MAJORIZATION ALGORITHMS FOR MIXED MODEL ANALYSIS

JAN DE LEEUW AND GUANGHAN LIU

ABSTRACT.

CONTENTS

1. MODEL

Consider the mixed linear model

$$\underline{\eta} = X\beta + Z\underline{\delta} + \underline{\epsilon}. \tag{1.1}$$

Date: December 28, 2011.

Here $\underline{\delta}$ and $\underline{\epsilon}$ are centered and uncorrelated random vectors, independent of each other. Vectors $\underline{\eta}$ and $\underline{\epsilon}$ have n elements, the matrix X has n rows and m columns, and the matrix Z has n rows and p columns. The fixed (i.e. non-random) vector β has m elements, and the random vector $\underline{\delta}$ has p elements.

Observe that we follow the convention of underlining random variables Hemelrijk [1966]. Other conventions are to use capitals for matrices, and Greek letters for unobserved quantities, which are either fixed (parameters) or random (outcomes, random parameters, error-terms). The only observables in model (**??**) are the matrices X and Z. The vector β contains unknown parameters, and random variables are always unobserved.

Other unknown parameters that interest us are the dispersion matrices

$$\mathbf{V}(\underline{\delta}) \stackrel{\Delta}{=} \Omega, \tag{1.2}$$

$$\mathbf{V}(\underline{\epsilon}) \stackrel{\Delta}{=} \Gamma. \tag{1.3}$$

Then

$$\mathbf{E}(\underline{\eta}) = X\beta, \tag{1.4}$$

$$\mathbf{V}(\underline{\eta}) = Z\Omega Z' + \Gamma \stackrel{\Delta}{=} \Lambda. \tag{1.5}$$

Throughout the paper we are not interested in modelling the higher-order moments, or any other aspect of the distribution, of $\underline{\eta}$. We only model expected values and dispersions. This could be interpreted as us having the multivariate normal in the back of our minds. And, indeed, that is exactly where we have the multivariate normal.

We use $a \models \alpha$ for "a is a realization of α." Finally $a \hat{=} \alpha$ means "a estimates α". We leave the interpretation of both relations unspecified. Some natural ways to combine notation is to write $\hat{\alpha}$ for an estimate of α. This means that Greek symbols with hats are always defined in terms of observed quantities. We can also have $\underline{\hat{\alpha}}$, which is a random variable (unobserved) estimating another random variable $\underline{\alpha}$, also unobserved.

To connect our model (**??**) to observed values of the outcome, we need to assume

$$y \models \underline{\eta}$$

but we won't tell you what this means.

Our notational apparatus may seem to be unnecessarily elaborate, but we think that it is essential to keep the distinction between random variables and their realizations in mind all the time. In many statistics papers calculations are only on random variables, which means that actual observations and statistics are left on the sidelines.

The Dutch Convention, i.e. the use of underlining, may be a bit confusing at first, because in other publications underlining is used to distinguish vectors from scalars. In most situations that we are aware of, however, there is no real need to emphasize the distinction between scalars and vectors. Distinguishing random variables from fixed quantities is very useful, because it shows immediately if a particular formula is used in the model context, or if it is used for actual computations. Any expression which contains an underlined symbol is necessarily part of the model. It does not refer to the actual data or experiment, but to a hypothetical framework (either frequentist or Bayesian). Any expression without underlined symbols can be potentially be used for computation, although it can only be used for actual computation if it does not contain any Greek symbols either (or if all the Greeks have hats). Of course in most statistics papers it is usually clear from the context if the author is talking about the model or about the data, but it does not hurt to be explicit.

The model in which the matrices Ω and Γ are not restricted any further have far too many parameters. Thus we are typically interested in models in which there are restrictions on these matrices. A very common one, for instance, is $\Gamma = \sigma^2 C$, with C a known positive definite matrix, but we shall encounter various other ones.

