Chapter 12
REML and ML Estimation

C. R. Henderson

1984 - Guelph

1 Iterative MIVQUE

The restricted maximum likelihood estimator (REML) of Patterson and Thompson (1971) can be obtained by iterating on MIVQUE, Harville (1977). Let the prior value of $\mathbf{g}$ and $\mathbf{r}$ be denoted by $\mathbf{g}_0$ and $\mathbf{r}_0$. Then compute MIVQUE and denote the estimates by $\mathbf{g}_1$ and $\mathbf{r}_1$. Next use these as priors in MIVQUE and denote the estimates $\mathbf{g}_2$ and $\mathbf{r}_2$. Continue this process until $\mathbf{g}_{k+1} = \mathbf{g}_k$ and $\mathbf{r}_{k+1} = \mathbf{r}_k$. Several problems must be recognized.

1. Convergence may be prohibitively slow or may not occur at all.

2. If convergence does occur, it may be to a local rather than a global maximum.

3. If convergence does occur, $\mathbf{g}$ and $\mathbf{r}$ may not fall in the parameter space.

We can check the last by noting that both $\mathbf{G}_k$ and $\mathbf{R}_k$ must be positive definite or positive semidefinite at convergence, where $\mathbf{G}_k$ and $\mathbf{R}_k$ are

$$\begin{pmatrix} \hat{g}_{11} & \hat{g}_{12} & \cdots \\ \hat{g}_{12} & \hat{g}_{22} & \cdots \\ . & . & \\ . & . & \end{pmatrix} \text{ and } \begin{pmatrix} \hat{r}_{11} & \hat{r}_{12} & \cdots \\ \hat{r}_{12} & \hat{r}_{22} & \cdots \\ . & . & \\ . & . & \end{pmatrix}.$$

For positive definitness or positive semidefiniteness all eigenvalues of $\mathbf{G}_k$ and $\mathbf{R}_k$ must be non-negative. Writing a computer program that will guarantee this is not trivial. One possibility is to check at each round, and if the requirement is not met, new starting values are chosen. Another possibility is to alter some elements of $\hat{\mathbf{G}}$ or $\hat{\mathbf{R}}$ at each round in which either $\hat{\mathbf{G}}$ or $\hat{\mathbf{R}}$ is not a valid estimator. (LRS note: Other possibilities are bending in which eigenvalues are modified to be positive and the covariance matrix is reformed using the new eigenvalues with the eigenvectors.)

Quadratic, unbiased estimators may lead to solutions not in the parameter space. This is the price to pay for unbiasedness. If the estimates are modified to force them into the parameter space, unbiasedness no longer can be claimed. What should be done in practice? If the purpose of estimation is to accumulate evidence on parameters with other research, one should report the invalid estimates, for otherwise the average of many estimates will be biased. On the other hand, if the results of the analysis are to be used immediately, for example, in BLUP, the estimate should be required to fall in the parameter space. It would seem illogical for example, to reduce the diagonals of $\hat{\mathbf{u}}$ in mixed model equations because the diagonals of $\bar{\mathbf{G}}^{-1}$ are negative.

2 An Alternative Algorithm For REML

An alternative algorithm for REML that is considerably easier per round of iteration than iterative MIVQUE will now be described. There is, however, some evidence that convergence is slower than in the iterative MIVQUE algorithm. The method is based on the following principle. At each round of iteration find the expectations of the quadratics under the pretense that the current solutions to $\hat{\mathbf{g}}$ and $\hat{\mathbf{r}}$ are equal to $\mathbf{g}$ and $\mathbf{r}$. This leads to much simpler expectations. Note, however, that the first iterate under this algorithm is not MIVQUE. This is the EM (expectation maximization) algorithm, Dempster et al. (1977).

From Henderson (1975a), when $\bar{\mathbf{g}} = \mathbf{g}$ and $\bar{\mathbf{r}} = \mathbf{r}$

$$\begin{aligned} Var(\hat{\mathbf{u}}) &= \mathbf{G} - \mathbf{C}_{11}. && (1) \\ Var(\hat{\mathbf{e}}) &= \mathbf{R} - \mathbf{WCW}' = \mathbf{R} - \mathbf{S}. && (2) \end{aligned}$$

The proof of this is

$$Var(\hat{\mathbf{e}}) = \text{Cov}\ (\hat{\mathbf{e}}, \mathbf{e}') = Cov[(\mathbf{y} - \mathbf{WCW}'\mathbf{R}^{-1}\mathbf{y}), \mathbf{e}'] = \mathbf{R} - \mathbf{WCW}'. \qquad (3)$$

