Variance Component Estimation Using Constrained Nonlinear Maximization

BU-1029-M June 1989

Franz Preitschopf
Universität Augsburg

George Casella
Cornell University

Key words and phrases: Mixed model, maximum likelihood, iterative procedures

Abstract

Variance component estimation, using likelihood techniques, is a nonlinear maximization problem with constraints on the solution. Iterative procedures that do not directly accommodate these constraints may have questionable statistical properties. Fast nonlinear maximizers are available, however, that maintain the solution constraints at each iteration. We apply a particular algorithm to likelihood based variance components estimation, and show that it outperforms standard procedures. Some algebraic reductions are also established.

1. Introduction

In variance component estimation, using a mixed linear model, likelihood estimation techniques will often necessitate some type of nonlinear maximization. Furthermore, this nonlinear maximization will be constrained by the structure of the problem. Algorithms for solution of these problems are necessarily iterative, but some do not directly take account of the constraints on the solution. Thus, it is possible to have one of the iteration steps yield an answer that is outside of the feasible region, the region of allowable solutions defined by the constraints. Although this situation is not of direct numerical concern, the statistical implication of such an occurrence is not clear.

For example, although variances are necessarily positive, an iteration step could result in negative estimates. Such estimates are not just numerical values outside of the feasible region, but are not solutions to any real statistical problem. The statistical path back into the feasible region is not clear, for the meaning of the solutions is not clear. There are some algorithms (such as EM, see Dempster, Laird and Rubin 1977) which avoid this problem, and hence provide a meaningful statistical solution. In the case of EM this solution is provided at the expense of time. An alternative is explored here, the use of a nonlinear, constrained optimizer (using the NAG library) applied to a mixed model problem. As suggested by Harville (1977), we treat the problem as true nonlinear optimization, and require all iterations to remain in the feasible region. The results, when compared with standard techniques found in SAS, are surprisingly good.

2. Variance Estimation in the Mixed Model

The classical mixed model is given by

$$Y = X\beta + Zu + \epsilon \,, \tag{2.1}$$

where $Y_{n\times 1}$ is the vector of observations, $X_{n\times p}$ and $Z_{n\times q} = \sum_{i=1}^{r} Z_i$ are known design matrices, Z_i having order $n \times q_i$, $\beta_{p\times 1}$ is a vector of unknown fixed effects, $u = (u_1,\cdots,u_r)$ is a vector of random effects, u_i having order $q_i \times 1$, and ϵ is an $n \times 1$ vector of random errors. We further assume that the random factors are normally distributed:

$$\begin{aligned} \epsilon &\sim N\left(0,\sigma_e^2 I_n\right) \\ u &\sim N\left(0,\mathrm{diag}(\sigma_1^2 I_{q_1},\ldots,\sigma_r^2 I_{q_2})\right) \\ \mathrm{Cov}(u,\epsilon) &= 0 \,. \end{aligned} \tag{2.2}$$

Under these assumptions we have

$$\begin{aligned} Y &\sim N(X\beta,V) \\ V &= \sum_{i=1}^{r} \sigma_i^2 Z_i Z_i' + \sigma_e^2 I_n \\ &= \sum_{i=0}^{r} \sigma_i^2 Z_i Z_i' \,, \end{aligned} \tag{2.3}$$

defining $\sigma_0^2 = \sigma_e^2$ and $Z_0 Z_0' = I_n$.

The constrained estimation problem is to estimate the variance components $\sigma_0^2,\cdots,\sigma_r^2$ subject to the constraint that V is positive definite. This is equivalent to requiring

$$\sigma_0^2 > 0, \quad \sigma_i^2 \geq 0, \quad i=1,\cdots,r \,. \tag{2.4}$$

2.1 Maximum Likelihood Estimation

The method of maximum likelihood possesses many desirable statistical properties. Maximum likelihood (ML) estimators are functions of the sufficient statistics of the model, are consistent and efficient estimators, and are asymptotically normal (Miller 1977). However, except in special cases, there are no closed-form ML estimates in the variance component problem. This situation should pose no problem in light of available computing power.

Under the model (2.3), the log likelihood function is proportional to

$$\log L(\beta,V) = -\tfrac{1}{2}\log|V| - \tfrac{1}{2}(Y - X\beta)'V^{-1}(Y - X\beta) , \tag{2.5}$$

where $|\cdot|$ denotes determinant and $(\ )'$ denotes transpose. We now want to maximize (2.5) subject to the constraints of (2.4). Using the standard technique of equating derivatives to zero, the solution of the following equations give a (possible) maximum of (2.5), and hence ML estimators of σ_i^2, i=0,$\cdots$,r:

$$\begin{aligned} X'V^{-1}X\beta &= X'V^{-1}Y \\ \text{trace}\left(V^{-1}Z_iZ_i'\right) &= Y'PZ_iZ_i'PY, \quad i=0,\cdots,r , \end{aligned} \tag{2.6}$$

where

$$P = V^{-1} - V^{-1}X\left(X'V^{-1}X\right)^{-}X'V^{-1} , \tag{2.7}$$

and $(\cdot)^-$ denotes generalized inverse.