2. Problem

The problem we study in this paper is estimating the unknown parameters β, Ω, Γ, and the unknown random variables $\underline{\delta}, \underline{\epsilon}$. There is nothing really new in the paper, except our emphasis. We try to steer away as far as possible from probabilistic assumptions, by taking the approach that we are fitting a model to the means and variances. This does not leave much room for inference, which is fine. Secondly, we derive augmentation algorithms in a completely algorithmic way, without using probabilistic notions. This makes our derivations look different from the usual EM-based ones. Thirdly, our matrix calculations, including the ones in the Appendix are very explicit and complete. They systematically use the singular value decomposition, and not the Schur complement, for instance. This also makes them slightly different from the usual ones. Finally, we put our money where our mouth is, and give actual `LISP-STAT` programs for the algorithms discussed in this paper.

Our starting point in model fitting is the simple case in which Λ is a *known* positive definite matrix G. In that case the usual estimate of β, which is both the weighted least squares estimate and the maximum likelihood estimate, is

$$\hat{\beta} = (X'GX)^{-1}X'G^{-1}y. \tag{2.1}$$

Clearly

$$\hat{\beta} \hat{=} (X'G^{-1}X)^{-1}X'G^{-1}\mathbf{E}\,(\underline{\eta}) = \beta, \tag{2.2a}$$

$$\hat{\beta} \models \underline{\hat{\beta}} \stackrel{\Delta}{=} \beta + (X'G^{-1}X)^{-1}X'G^{-1}\{Z\underline{\delta} + \underline{\epsilon}\}, \tag{2.2b}$$

and

$$\mathbf{E}\,(\underline{\hat{\beta}}) = \beta, \tag{2.2c}$$

$$\mathbf{V}\,(\underline{\hat{\beta}}) = (X'G^{-1}X)^{-1}. \tag{2.2d}$$

As we indicated above, the model with Ω and Γ unrestricted has too many parameters for actual data analysis, but the model with Λ completely known does not have enough. There simply is not that much prior knowledge. We need something intermediate in most practical cases.

One easy compromise is to suppose $\Gamma = \sigma^2 C$ and $\Omega = \sigma^2 W$, with C and W known, and σ^2 unknown. Thus $\Lambda = \sigma^2 G$, where $G = ZWZ' + C$. Throughout we suppose $C > 0$, which implies $G > 0$. The classical estimate of β is still given by (2.1), but

$$\mathbf{V}\,(\underline{\hat{\beta}}) = \sigma^2 (X'G^{-1}X)^{-1}.$$

In this case, consequently, we also want to estimate σ^2. This is typically done by defining the *residuals*

$$r \stackrel{\Delta}{=} y - X\hat{\beta} \models \underline{\rho} \stackrel{\Delta}{=} (\mathcal{I}_n - X(X'G^{-1}X)^{-1}X'G^{-1})\{Z\underline{\delta} + \underline{\epsilon}\},$$

and the *residual sum of squares*

$$r'G^{-1}r \models \underline{\rho}'G^{-1}\underline{\rho}.$$

We have

$$\mathbf{V}\,(\underline{\rho}) = \sigma^2\{G - X(X'G^{-1}X)^{-1}X'\}, \mathbf{E}\,(\underline{\rho}'G^{-1}\underline{\rho}) = \sigma^2(n-m),$$

which makes it straightforward to define

$$\hat{\sigma}^2 \stackrel{\Delta}{=} \frac{r'G^{-1}r}{n-m}.$$

Also

$$\hat{\sigma}^2 (X'G^{-1}X)^{-1} \triangleq \mathbf{V}\,(\underline{\hat{\beta}}).$$

Unfortunately, the situation in which we have only one single unknown dispersion parameter σ^2 is also too restricted in many situations.

