of $\underset{\sim}{\theta}$, so that there is a $\underset{\sim}{\theta} \in \Omega$ satisfying the REML equations $\partial L_1^*/\partial\underset{\sim}{\theta} = \underset{\sim}{0}$ if and only if $\underset{\sim}{B}^{*-1}\underset{\sim}{d} \in \Omega$, in which case the REML equations admit the explicit solution $\underset{\sim}{\theta} = \underset{\sim}{B}^{*-1}\underset{\sim}{d}$, and the REML analogue of Anderson's iterative procedure converges (to the UMIVQUE of $\underset{\sim}{\theta}$) in a single iteration.

Some results are available in the literature on the existence of uniformly minimum variance quadratic unbiased estimators of variance components (see, e.g., [19] and [24]). It is known in particular that such estimators exist for all four of the variance components associated with the balanced two-way crossed random-effects ANOVA model with interaction (they are the ANOVA estimators). Thus, in the case of this model, the REML equations admit an explicit solution and the REML analogue of Anderson's algorithm converges in a single iteration, even though the ML equations do not have an explicit solution and the ML version of Anderson's algorithm need not converge in one iteration (refer to Section 6).

Rao's MINQUE approach does not require any normality assumptions, nor is its intuitive appeal diminished by

non-normality. 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 $\underset{\sim}{\theta}$ derived under normality assumptions are reasonable estimators even when the form of the distribution of $\underset{\sim}{b}$ and $\underset{\sim}{e}$ is unspecified.

8.2 Henderson's Methods

The most commonly used methods for estimating variance components are the Methods 1, 2, and 3 set forth by Henderson in [34]. 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 ([65], [67], and [68]) gave excellent descriptions of Henderson's methods and indicated various generalizations. Henderson's methods yield translation-invariant quadratic unbiased estimators. In certain balanced data cases, these estimators are UMIVQUE's and thus, if the non-negativity constraints on the variance components do not come into play, agree with REML estimates (refer to Section 8.1). 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 non-negativity

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 [14] proposed a modified version (subsequently corrected by Thompson [73]) of Henderson's Method 3 which seems more akin to REML. It uses equations of the form (3.7) 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 $\underset{\sim}{\theta}$ and an iterative process is necessary.

W. A. Thompson [75] considered the problem of estimating variance components for balanced cases of random effects models where UMIVQUE's exist. He sought a procedure that would produce estimates which are inherently non-negative but which would agree with the ordinary ANOVA estimators for data sets where the latter estimates are all non-negative. He proposed a method based on an idea credited to Anderson and Bancroft [1]. Thompson's procedure applies when there exists a sufficient set of c+2 independent statistics consisting of the grand (arithmetic) mean and the c+1 mean squares from the relevant ANOVA table. It consists of maximizing the joint likelihood of the mean squares subject to non-negativity constraints on the variance components. Thompson indicated an extension of his approach to the estimation of covariance components and hinted at a complete generalization. It is

easy to show that Thompson's procedure agrees with the REML procedure in any case where Thompson's procedure applies. Moreover, his 'hinted-at' generalization would seem to be the REML procedure, so that it can be argued that the idea underlying REML is due to W. A. Thompson and perhaps to Anderson and Bancroft. For the balanced random-effects ANOVA models for which Thompson proposed his technique, exact expressions are possible for his estimators (and thus for the REML estimators). Thompson worked these out for several special cases including the balanced two-way crossed random-effects ANOVA models, both with and without interaction.

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 [67] or [68]). How these methods should be extended to the general problem of estimating $\underset{\sim}{\theta}$ is even less clear. In contrast, ML and REML estimators are always well-defined (at least conceptually). Moreover, except for 'balanced' cases, little is known about the goodness of the Henderson estimators, other than that they are unbiased and translation invariant. It is well-known that, at least in particular cases, there are biased estimators that have uniformly smaller MSE's than the Henderson estimators (see, e.g., [43]). What is more surprising is the recent discovery by Olsen, Seely, and Birkes [54] 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 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. When the computations necessary to implement REML are prohibitive, we can estimate variance and covariance components by various of Henderson's methods or by trying the approximate REML approach outlined in Section 7. The latter approach might have the better MSE properties, depending on what approximations were used. In any case, it has the advantage that the constraints on the parameter space are handled in a 'natural' way, though possibly Henderson's methods could satisfactorily be modified to accommodate constraints by a pseudo-maximization technique analogous to that imposed in Section 7 on the approximate REML procedure.