Solution of the equations in (2.6) is an iterative process, one that is complicated by the constraints (2.4). Doing this iteration can be more computationally intensive than a direct nonlinear maximization of (2.5) using a modern algorithm.

2.2 Restricted Maximum Likelihood Estimation

A popular alternative to ML estimation is known as restricted maximum likelihood (REML) estimation. There are many justifications of REML estimation, perhaps the most straightforward is that it is equivalent to marginal likelihood. That is, first append to the model (2.3) the additional assumption that

$$\beta_{p\times 1} \sim \text{uniform in } \mathbb{R}^p , \tag{2.8}$$

and calculate the marginal distribution of Y by integrating out β. Then perform ordinary maximum likelihood estimation on the resulting model. This is REML estimation.

After the algebraic dust has cleared, the REML log likelihood function is proportional to

$$\log L(V) = -\tfrac{1}{2}\log|K'VK| - \tfrac{1}{2}Y'K(K'VK)^{-1}K'Y , \tag{2.9}$$

where K is any full-rank matrix satisfying $K'X = 0$. It then follows that (2.9) is independent of the choice of K.

The likelihood in (2.9) could have, alternatively, been derived by first multiplying both sides of (2.1) by K, which then yields $Y \sim N(0, K'VK)$ and the likelihood of (2.9). The

justification of this derivation is that estimation of the variance components is then based on "error contrasts," that is, the part of the data that is orthogonal to the fixed effects. In either case, the resulting REML variance estimates are translation invariant (not dependent on the values of the fixed effect estimates) and their degrees of freedom take account the estimation of the fixed effects, two very desirable properties.

Differentiating (2.9) and equating to zero yields, after much algebra, the REML equations

$$\text{trace}\left(PZ_iZ_i'\right) = Y'PZ_iZ_i'PY \qquad i=0,\cdots,r \tag{2.10}$$

where P is given by (2.7), or equivalently

$$P = K(K'VK)^{-1}K' , \tag{2.11}$$

a fact established by Khatri (1966) if X has full rank, and Pukelsheim (1988) in general.

As can be seen, solution of the equations in (2.10) will be quite similar to those in (2.6). Thus, any advantages enjoyed by nonlinear maximization methods in ML estimation will carry over to REML estimation.

3. The Twoway Model with Interactions

To evaluate the computational improvement offered by constrained nonlinear optimization, we investigated the twoway model in detail. As a side benefit, algebraic reductions were obtained in solving the estimation equations.

3.1 The Optimization Problem

The twoway model with interactions is a special case of the model (2.1) which is commonly written

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \qquad \begin{matrix} i=1,\cdots,a \\ j=1,\cdots,b \\ k=1,\cdots,n \end{matrix} \tag{3.1}$$

where μ is the fixed, overall mean and α_i, β_j and γ_{ij} denote random effects. Notice also that we have restricted attention to the balanced case. In matrix notation, model (3.1) can be written

$$Y = 1\mu + Z_1\alpha + Z_2\beta + Z_3\gamma + Z_0\epsilon \tag{3.2}$$

where 1 is vector of 1s, $\alpha = (\alpha_1,\cdots,\alpha_a)'$, $\beta = (\beta_1,\cdots,\beta_b)'$, $\gamma = (\gamma_{11},\cdots,\gamma_{ab})'$ and

$$\begin{aligned} Z_0 &= I_a \otimes I_b \otimes I_n \\ Z_1 &= I_a \otimes 1_b \otimes 1_n \\ Z_2 &= 1_a \otimes I_b \otimes 1_n \\ Z_3 &= I_a \otimes I_b \otimes 1_n \ , \end{aligned} \tag{3.3}$$

where $\otimes$ denotes Kronecker product.

To obtain likelihood estimates, we need an expression for both log L and its gradient, where

$$\log L\left(\mu,\sigma_0^2,\sigma_1^2,\sigma_2^2,\sigma_3^2\right) \propto -\tfrac{1}{2}\log|V| - \tfrac{1}{2}(Y - 1\mu)'V^{-1}(Y - 1\mu) \tag{3.4}$$

$$V = \sum_{i=0}^{3} \sigma_i^2 Z_i Z_i' \ ,$$

and the gradient is given by

(3.5) $$\frac{\partial}{\partial \mu}\log L = 1'V^{-1}(Y - 1\mu)$$

$$\frac{\partial}{\partial \sigma_i^2}\log L = -\text{trace}\left(V^{-1}Z_iZ_i'\right) + (Y - 1\mu)'V^{-1}Z_iZ_i'V^{-1}(Y - 1\mu) \quad i=0,\cdots,3 .$$

The equations in (3.5) define the ML equations for this case, and their solutions, subject to $\sigma_0 > 0$, $\sigma_i^2 \geq 0$, i=1,2,3, yield the ML estimates. Maximizing (3.4) subject to this constraint defines the nonlinear maximization problem.