If Λ is only partially known, then we have to estimate the unknown part in some way or another. This gives

$$\hat{\Lambda} = Z\hat{\Omega}Z' + \hat{\Gamma}.$$

We then use

$$\hat{\beta} = (X'\hat{\Lambda}^{-1}X)^{-1}X'\hat{\Lambda}^{-1}y.blue2$$

The various methods for mixed model estimation differ in the way they compute $\hat{\Lambda}$, but they all use (**??**). Of course (**??**) is no longer true, which makes the statistical stability analysis of the estimates (much) more complicated. The only thing that is still true, formally, is

$$\hat{\beta} \hat{=} (X'\Lambda^{-1}X)^{-1}X'\Lambda^{-1}\mathbf{E}\,(\underline{\eta}) = \beta,$$

but this does not mean much, because we have not specified precisely what $\hat{=}$ means. For extensive overviews of the various different algorithms we refer to the review papers Searle [1971a], Harville [1977], Henderson [1984], Engel [1990], or to the books by Searle Searle [1971b], Searle et al. [1992] and Rao and Kleffe Rao and Kleffe [1988].

3. The BLUE, the BLUP, and the MME

In this section we review some theory which applies to the case in which Λ is known. As we have said, this does not really occur in practice, but much of the statistical computation is motivated by these results anyway. We first estimate Λ, and then we act *as if* Λ is known (and eqaul to the estimate).

Without using normality or esoteric assumptionsd like that, we get on our way by using the Gauss-Laplace-Markov theory of linear estimation. Let us estimate β by using a linear function of the form $A'y$. Clearly $A'y \models A'\underline{\eta}$. Choose A in such a way that cond

$$\mathbf{E}\,(A'\underline{\eta}) = \beta, cond1 \mathbf{V}\,(A'\underline{\eta}) \text{ is minimal}. cond2$$

Any A satisfying (**??**) and (**??**) defines a *best linear unbiased estimate* $\hat{\beta} = Ay$, also known as a BLUE. The *minimal* in (**??**) is meant in the Loewner sense, of course, i.e. for two square symmetric matrices A and B we have $A \geq B$ if $A - B$ is positive semi-definite, and $A > B$ if $A - B$ is positive definite.

Theorem 3.1. *If* $A = \Lambda^{-1}X(X'\Lambda^{-1}X)^{-1}$ *then* Ay *is BLUE.*

Proof. From (**??**) we must have $A'X\beta = \beta$, i.e. $A'X = \mathcal{I}$. Also

$$\mathbf{V}\,(A'\underline{\eta}) = A'\mathbf{V}\,(\underline{\eta})A = A'\Lambda A.$$

Define the matrix

$$T = \Big(\Lambda^{1/2}A \mid \Lambda^{-1/2}X.$$

Then

$$T'T = \Big(A'\Lambda A\mathcal{I}\mathcal{I}X'\Lambda^{-1}X.$$

Because $T'T \geq 0$, we must have,

$$A'\Lambda A - (X'\Lambda^{-1}X)^{-1} \geq 0.$$

For the A mentioned in the Theorem, we have equality, and the lower bound is attained. □

If we want to estimate the random parameters of the model, things become a bit more complicated. In the first place the concept of "estimating" a random variable should be defined carefully. The concept corresponding to the BLUE is the BLUP, or best linear unbiased prediction. The name is due to Goldberger Goldberger [1962], and the silly acronym BLUP has been in use since the seventies Robinson [1991].

Let us find the BLUP of $\underline{\delta}$. The idea is simply to measure prediction error by pred

$$\mathbf{E}\,((A\underline{\eta} + b - \underline{\delta})(A\underline{\eta} + b - \underline{\delta})'), pred1$$

and to choose A and b such that this is minimized, on the condition that

$$\mathbf{E}\,(A\underline{\eta} + b - \underline{\delta}) = 0.pred2$$

The BLUP of $\underline{\epsilon}$ is defined in the same way.