8.3 Bayesian Methods

A review of pre-1970 Bayesian results on inference for variance components can be found in [29]. 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 $\underset{\sim}{\theta}$ 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. 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 $\underset{\sim}{y}$ has the representation (2.1). If we wish to analyze the data by Bayesian techniques, we need to specify a prior distribution for $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$. Lindley and Smith [49] suggested in effect that, in many cases, it is possible to re-define 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 $\underset{\sim}{\alpha}$ to be independently and uniformly distributed over the real line and to be independent of $\underset{\sim}{\theta}$, 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 re-defined model, the highest order hyperparameters comprise the components of $\underset{\sim}{\alpha}$; the components of $\underset{\sim}{\beta}$ include the deviations of various orders or, possibly, appropriate linear combinations of those deviations, together with the components of the original $\underset{\sim}{\beta}$-vector; and additional 'parameters' may be inserted into the original $\underset{\sim}{\theta}$ vector to accommodate the new entries in the vector $\underset{\sim}{\beta}$. Of course, in the new model, some components of $\underset{\sim}{b}$ are random variables only in a subjective sense, but this is not objectionable because of the way in which the model is to be employed. It is supposed that the hyperparameters of various orders have been introduced in such a way that a prior distribution for $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ of the sought-after form is now appropriate.

If we wish to estimate $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and/or $\underset{\sim}{\theta}$ by their posterior means or at least by approximations to their posterior means, one way to proceed is to first estimate $\underset{\sim}{\theta}$ and to then estimate $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ by evaluating their conditional posterior means given $\underset{\sim}{\theta}$ at $\underset{\sim}{\theta} = \underset{\sim}{\hat{\theta}}$, where $\underset{\sim}{\hat{\theta}}$ is the estimate of $\underset{\sim}{\theta}$. This approach is

exact if the conditional posterior means of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ are linear in $\underset{\sim}{\theta}$ and if $\hat{\underset{\sim}{\theta}} = E(\underset{\sim}{\theta}|\underset{\sim}{y})$; and approximate otherwise. Suppose now that $p^* = p$ and that a priori the components of $\underset{\sim}{\alpha}$ are independently and uniformly distributed over the real line and are independent of $\underset{\sim}{\theta}$. If in addition the conditional prior distribution of $\underset{\sim}{b}$ and $\underset{\sim}{e}$ given $\underset{\sim}{\theta}$ (and $\underset{\sim}{\alpha}$) were normal, then the mean of the conditional posterior distribution of α and β for given θ is provided by $\hat{\underset{\sim}{\alpha}}$ and $\hat{\underset{\sim}{\beta}}^*$, where $\hat{\underset{\sim}{\alpha}}$ and $\hat{\underset{\sim}{\beta}}^*$ are as defined in Section 3. Even if the conditional prior distribution of $\underset{\sim}{b}$ and $\underset{\sim}{e}$ were non-normal, $\hat{\underset{\sim}{\alpha}}$ and $\hat{\underset{\sim}{\beta}}^*$ are still the linear expectations (in the sense described by Hartigan [25]) of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ with respect to their conditional posterior distributions, so that, in either case, it may make sense to estimate $\underset{\sim}{\alpha}$ and $\underset{\sim}{\beta}$ by $\hat{\underset{\sim}{\alpha}}(\hat{\underset{\sim}{\theta}})$ and $\hat{\underset{\sim}{\beta}}^*(\hat{\underset{\sim}{\theta}})$.