A g-inverse of the mixed model coefficient matrix is

$$\begin{pmatrix} \mathbf{C}_{00} & \mathbf{C}_{01} \\ \mathbf{C}_{10} & \mathbf{C}_{11} \end{pmatrix} = \mathbf{C}.$$

Note that if we proceed as in Section 11.5 we will need only diagonal blocks of $\mathbf{WCW}'$ corresponding to the diagonal blocks of $\mathbf{R}$.

$$Var(\hat{\mathbf{u}}) = \sum_i \sum_{j \geq i} \mathbf{G}^*_{ij} g_{ij} - \mathbf{C}_{11} \qquad (4)$$

See Chapter 11, Section 3 for definition of $\mathbf{G}^*$. Let $\mathbf{C},\ \mathbf{S},\ \hat{\mathbf{u}}$, and $\hat{\mathbf{e}}$ be the values computed for the k^{th} round of iteration. Then solve in the k+l round of iteration for values of $\mathbf{g}, \mathbf{r}$ from the following set of equations.

$$\begin{aligned} tr\mathbf{Q}_1\mathbf{G} &= \hat{\mathbf{u}}'\mathbf{Q}_1\hat{\mathbf{u}} + tr\mathbf{Q}_1\mathbf{C}_{11} \\ &\vdots \end{aligned}$$

$$\begin{aligned} tr\mathbf{Q}_b\mathbf{G} &= \hat{\mathbf{u}}'\mathbf{Q}_b\hat{\mathbf{u}}' \ + \ tr\mathbf{Q}_b\mathbf{C}_{11} \\ tr\mathbf{Q}_{b+1}\mathbf{R} &= \hat{\mathbf{e}}'\mathbf{Q}_{b+1}\hat{\mathbf{e}} \ + \ tr\mathbf{Q}_{b+1}\mathbf{S} \\ &\vdots \\ tr\mathbf{Q}_c\mathbf{R} &= \hat{\mathbf{e}}'\mathbf{Q}_c\hat{\mathbf{e}} \ + \ tr\mathbf{Q}_c\mathbf{S}. \end{aligned} \tag{5}$$

Note that at each round a set of equations must be solved for all elements of **g** , and another set of all elements of **r** . In some cases, however, **Q** 's can be found such that only one element of g_{ij} (or r_{ij}) appears on the left hand side of each equation of (5). Note also that if $\bar{\mathbf{G}}^{-1}$ appears in a $\mathbf{Q}_i$ the value of $\mathbf{Q}_i$ changes at each round of iteration. The same applies to $\bar{\mathbf{R}}^{-1}$ appearing in $\mathbf{Q}_i$ for $\hat{\mathbf{e}}$. Consequently it is desirable to find $\mathbf{Q}_i$ that isolate a single g_{ij} or r_{ij} in each left hand side of (5) and that are not dependent upon $\bar{\mathbf{G}}$ and $\bar{\mathbf{R}}$. This can be done for the g_{ij} in all genetic problems with which I am familiar.

The second algorithm for REML appears to have the property that if positive definite **G** and **R** are chosen for starting values, convergence, if it occurs, will always be to positive definite $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}$. This suggestion has been made by Smith (1982).

3 ML Estimation

A slight change in the second algorithm for REML, presented in Section 2 results in an EM type ML algorithm. In place of $\mathbf{C}_{11}$ substitute $(\mathbf{Z}'\bar{\mathbf{R}}^{-1}\mathbf{Z}+\mathbf{G}^{-1})^{-1}$. In place of $\mathbf{WCW}'$ substitute $\mathbf{Z}(\mathbf{Z}'\bar{\mathbf{R}}^{-1}\mathbf{Z}+\bar{\mathbf{G}}^{-1})^{-1}\mathbf{Z}'$. Using a result reported by Laird and Ware (1982) substituting ML estimates of **G** and **R** for the corresponding parameters in the mixed model equations yields empirical Bayes estimates of **u** . As stated in Chapter 8 the $\hat{\mathbf{u}}$ are also ML estimates of the conditional means of **u** .

If one wishes to use the LaMotte type quadratics for REML and ML, the procedure is as follows. For REML iterate on

$$tr\mathbf{Q}_j \ \sum_i \ \mathbf{V}_i^* \ \theta_i \ = \ (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^o)'\mathbf{Q}_j(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}^o) \ + \ tr\mathbf{Q}_j \ \ \mathbf{X}(\mathbf{X}'\bar{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'.$$

$\mathbf{Q}_j$ are the quadratics computed by the LaMotte method described in Chapter 11. Also this chapter describes the $\mathbf{V}_i^*$. Further, $\boldsymbol{\beta}^o$ is a GLS solution.