Theorem: The BLUPS of $\underline{\delta}$ and $\underline{\epsilon}$ are given by

$$\hat{\underline{\delta}} = (\Omega^{-1} + Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1}(\underline{\eta} - X\beta), \hat{\underline{\epsilon}} = (\mathcal{I} - Z(\Omega^{-1} + Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1})(\underline{\eta} - X\beta).$$

Proof: From (**??**) we obtain $b = -AX\beta$. Now for (**??**) this gives

$$A\Lambda A' + \Omega - AZ\Omega - \Omega Z'A' \geq \Omega - \Omega Z'\Lambda^{-1}Z\Omega,$$

with equality for

$$\hat{A} = \Omega Z'\Lambda^{-1},$$

i.e.

$$\hat{\underline{\delta}} = \Omega Z'\Lambda^{-1}(\underline{\eta} - X\beta).$$

In the same way the BLUP of $\underline{\epsilon}$ is computed from

$$A\Lambda A' + \Gamma - A\Gamma - \Gamma A' \geq \Gamma - \Gamma\Lambda^{-1}\Gamma,$$

with equality for

$$\hat{A} = \Gamma\Lambda^{-1}.$$

Thus

$$\underline{\hat{\epsilon}} = \Gamma\Lambda^{-1}(\underline{\eta} - X\beta).$$

We can eliminate Λ^{-1} by using the results in the Appendix. This gives the formulas in the Theorem. **Q.E.D.**

Observe that indeed

$$\underline{\eta} = X\beta + Z\underline{\hat{\delta}} + \underline{\hat{\epsilon}}.yes$$

The calculations are, of course, completely within the model. The next step is to definebp

$$\hat{\delta} = (\hat{\Omega}^{-1} + Z'\hat{\Gamma}^{-1}Z)^{-1}Z'\hat{\Gamma}^{-1}(y - X\hat{\beta}), bp1\hat{\epsilon} = (\mathcal{I} - Z(\hat{\Omega}^{-1} + Z'\hat{\Gamma}^{-1}Z)^{-1}Z'\hat{\Gamma}^{-1})(y - X\hat{\beta}).bp2$$

Thenqp

$$\hat{\delta}\hat{=}(\Omega^{-1} + Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1}(y - X\beta) \models \underline{\hat{\delta}}, qp1\hat{\epsilon}\hat{=}(\mathcal{I} - Z(\Omega^{-1} + Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1})(y - X\beta) \models \underline{\hat{\epsilon}}.qp2$$

It is actually possible to derive a slightly more pleasing form for these estimated BLUPS. Let

$$W \stackrel{\Delta}{=} \Gamma^{-1} - \Gamma^{-1}X(X'\Gamma^{-1}X)^{-1}\Gamma^{-1}.$$

rp**Theorem:**

$$\hat{\delta} = (\hat{\Omega}^{-1} + Z'\hat{W}Z)^{-1}Z'\hat{W}y, rp1\hat{\epsilon} = \hat{\Gamma}(\hat{\Lambda}^{-1} - \hat{\Lambda}^{-1}X(X'\hat{\Lambda}^{-1}X)^{-1}X'\hat{\Lambda}^{-1})yrp2$$

Proof: Just substitute the BLUE in (**??**) and simplify. **Q.E.D.**

Observe that from (**??**) we see that $X'\hat{\Gamma}^{-1}\hat{\epsilon} = 0$, which is nice. Also, analogous to (**??**), we still have

$$y = X\hat{\beta} + Z\hat{\delta} + \hat{\epsilon}.$$

It is also possible to derive BLUE and BLUP in a single swoop. For this we use the Henderson's mixed model equations Henderson [1950], i.e. the least squares solutions to the system $C\gamma = z$, where

$$C = \Big(\hat{\Gamma}^{-1/2}X\hat{\Gamma}^{-1/2}Z0\hat{\Omega}^{-1/2}, z = \Big(\hat{\Gamma}^{-1/2}y0, \gamma = (\beta\delta.$$