It remains to estimate $\underset{\sim}{\theta}$. Lindley and Smith would have us take the estimate of $\underset{\sim}{\theta}$ to be the $\underset{\sim}{\theta}$-component of the mode of the joint posterior density of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\theta}$. They acknowledge that their approach may be unsatisfactory if vague priors are assumed for certain components of $\underset{\sim}{\theta}$. In particular, 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 nonsensical estimators when used with informative priors. The problem with their approach may stem from the severe 'dependencies' that undoubtedly exist between components of $\underset{\sim}{\theta}$ and components of $\underset{\sim}{\beta}$ in the joint posterior

density of $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\theta}$, which may lead to the $\underset{\sim}{\theta}$-component of the mode of the joint posterior density being far removed from $E(\underset{\sim}{\theta}|\underset{\sim}{y})$.

A seemingly superior approach would be to take the estimate of $\underset{\sim}{\theta}$ to be the $\underset{\sim}{\theta}$-component of the mode of the marginal posterior density of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ or, better yet, the mode of the marginal posterior density of $\underset{\sim}{\theta}$. Suppose again that $p^* = p$, and that *a priori* the components of $\underset{\sim}{\alpha}$ are independently and uniformly distributed over the real line and are independent of $\underset{\sim}{\theta}$ so that the joint prior density of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ is proportional to $h(\underset{\sim}{\theta})$ for some function h. For purposes of estimating $\underset{\sim}{\theta}$ alone, it can, for reasons noted in [31], make sense to adopt such a prior density even if prior information on $\underset{\sim}{\alpha}$ is available. From [31], we have that the marginal posterior density of $\underset{\sim}{\theta}$ (the density obtained by formally integrating out $\underset{\sim}{\alpha}$) is proportional to the product of $h(\underset{\sim}{\theta})$ and the likelihood function of an arbitrary set of $(n-p^*)$ linearly independent error contrasts. For $h(\underset{\sim}{\theta}) \equiv 1$, the $\underset{\sim}{\theta}$-component of the mode of the marginal posterior density of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ is simply the ML estimate, and the mode of the marginal posterior density of $\underset{\sim}{\theta}$ is the REML estimate. Among noninformative priors, the Jeffreys' prior is sometimes recommended. The Jeffreys' prior for $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ is $[\det\{\underset{\sim}{X}'\underset{\sim}{V}^{-1}\underset{\sim}{X}\} \cdot \det\{\underset{\sim}{B}(\underset{\sim}{\theta})\}]^{\frac{1}{2}}$, which is consistent with our assumption about the form of the joint prior of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$. For the ordinary fixed ANOVA or regression models, where $q = 0$, $m = 1$, $\underset{\sim}{V} = \theta_1\underset{\sim}{I}$, and $\Omega = \{\underset{\sim}{\theta} : \theta_1 > 0\}$, this prior leads to

$[1/(n+p^*+2)](\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})$ as the $\underset{\sim}{\theta}$-component of the mode of the marginal posterior density of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ and to

$$[1/(n+2)](\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}}) \qquad (8.3)$$

as the mode of the marginal posterior density of $\underset{\sim}{\theta}$. Neither of these estimators is particularly appealing. Instead of using the Jeffreys' prior derived from L, we could consider taking $h(\underset{\sim}{\theta}) = [\det\{\underset{\sim}{B}^*(\underset{\sim}{\theta})\}]^{\frac{1}{2}}$, which is the Jeffreys' prior for $\underset{\sim}{\theta}$ based on L_1^*. For the ordinary fixed ANOVA models, we then have that the $\underset{\sim}{\theta}$-component of the mode of the marginal posterior density of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ is the estimator (8.3) and the mode of the marginal posterior density of $\underset{\sim}{\theta}$ is the estimator (4.7). As indicated in Section 4.3, the latter estimator has a downward bias of 'only' $2\theta_1/(n-p^*+2)$ and has uniformly smallest MSE among estimators of the form $(1/k)(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})'(\underset{\sim}{y} - \underset{\sim}{X}\hat{\underset{\sim}{\alpha}})$, so that it has appeal for frequentists who care about MSE but not about small biases. That the pesudo-Bayesian procedure that estimates θ by maximizing

$$L_1^*(\underset{\sim}{\theta};\ \underset{\sim}{y}) + (\tfrac{1}{2})\ \log[\det\{\underset{\sim}{B}^*(\theta)\}], \qquad (8.4)$$

for $\theta \in \Omega$, might be an equally satisfying procedure in more complicated settings is an intriguing possibility.