ML is computed in the same way as REML except that

$$tr\mathbf{Q}_j \ \ \mathbf{X}(\mathbf{X}'\bar{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}' \text{ is deleted.}$$

The EM type algorithm converges slowly if the maximizing value of one or more parameters is near the boundary of the parameter space, eg. $\hat{\sigma}_i^2 \rightarrow 0$. The result of Hartley and Rao (1967) can be derived by this general EM algorithm.

4 Approximate REML

REML by either iterative MIVQUE or by the method of Section 2 is costly because every round of iteration requires a g-inverse of the mixed model coefficient matrix. The cost could be reduced markedly by iterating on approximate MIVQUE of Section 11.7. Further simplification would result in the $\mathbf{R} = \mathbf{I}\sigma_e^2$ case by using the residual mean square of OLS as the estimate of σ_e^2. Another possibility is to use the method of Section 2, but with an approximate g-inverse and solution at each round of iteration. The properties of such an estimation are unknown.

5 A Simple Result For Expectation Of Residual Sum Of Squares

Section 11.13 shows that in a model with $\mathbf{R} = \mathbf{R}_*\sigma_e^2$, $Var(\mathbf{u}_i) = \mathbf{G}_{*i}\sigma_e^2$, and $Cov(\mathbf{u}_i, \mathbf{u}_j') = \mathbf{0}$, one of the quadratics that can be used is

$$\mathbf{y}'\mathbf{R}_*^{-1}\mathbf{y} - \text{(soln. vector)}' \text{ (r.h.s. vector)} \tag{6}$$

with equations written as (77) in Chapter 11. $\mathbf{R}_*$ and $\mathbf{G}_{*i}$ are known. Then if $\alpha = \sigma_e^2/\sigma_i^2$, as is defined in taking expectations for the computations of Section 2, the expectation of (6) is

$$[n - \text{rank } (\mathbf{X})]\sigma_e^2. \tag{7}$$

6 Biased Estimation With Few Iterations

What if one has only limited data to estimate a set of variances and covariances, but prior estimates of these parameters have utilized much more data? In that case it might be logical to iterate only a few rounds using the EM type algorithm for REML or ML. Then the estimates would be a compromise between the priors and those that would be obtained by iterating to convergence. This is similar to the consequences of Bayesian estimation. If the priors are good, it is likely that the MSE will be smaller than those for ML or REML. A small simulation trial illustrates this. The model assumed was

$$\begin{aligned}
y_{ij} &= \beta\mathbf{X}_{ij} + a_i + e_{ij}.\\
\mathbf{X}' &= (3, 2, 5, 1, 3, 2, 3, 6, 7, 2, 3, 5, 3, 2).\\
n_i &= (3, 2, 4, 5).\\
Var(\mathbf{e}) &= 4\ \mathbf{I},\\
Var(\mathbf{a}) &= \mathbf{I},
\end{aligned}$$

$$Cov(\mathbf{a}, \mathbf{e}') \quad = \quad \mathbf{0}.$$

5000 samples were generated under this model and EM type REML was carried out with starting values of $\sigma_e^2/\sigma_a^2 \;=\; 4, \;\; .5,$ and 100. Average values and MSE were computed for rounds $1, 2, ..., 9$ of iteration.

	Starting Value $\sigma_e^2/\sigma_a^2 = 4$					
	$\hat{\sigma}_e^2$		$\hat{\sigma}_a^2$		$\hat{\sigma}_e^2/\hat{\sigma}_a^2$	
Rounds	Av.	MSE	Av.	MSE	Av.	MSE
1	3.98	2.37	1.00	.22	4.18	.81
2	3.93	2.31	1.04	.40	4.44	2.97
3	3.88	2.31	1.10	.70	4.75	6.26
4	3.83	2.34	1.16	1.08	5.09	10.60
5	3.79	2.39	1.22	1.48	5.47	15.97
6	3.77	2.44	1.27	1.86	5.86	22.40
7	3.75	2.48	1.31	2.18	6.26	29.90
8	3.74	2.51	1.34	2.43	6.67	38.49
9	3.73	2.53	1.35	2.62	7.09	48.17

In this case only one round appears to be best for estimating σ_e^2/σ_a^2 .