The MME are reviewed extensively in a recent paper by Robinson Robinson [1991], and they also appear prominently throughout Searle's new book Searle et al. [1992]. Of course the normal equations are

$$\Big(X'\hat{\Gamma}^{-1}XX'\hat{\Gamma}^{-1}ZZ'\hat{\Gamma}^{-1}X\hat{\Omega}^{-1} + Z'\hat{\Gamma}^{-1}Z\Big(\beta\delta = \Big(X'\hat{\Gamma}^{-1}yZ'\hat{\Gamma}^{-1}y.$$

Solving the normal equations can be done in two steps. We first solve for β in terms of δ, and vice versa. This gives

$$\hat{\beta} = (X'\hat{\Gamma}^{-1}X)^{-1}X'\hat{\Gamma}^{-1}(y - Z\hat{\delta}), \hat{\delta} = (\hat{\Omega}^{-1} + Z'\hat{\Gamma}^{-1}Z)^{-1}Z'\hat{\Gamma}^{-1}(y - X\hat{\beta}).$$

Now substitute, and simplify. This gives, after some work,

$$\hat{\beta} = (X'\hat{\Lambda}^{-1}X)^{-1}X'\hat{\Lambda}^{-1}y,$$

which is the BLUE, and

$$\delta = (\hat{\Omega}^{-1} + Z'\hat{W}Z)^{-1}Z'\hat{W}y,$$

which is the BLUP.

4. Deviance of the Data

The deviance (i.e. twice the negative log-likelihood) of the observed y is (except for uninteresting constants)

$$\Delta(\beta, \Gamma, \Omega) = \log \mid \Lambda \mid +(y - X\beta)'\Lambda^{-1}(y - X\beta).ddev$$

We compute maximum likelihood estimates by minimizing this deviance over all the unknown parameters. By minimizing over β for fixed Ω and Γ we can also define the *marginal deviance.* The minimum over β is of course attained by the BLUE.

Lemma 1:

$$(y - X\beta)'\Lambda^{-1}(y - X\beta) = \min_{\delta} \quad (y - X\beta - Z\delta)'\Gamma^{-1}(y - X\beta - Z\delta) + \delta'\Omega^{-1}\delta.$$

Proof: The minimum over δ is attained at the BLUP

$$\hat{\delta} = (\Omega^{-1} + Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1}(y - X\beta),$$

and substituting this gives the required result. **Q.E.D.**

Theorem 1:

$$\Delta(\beta, \Omega, \Gamma) = \log \mid \Gamma \mid + \log \mid \Omega \mid + \log \mid \Omega^{-1} + Z'\Gamma^{-1}Z \mid + + \min_{\delta} (y - X\beta - Z\delta)'\Gamma^{-1}(y - X\beta - Z\delta) + \delta'\Omega$$

Proof: We use the expression for the determinant from Lemma A1 in the Appendix, and the result from the previous lemma. **Q.E.D.**

Corrolary: Suppose $(\hat{\beta}, \hat{\delta})$ are the BLUE and BLUP, i.e. the solutions to the MME. Then

$$\Delta^{\star}(\Omega, \Gamma) = \log \mid \Gamma \mid + \log \mid \Omega \mid + \log \mid \Omega^{-1} + Z'\Gamma^{-1}Z \mid + + (y - X\hat{\beta} - Z\hat{\delta})'\Gamma^{-1}(y - X\hat{\beta} - Z\hat{\delta}) + \hat{\delta}'\Omega^{-1}\hat{\delta}.$$

5. Deviance of Residuals

Suppose K is any matrix satisfying $K'X = 0$ and $\text{rank}(K) = n - \text{rank}(X)$. Thus K is a basis for the null-space of X. Define $\underline{\rho} = K'\underline{\eta}$ and $r = K'y$. Clearly

$$r \models \underline{\rho} \sim \mathcal{N}(0, K'\Lambda K),$$

and the distribution of ρ does not contain β any more. The deviance is

$$\Delta^{\circ}(\Omega, \Gamma) = \log \mid K'\Lambda K \mid + r'(K'\Lambda K)^{-1} r.rdev$$

We compute REML estimates Patterson and Thomson [1971] by minimizing this deviance.