Suppose that we wish to estimate $\underset{\sim}{\alpha}$, $\underset{\sim}{\beta}$, and $\underset{\sim}{\theta}$, but that it is impossible or inconvenient to re-define the terms in (2.1) in such a way that a vague prior is appropriate for $\underset{\sim}{\alpha}$. We might then consider jointly estimating $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ as the mode of

their marginal posterior distribution. Using reasoning similar to that employed earlier, $\hat{\underset{\sim}{\beta}}$ evaluated at the modal values of $\underset{\sim}{\alpha}$ and $\underset{\sim}{\theta}$ could serve as the estimate of $\underset{\sim}{\beta}$.

In computing modal estimates for $\underset{\sim}{\theta}$ (and possibly $\underset{\sim}{\alpha}$), we have, in every case discussed above, that the logarithm of the marginal posterior density to be maximized consists of the logarithm of a prior density plus the logarithm of a likelihood function--either the full likelihood function or the likelihood function of a set of $(n-p^*)$ linearly independent error contrasts. Thus, in applying the algorithms described in Sections 6.2 and 6.3 to the problem of computing the mode of the relevant posterior density, we can, for purposes of simplifying the determination of the iterates, use essentially the same techniques as in ML or REML estimation to exploit structure in the $\underset{\sim}{R}$, $\underset{\sim}{D}$, $\underset{\sim}{Z}$, and $\underset{\sim}{X}$ matrices.

SECTION 9

FURTHER RESEARCH

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

Miller [52] made an impressive beginning on the problem of developing a realistic asymptotic theory for ML estimators

of $\underset{\sim}{\theta}$. His results apply to the ordinary ANOVA models. Extensions to a broader class of models of the form (2.1) and to REML estimators would be useful. For particular models, such as the ordinary ANOVA models, it would be nice to know what parameterizations produce the 'fastest convergence' to asymptotic normality. Still other unsolved problems with the asymptotic theory are indicated in [52].

In Sections 3 and 5, results were described which can be used to exploit structure in the $\underset{\sim}{R}$, $\underset{\sim}{D}$, $\underset{\sim}{Z}$, and $\underset{\sim}{X}$ 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. Explicit simplifications were given for the ordinary ANOVA models. It might be worthwhile to work out detailed procedures for other commonly used models. Thompson [74] did essentially this for MANOVA models.

While it is unlikely that any one of the iterative procedures for computing ML or REML estimates of $\underset{\sim}{\theta}$ 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 [7] for non-linear least squares problems, will ultimately be needed.

If Henderson's iterative algorithm for computing ML estimates of variance components and its various analogues 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 $\tilde{\alpha}$ and $\tilde{\beta}^*$ or, equivalently, for $\underset{\sim}{A}$ and $\underset{\sim}{H}$. It would be nice to know how the approximate REML scheme compares with Henderson's Methods 1, 2, and 3 as a procedure for estimating variance and covariance components.

The pseudo-Bayesian procedure that estimates $\underset{\sim}{\theta}$ by maximizing the expression (8.4) 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.

REFERENCES

[1] Anderson, R. L. and Bancroft, T. A., Statistical Theory in Research, New York: McGraw-Hill Book Co., 1952.

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

[3] Anderson, T. W., "Estimation of Covariance Matrices Which Are Linear Combinations or Whose Inverses Are Linear Combinations of Given Matrices," in R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao and K. J. C. Smith, ed., Essays in Probability and Statistics, Chapel Hill, North Carolina: University of North Carolina Press, 1970, 1-24.

[4] Anderson, T. W., "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, 1971.

[5] Anderson, T. W., The Statistical Analysis of Time Series, New York: John Wiley and Sons, Inc., 1971.

[6] Anderson, T. W., "Asymptotically Efficient Estimation of Covariance Matrices with Linear Structure," Annals of Statistics, 1 (January 1973), 135-41.