	Starting Value $\sigma_e^2/\sigma_a^2 = .5$					
	$\hat{\sigma}_e^2$		$\hat{\sigma}_a^2$		$\hat{\sigma}_e^2/\hat{\sigma}_a^2$	
Rounds	Av.	MSE	Av.	MSE	Av.	MSE
1	3.14	2.42	3.21	7.27	1.08	8.70
2	3.30	2.37	2.53	4.79	1.66	6.27
3	3.40	2.38	2.20	4.05	2.22	5.11
4	3.46	2.41	2.01	3.77	2.75	5.23
5	3.51	2.43	1.88	3.64	3.28	6.60
6	3.55	2.45	1.79	3.58	3.78	9.20
7	3.57	2.47	1.73	3.54	4.28	12.99
8	3.60	2.49	1.67	3.51	4.76	17.97
9	3.61	2.51	1.63	3.50	5.23	24.11

	Starting Value $\hat{\sigma}_e^2/\hat{\sigma}_a^2 = 100$					
	$\hat{\sigma}_e^2$		$\hat{\sigma}_a^2$		$\hat{\sigma}_e^2/\hat{\sigma}_a^2$	
Rounds	Av.	MSE	Av.	MSE	Av.	MSE
1	4.76	4.40	.05	.91	.99	9011
2	4.76	4.39	.05	.90	.98	8818
3	4.76	4.38	.05	.90	.97	8638
4	4.75	4.37	.05	.90	.96	8470
5	4.75	4.35	.05	.90	.95	8315
6	4.75	4.34	.05	.90	.94	8172
7	4.75	4.32	.05	.90	.92	8042
8	4.74	4.31	.06	.89	.91	7923
9	4.74	4.28	.06	.89	.90	7816

Convergence with this very high starting value of σ_e^2/σ_a^2 relative to the true value of 4 is very slow but the estimates were improving with each round.

7 The Problem Of Finding Permissible Estimates

Statisticians and users of statistics have for many years discussed the problem of ”estimates” of variances that are less than zero. Most commonly employed methods of estimation are quadratic, unbiased, and translation invariant, for example ANOVA estimators, Methods 1,2, and 3 of Henderson, and MIVQUE. In all of these methods there is a positive probability that a solution to one or more variances will be negative. Strictly speaking, these are not really estimates if we define, as some do, that an estimate must lie in the parameter space. But, in general, we cannot obtain unbiasedness unless we are prepared to accept such solutions. The argument used is that such ”estimates” should be reported because eventually there may be other estimates of the same parameters obtained by unbiased methods, and then these can be averaged to obtain better unbiased estimates.

Other workers obtain truncated estimates. That is, given estimates $\hat{\sigma}_1^2, ..., \hat{\sigma}_q^2$, with say $\hat{\sigma}_q^2 < 0$, the estimates are taken as $\hat{\sigma}_1^2, ..., \hat{\sigma}_{q-1}^2$, 0. Still others revise the model so that the offending variable is deleted from the model, and new estimates are then obtained of the remaining variances. If these all turn out to be non-negative, the process stops. If some new estimate turns negative, then that variance is dropped from the model and a new set of estimates obtained.

These truncated estimators can no longer be defined as unbiased. Verdooren (1980) in an interesting review of variance component estimation uses the terms ”permissible” and ”impermissible” to characterize estimators. Permissible estimators are those in which the solution is guaranteed to fall in the parameter space, that is all estimates of variances are non-negative. Impermissible estimators are those in which there is a probability greater than 0 that the solution will be negative.

If one insists on permissible estimators, why not then use some method that guarantees this property while at the same time invoking, if possible, other desirable properties of estimators such as consistency, minimum mean squared error, etc.? Of course unbiasedness cannot, in general, be invoked. For example, an algorithm for ML, Henderson (1973), guarantees a permissible estimator provided convergence occurs. A simple extension of this method due to Harville (1977), yields permissible estimators by REML. The problem of permissible estimators is especially acute in multiple trait models. For example, in a two trait phenotypic model say

$$\begin{aligned} y_{ij} &= \mu_i + e_{1j} \\ y_{2j} &= \mu_2 + e_{2j} \end{aligned}$$

we need to estimate

$$Var \begin{pmatrix} e_{1j} \\ e_{2j} \end{pmatrix} = \begin{pmatrix} c_{11} & c_{12} \\ c_{12} & c_{22} \end{pmatrix}. \;\; c_{11} \geq 0, \;\; c_{22} \geq 0, \;\; c_{11}c_{22} \geq c_{12}^2.$$