The analogous equations can be derived for REML estimation in the twoway case, yielding a log likelihood

(3.6) $$\log L\left(\sigma_0^2,\sigma_1^2,\sigma_2^2,\sigma_3^2\right) \propto -\tfrac{1}{2}\log|K'VK| - \tfrac{1}{2}Y'PY ,$$

and gradient

$$\frac{\partial}{\partial \sigma_i^2}\log L = -\tfrac{1}{2}\,\text{trace}\left(PZZ_i'\right) + \tfrac{1}{2}Y'PZ_iZ_i'Y \quad i=0,\cdots,3 ,$$

which also define the optimization problem, subject to the variance constraint.

3.2 Algebraic Reductions

Using the Kronecker product structure of Z_0 - Z_3, it is possible to simplify expressions for V and V^{-1}, and thus greatly reduce computation effort in the likelihood calculations.

For $V = \sum_{i=0}^{3}\sigma_i^2Z_iZ_i'$, where the Z_i are given by (3.3), Searle (1988) derived the inverse of V as

(3.7) $$V^{-1} = \sum_{i=0}^{4}\eta_i^{-1}S_i$$

where

(3.8)
$$\begin{aligned}
\eta_0 &= \sigma_0^2 \\
\eta_1 &= \sigma_0^2 + n\sigma_3^2 + bn\sigma_1^2 \\
\eta_2 &= \sigma_0^2 + n\sigma_3^2 + an\sigma_2^2 \\
\eta_3 &= \sigma_0^2 + n\sigma_3^2 \\
\eta_4 &= \eta_1 + \eta_2 - \eta_3
\end{aligned}$$

and

$$
\begin{aligned}
S_0 &= I_a \otimes I_b \otimes C_n \\
S_1 &= C_a \otimes \tfrac{1}{b}J_k \otimes \tfrac{1}{n}J_n \\
S_2 &= \tfrac{1}{a}J_a \otimes C_b \otimes \tfrac{1}{n}J_n \\
S_3 &= C_a \otimes C_b \otimes \tfrac{1}{n}J_n \\
S_4 &= \tfrac{1}{a}J_a \otimes \tfrac{1}{b}J_b \otimes \tfrac{1}{n}J_n
\end{aligned}
\tag{3.9}
$$

where $J_k = 1_k 1_k'$ and $C_k = I_k - \frac{1}{k}J_k$. Note that the parameters $\eta_0,\cdots,\eta_3$ are the expected values of the mean squares in a twoway random analysis of variance. It is also the case that the quadratic forms $Y'S_iY$, i=0,$\cdots$,3 yield the usual sums of squares, with $Y'S_4Y$ being the correction term. More precisely,

$$
\begin{aligned}
Y'S_0Y &= SSE = \sum_{i,j,k}\left(y_{ijk} - \overline{y}_{ij\cdot}\right)^2 \\
Y'S_1Y &= SSA = bn\sum_{i}\left(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdots}\right)^2 \\
Y'S_2Y &= SSB = an\sum_{j}\left(\overline{y}_{\cdot j\cdot} - \overline{y}_{\cdots}\right)^2 \\
Y'S_3Y &= SSAB = n\sum_{i,j}\left(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot j\cdot} + \overline{y}_{\cdots}\right)^2 \\
Y'S_4Y &= CT = abn\,\overline{y}^2_{\cdots} .
\end{aligned}
\tag{3.10}
$$

If we now apply (3.7) - (3.10) to the likelihood function of (3.4), we can write

$$
\log|V| \propto ab(n-1)\log\eta_0 + (a-1)\log\eta_1 + (b-1)\log\eta_2 + (a-1)(b-1)\log\eta_3 + \log\eta_4
\tag{3.11}
$$

$$
(Y - 1\mu)'V^{-1}(Y - 1\mu) = \eta_0^{-1}SSE + \eta_1^{-1}SSA + \eta_2^{-1}SSB + \eta_3^{-1}SSAB + \eta_4^{-1}CT ,
$$

greatly simplifying the likelihood calculations.