[7] Bard, Y., "Comparison of Gradient Methods for the Solution of Nonlinear Parameter Estimation Problems," SIAM Journal on Numerical Analysis, 7 (March 1970), 157-86.

[8] Bard, Y., Nonlinear Parameter Estimation, New York: Academic Press, 1974.

[9] Beltrami, E. J., An Algorithmic Approach to Nonlinear Analysis and Optimization, New York: Academic Press, 1970.

[10] Bliss, C. I., Statistics in Biology, Volume I, New York: McGraw-Hill Book Company, 1967.

[11] Box, G. E. P. and Jenkins, G. M., Time Series Analysis, San Francisco: Holden-Day, 1970.

[12] Carroll, C. W., "The Created Response Surface Technique for Optimizing Nonlinear, Restrained Systems," Operations Research, 9 (March-April 1961), 169-184.

[13] Corbeil, R. R. and Searle, S. R., "Restricted Maximum Likelihood (REML) Estimation of Variance Components in the Mixed Model," Paper No. BU-538-M, Biometrics Unit, Cornell University, Ithaca, New York, 1974.

[14] Cunningham, E. P. and Henderson, C. R., "An Iterative Procedure for Estimating Fixed Effects and Variance Components in Mixed Model Situations," Biometrics, 24 (March 1968), 13-25.

[15] Davies, O. L. (editor), Statistical Methods in Research and Production, Third Edition, London: Oliver and Boyd, 1967.

[16] Duncan, D. B. and Horn, S. D., "Linear Dynamic Recursive Estimation from the Viewpoint of Regression Analysis," Journal of the American Statistical Association, 67 (December 1972), 815-21.

[17] Fletcher, R. and Powell, M. J. D., "A Rapidly Convergent Descent Method for Minimization," Computer Journal, 6 (July 1963), 163-8.

[18] Fox, R. L. and Kapoor, M. P., "Rate of Change of Eigenvalues and Eigenvectors," AIAA Journal, 6 (December 1968), 2426-9.

[19] Gautschi, W., "Some Remarks on Herbach's Paper, 'Optimum Nature of the F-Test for Model II in the Balanced Case'," Annals of Mathematical Statistics, 30 (December 1959), 960-3.

[20] Goldfarb, D., "Extension of Davidon's Variable Metric Method to Maximization under Linear Inequality and Equality Constraints," SIAM Journal on Applied Mathematics, 17 (July 1969), 739-64.

[21] Goldfarb, D. and Lapidus, L., "Conjugate Gradient Method for Nonlinear Programming Problems with Linear Constraints," Industrial and Engineering Chemistry Fundamentals, 7 (February 1968), 142-51.

[22] Goldfeld, S. M., Quandt, R. E. and Trotter, H. F., "Maximization by Quadratic Hill Climbing," Econometrica, 34 (July 1966), 541-51.

[23] Graybill, F. A., Introduction to Matrices with Applications in Statistics, Belmont, California: Wadsworth Publishing Company, Inc., 1969.

[24] Graybill, F. A. and Hultquist, R. A., "Theorems Concerning Eisenhart's Model II," Annals of Mathematical Statistics, 32 (March 1961), 261-9.

[25] Hartigan, J. A., "Linear Bayesian Methods," Journal of the Royal Statistical Society, Series B, 31, No. 3 (1969), 446-54.

[26] Hartley, H. O. and Rao, J. N. K., "Maximum-Likelihood Estimation for the Mixed Analysis of Variance Model," Biometrika, 54 (June 1967), 93-108.

[27] Hartley, H. O. and Vaughn, W. K., "A Computer Program for the Mixed Analysis of Variance Model Based on Maximum Likelihood," in T. A. Bancroft, ed., Statistical Papers in Honor of George W. Snedecor, Ames, Iowa: Iowa State University Press, 1972.

[28] Harville, D. A., "Quadratic Unbiased Estimation of Variance Components for the One-Way Classification," Biometrika, 56 (August 1969), 313-26. (Correction, Biometrika, 57 (April 1970), 226).