The last of these criteria insures that the estimated correlation between e_{1j} and e_{2j} falls in the range -1 to 1. The literature reporting genetic correlation estimates contains many cases in which the criteria are not met, this in spite of probable lack of reporting of many other sets of computations with such results. The problem is particularly difficult when there are more than 2 variates. Now it is not sufficient for all estimates of variances to be non- negative and all pairs of estimated correlations to fall in the proper range. The requirement rather is that the estimated variance-covariance matrix be either positive definite or at worst positive semi-definite. A condition guaranteeing this is that all latent roots (eigenvalues) be positive for positive definiteness or be non-negative for positive semidefiteness. Most computing centers have available a good subroutine for computing eigenvalues. We illustrate with a 3 × 3 matrix in which all correlations are permissible, but the matrix is negative definite.

$$\begin{pmatrix} 3 & -3 & 4 \\ -3 & 4 & 4 \\ 4 & 4 & 6 \end{pmatrix}$$

The eigenvalues for this matrix are (9.563, 6.496, -3.059), proving that the matrix is negative definite. If this matrix represented an estimated $\mathbf{G}$ for use in mixed model equations, one would add $\mathbf{G}^{-1}$ to an appropriate submatrix, of OLS equations, but

$$\mathbf{G}^{-1} = \begin{pmatrix} -.042 & -.139 & .147 \\ & -.011 & .126 \\ & & -.016 \end{pmatrix},$$

so one would add negative quantities to the diagonal elements, and this would make no sense. If the purpose of variance-covariance estimation is to use the estimates in setting up mixed model equations, it is essential that permissible estimators be used.

Another difficult problem arises when variance estimates are to be used in estimating h^2. For example, in a sire model, an estimate of h^2 often used is

$$\hat{h}^2 \ = \ 4 \ \hat{\sigma}_s^2/(\hat{\sigma}_s^2 \ + \ \hat{\sigma}_e^2).$$

By definition $0 \ < \ h^2 \ < 1$, the requirement that $\hat{\sigma}_s^2 \ > \ 0$ and $\hat{\sigma}_e^2 \ > \ 0$ does not insure that $\hat{h}^2$ is permissible. For this to be true the permissible range of $\hat{\sigma}_s^2/\hat{\sigma}_e^2$ is 0 to 3^{-1}. This would suggest using an estimation method that guarantees that the estimated ratio falls in the appropriate range.

In the multivariate case a method might be derived along these lines. Let some translation invariant unbiased estimator be the solution to

$$\mathbf{C}\hat{\mathbf{v}} \ = \ \mathbf{q},$$

where $\mathbf{q}$ is a set of quadratics and $\mathbf{Cv}$ is $E(\mathbf{q})$. Then solve these equations subject to a set of inequalities that forces $\hat{\mathbf{v}}$ to fall in the parameter space, as a minimum, all eigenvalues $\geq \ 0$ where $\hat{\mathbf{v}}$ comprises the elements of the variance-covariance matrix.

8 Method For Singular G

When $\mathbf{G}$ is singular we can use a method for EM type REML that is similar to MIVQUE in Section 11.16. We iterate on $\hat{\boldsymbol{\alpha}}'\mathbf{G}_i^*\hat{\boldsymbol{\alpha}}$, and the expectation is $tr\mathbf{G}_i^* \ Var(\hat{\boldsymbol{\alpha}})$. Under the pretense that $\tilde{\mathbf{G}} \ = \ \mathbf{G}$ and $\tilde{\mathbf{R}} \ = \ \mathbf{R}$

$$Var(\hat{\boldsymbol{\alpha}}) \ = \ \tilde{\mathbf{G}}^{-}\mathbf{G}\tilde{\mathbf{G}}^{-} \ - \ \mathbf{C}_{22}.$$

$\mathbf{C}_{22}$ is the lower q^2 submatrix of a g-inverse of the coefficient matrix of (5.51), which has rank, $r(\mathbf{X}) \ + \ r(\mathbf{G})$. Use a g-inverse with $q-$rank($\mathbf{G}$) rows (and cols.) zeroed in the last q rows (and cols.). Let $\mathbf{G}^-$ be a g-inverse of $\mathbf{G}$ with the same $q-$rank($\mathbf{G}$) rows (and cols.) zeroed as in $\mathbf{C}_{22}$. For ML substitute $(\mathbf{GZ}'\mathbf{R}^{-1}\mathbf{ZG})^-$ for $\mathbf{C}_{22}$.