Analogous reductions hold for REML estimation in this case. Recall the definition of the matrix P, given in expression (2.7). For Z_0 - Z_3 of (3.3), we have

$$
V^{-1}1\left(1'V^{-1}1\right)^{-1}1'V^{-1} = \eta_4^{-1}\frac{1}{abn}J ,
\tag{3.12}
$$

and hence

$$
P = V^{-1} - \eta_4^{-1}\frac{1}{abn}J .
\tag{3.13}
$$

Using this expression for P, we can simplify the REML likelihood of (3.6) by using the expressions

$$\log|K'VK| \propto ab(n\text{-}1)\log \eta_0 + (a\text{-}1)\log \eta_1 + (b\text{-}1)\log \eta_2 + (a\text{-}1)(b\text{-}1)\log \eta_3$$

(3.14)

$$Y'PY = \eta_0^{-1}SSE + \eta_1^{-1}SSA + \eta_2^{-1}SSB + \eta_3^{-1}SSAB\ .$$

4. Constrained Nonlinear Maximization

The nonlinear maximizer used is part of the National Algorithms Group (NAG) library, routine E04KDF, which is based on work of Gill and Murray (1974). It requires only the function, gradient, and feasible region as input. The maximizer is extremely easy to use, and gave excellent results whether it was used for ML or REML estimation.

We compared the NAG maximizer to a number of estimation procedures found in SAS – ML and REML based on Hemmerle and Hartley (1973), ANOVA estimation, and MIVQUE. The outcome of our comparisons were quite similar, so we will only discuss one data set in detail. It is a small (a=3, b=4, n=2) data set from Searle (1988). Results of the variance estimation are summarized in Table 1.

Table 1: Comparison of Variance Estimates

	SAS				
	ML	REML	ANOVA	MIVQUE	NAG-ML
σ_0^2	20.10	20.10	20.83	20.83	20.10
σ_1^2	16.15	25.49	25.17	25.17	16.15
σ_2^2	0	0	-1.78	-1.78	0
σ_3^2	0	0	.92	.92	0

The NAG-ML estimation agreed with the ML estimates in all cases that we tried. We also varied starting points and started at zero estimates known to be positive. Our usual starting point was the ANOVA estimates with negative values set to zero. These all worked fine.

The most impressive improvement, in terms of numerical considerations, was in computation time. On Cornell University's IBM mainframe SAS-ML took .31 seconds of CPU time on the average, while NAG-ML took .03 seconds. Thus, statistical considerations aside, NAG-ML (and NAG-REML) are ten times faster than their SAS counterparts.

5. Discussion

There are a number of algorithms to choose from when faced with a nonlinear optimization problem. From a numerical view, the important consideration is speed of convergence to the optimum. For the statistician, however, their is another concern. Any iterative procedure should yield feasible results at every iteration. If not, then the statistical interpretation of that iteration is meaningless, and hence the statistical interpretation of the solution is questionable. In particular, it is not known, in general, if an iteration path that goes outside of the feasible region will always reach a maximum. Furthermore, some algorithms truncate nonfeasible iterations back to the feasible region. It is not known if this practice will guarantee a path to the maximum The only sure way to avoid these problems is to never allow the iteration path to jump out of the feasible region.

Naive iteration of the ML equations of (2.6) or the REML equations of (2.10) could have the problem of going outside of the feasible region. A statistical-based algorithm such as EM is wonderful in that every step of the process remains feasible, and hence the statistical interpretation of the solution is kept intact. However, the convergence of EM can be very slow, making it impractical for large problems.

The advantages of the algorithms considered in this paper, either NAG-ML or NAG-REML, are twofold. Like EM, each iteration results in a feasible solution, leaving a clear statistical meaning to the solution. However, these algorithms are fast, resulting in order of magnitude improvements in computation time.

Once we have satisfied our statistical interpretational needs with an algorithm that always remains feasible, we can then exploit its numerical properties. For example, when doing iterations on the ML equations, we noticed that if a variance component was set to zero, it remained there. This need not be the case with the NAG maximizers, as they have the ability to move iterates away from the boundary. Another interesting application is the following. The NAG algorithm is so fast that it can easily be imbedded in the EM algorithm to solve unbalanced maximum likelihood problems. Given an unbalanced problem, use the E-

step to fill in data to balance the problem, and use NAG-ML (or REML) to do the M-step. Given the speed of the NAG algorithm, this NAG/EM approach should be reasonable in a wide variety of cases.

References

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

Gill, P.E. and Murray, W. (1974). Newton-type methods for unconstrained and linearly constrained optimization. *Mathematical Programming* **7**, 311-350.

Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72**, 320-339.

Hemmerle, W.J. and Hartley, H.O. (1973). Computing maximum likelihood estimates for the mixed A.O.V. model using the W transformation. *Technometrics* 15, 819-831.

Khatri, C.G. (1966). A note on a MANOVA model applied to problems in growth curves. *Annals Inst. Statist. Math.* 18, 75.

Miller, J.J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Ann. Statist.* 5, 746-762.

Pukelsheim, F. (1988). Personal communication.

Searle, S.R. (1988). Lecture notes on variance components. Biometrics Unit, Cornell University, Ithaca, NY.