[29] Harville, D. A., "Variance-Component Estimation for the Unbalanced One-Way Random Classification--A Critique," Technical Report No. 69-0180, Aerospace Research Laboratories, Wright-Patterson AFB, Ohio, 1969.

[30] Harville, D. A., "Generalization of the Gauss-Markov Theorem to Include the Estimation of Random Effects," Presented at the 36th Annual Meeting of the Institute of Mathematical Statistics, New York, 1973. (Abstract, IMS Bulletin, 2 (November 1973), 231).

[31] Harville, D. A., "Bayesian Inference for Variance Components Using Only Error Contrasts," Biometrika, 61 (August 1974), 383-5.

[32] Harville, D. A., "Linear Models as a Basis for Rating High School and College Football Teams," In Preparation, 1975.

[33] Hemmerle, W. J. and Hartley, H. O., "Computing Maximum Likelihood Estimates for the Mixed A.O.V. Model Using the W Transformation," Technometrics, 15 (November 1973), 819-31.

[34] Henderson, C. R., "Estimation of Variance and Covariance Components," Biometrics, 9 (June 1953), 226-52.

[35] Henderson, C. R., "Selection Index and Expected Genetic Advance," in Statistical Genetics and Plant Breeding, National Academy of Sciences--National Research Council Publication No. 982, 1963, 141-63.

[36] Henderson, C. R., "Maximum Likelihood Estimation of Variance Components," Unpublished Manuscript, 1973.

[37] Henderson, C. R., "MINQUE of Variance Components," Unpublished Manuscript, 1973.

[38] Henderson, C. R., "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, 1973, 10-41.

[39] Herbach, L. H., "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 (December 1959), 939-59.

[40] Hocking, R. R. and Kutner, M. H., "Some Analytical and Numerical Comparisons of Estimators for the Mixed A.O.V. Model," To appear in Biometrics, 1975.

[41] Joreskog, K. G., "Analysis of Covariance Structures," in P. R. Krishnaiah, ed., Multivariate Analysis-III, New York: Academic Press, 1973, 263-85.

[42] Kempthorne, O., An Introduction to Genetic Statistics, New York: John Wiley and Sons, Inc., 1957.

[43] Klotz, J. H., Milton, R. C. and Zacks, S., "Mean Square Efficiency of Estimators of Variance Components," Journal of the American Statistical Association, 64 (December 1969), 1383-402.

[44] LaMotte, L. R., "A Class of Estimators of Variance Components," Technical Report No. 10, Department of Statistics, University of Kentucky, Lexington, Kentucky, 1970.

[45] LaMotte, L. R., "Locally Best Quadratic Estimators of Variance Components," Technical Report No. 22, Department of Statistics, University of Kentucky, Lexington, Kentucky, 1971.

[46] LaMotte, L. R., "Quadratic Estimation of Variance Components," Biometrics, 29 (June 1973), 311-30.

[47] Lawton, W. H. and Sylvestre, E. A., "Elimination of Linear Parameters in Nonlinear Regression," Technometrics, 13 (August 1971), 461-7.

[48] Levenberg, K., "A Method for the Solution of Certain Non-Linear Problems in Least Squares," Quarterly of Applied Mathematics, 2 (July 1944), 164-8.

[49] Lindley, D. V. and Smith, A. F. M., "Bayes Estimates for the Linear Model," Journal of the Royal Statistical Society, Series B, 34, No. 1 (1972), 1-18.

[50] Marcus, M. and Minc, H., Introduction to Linear Algebra, New York: The Macmillan Company, 1965.

[51] Marquardt, D. W., "An Algorithm for Least Squares Estimation of Nonlinear Parameters," SIAM Journal, 11 (June 1963), 431-41.

[52] Miller, J. J., "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, 1973.

[53] Nering, E. D., Linear Algebra and Matrix Theory, Second ed., New York: John Wiley and Sons, Inc., 1970.

[54] Olsen, A., Seely, J. and Birkes, D., "Invariant Quadratic Unbiased Estimation for Two Variance Components," Unpublished Manuscript, 1975.

[55] Patterson, H. D. and Thompson, R., "Recovery of Inter-Block Information when Block Sizes Are Unequal," Biometrika, 58 (December 1971), 545-54.

[56] Powell, M. J. D., "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives," Computer Journal 7 (July 1964), 155-62.

[57] Powell, M. J. D., "A Survey of Numerical Methods for Unconstrained Optimization," SIAM Review, 12 (January 1970), 79-97.

[58] Rao, C. R., Linear Statistical Inference and Its Applications, New York: John Wiley and Sons, Inc., 1965.

[59] Rao, C. R., "Estimation of Heteroscedastic Variances in Linear Models," Journal of the American Statistical Association, 65 (March 1970), 161-72.

[60] Rao, C. R., "Estimation of Variance and Covariance Components--MINQUE Theory," Journal of Multivariate Analysis, 1 (September 1971), 257-75.

[61] Rao, C. R., "Minimum Variance Quadratic Unbiased Estimation of Variance Components," Journal of Multivariate Analysis, 1 (December 1971), 445-56.

[62] Rao, C. R., "Estimation of Variance and Covariance Components in Linear Models," Journal of the American Statistical Association, 67 (March 1972), 112-15.

[63] Rao, J. N. K., "On Expectations, Variances, and Covariances of ANOVA Mean Squares by 'Synthesis'," Biometrics, 24 (December 1968), 963-78.

[64] Rosen, J. B., "The Gradient Projection Method for Nonlinear Programming. Part I. Linear Constraints," SIAM Journal, 8 (March 1960), 181-217.

[55] Searle, S. R., "Another Look at Henderson's Methods of Estimating Variance Components" (with discussion), Biometrics, 24 (December 1968), 749-87.

[66] Searle, S. R., "Large Sample Variances of Maximum Likelihood Estimators of Variance Components Using Unbalanced Data," Biometrics, 26 (September 1970), 505-24.

[67] Searle, S. R., Linear Models, New York: John Wiley and Sons, Inc., 1971.

[68] Searle, S. R., "Topics in Variance Component Estimation," Biometrics, 27 (March 1971), 1-76.

[69] Searle, S. R., "Prediction, Mixed Models, and Variance Components," in F. Proschan and R. J. Serfling, ed., Reliability and Biometry, Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, 1974, 229-66.

[70] Smith, F. B. and Shanno, D. F., "An Improved Marquardt Procedure for Nonlinear Regressions," Technometrics, 13 (February 1971), 63-74.

[71] Snedecor, G. W. and Cochran, W. G., Statistical Methods, Sixth Edition, Ames, Iowa: Iowa State University Press, 1967.

[72] Stewart III, G. W., "A Modification of Davidon's Minimization Method to Accept Difference Approximations of Derivatives," Journal of the Association for Computing Machinery, 14 (January 1967), 72-83.

[73] Thompson, R., "Iterative Estimation of Variance Components for Non-Orthogonal Data," Biometrics, 25 (December 1969), 767-73.

[74] Thompson, R., "The Estimation of Variance and Covariance Components with an Application when Records Are Subject to Culling," Biometrics, 29 (September 1973), 527-50.

[75] Thompson, W. A., Jr., "The Problem of Negative Estimates of Variance Components," Annals of Mathematical Statistics, 33 (March 1962), 273-89.

[76] Townsend, E. C., "Unbiased Estimators of Variance Components in Simple Unbalanced Designs," Ph.D. Dissertation, Cornell University, Ithaca, New York, 1968.

[77] Townsend, E. C. and Searle, S. R., "Best Quadratic Unbiased Estimation of Variance Components from Unbalanced Data in the 1-Way Classification," Biometrics, 27 (September 1971), 643-57.

[78] Westlake, J. R., A Handbook of Numerical Matrix Inversion and Solution of Linear Equations, New York: John Wiley and Sons, Inc., 1968.

[79] Zacks, S., The Theory of Statistical Inference, New York: John Wiley and Sons, Inc., 1971.

☆U.S.Government Printing Office: 1975 — 657-023/696