Lemma:

$$r'(K'\Lambda K)^{-1}r = y'K(K'\Lambda K)^{-1}K'y = \min_{\beta} (y - X\beta)'\Lambda^{-1}(y - X\beta).$$

Proof: Obviously

$$\min_{\beta}(y - X\beta)'\Lambda^{-1}(y - X\beta) = \min_{K'h=0}(y - h)'\Lambda^{-1}(y - h).$$

By using Lagrange multipliers ν this amounts to solving

$$\Lambda^{-1}(y-h) = K\nu, K'h = 0.$$

The solution to this system is

$$\hat{\nu} = (K'\Lambda K)^{-1}r, \hat{h} = y - \Lambda K(K'\Lambda K)^{-1}r.$$

It follows that

$$\min_{K'h=0}(y-h)'\Lambda^{-1}(y-h) = \hat{\nu}'K\Lambda K'\hat{\nu} = r'(K'\Lambda K)^{-1}r.$$

Q.E.D.

Lemma:

$$\log \mid K'\Lambda K \mid = \log \mid K'\Gamma K \mid + \log \mid \Omega \mid + \log \mid \Omega^{-1} + Z'WZ \mid .$$

Proof: Obviously $K'\Lambda K = K'Z\Omega Z'K + K'\Gamma K$. Now apply Lemma A1 in the Appendix. This gives

$$\log \mid K'\Lambda K \mid = \log \mid K'\Gamma K \mid + \log \mid \Omega \mid + \log \mid \Omega^{-1} + Z'K(K'\Gamma K)^{-1}K'Z \mid .$$

Consider the matrix

$$T = \Big(\Gamma^{1/2}K(K'\Gamma K)^{-1/2} \mid \Gamma^{-1/2}X(X'\Gamma^{-1}X)^{-1/2}.$$

Then $T'T = \mathcal{I}_n$, which implies

$$TT' = \Gamma^{1/2}K(K'\Gamma K)^{-1}K'\Gamma^{1/2} + \Gamma^{-1/2}X(X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1/2} = \mathcal{I}_n.$$

It follows that

$$K(K'\Gamma K)^{-1}K' = \Gamma^{-1} - \Gamma^{-1}X(X'\Gamma^{-1}X)^{-1}X'\Gamma^{-1} = W.$$

Q.E.D.

Corrolary: Suppose $(\hat{\beta}, \hat{\delta})$ are the BLUE and BLUP, i.e. the solutions to the MME. Then

$$\Delta^{\circ}(\Omega, \Gamma) = \log \mid K'\Gamma K \mid + \log \mid \Omega \mid + \log \mid \Omega^{-1} + Z'WZ \mid + (y - X\hat{\beta} - Z\hat{\delta})'\Gamma^{-1}(y - X\hat{\beta} - Z\hat{\delta}) + \hat{\delta}'\Omega^{-1}\hat{\delta}.$$

6. Algorithm

The reductions and derivations in the previous sections have lead us to the following two problems.

The problem of minimizing the data-deviance (the FIML problem) has been shown to be equivalent to minimizing

$$\Delta^{\star}(\Omega,\Gamma,\beta,\delta) \stackrel{\Delta}{=} \log \mid \Gamma \mid + \log \mid \Omega \mid + \log \mid \Omega^{-1} + Z'\Gamma^{-1}Z \mid + + (y - X\beta - Z\delta)'\Gamma^{-1}(y - X\beta - Z\delta) + \delta'\Omega^{-1}$$

The problem of minimizing the residual-deviance (the REML problem) amounts to minimizing

$$\Delta^{\circ}(\Omega,\Gamma,\beta,\delta) \stackrel{\Delta}{=} \log \mid K'\Gamma K \mid + \log \mid \Omega \mid + \log \mid \Omega^{-1} + Z'WZ \mid + + (y - X\beta - Z\delta)'\Gamma^{-1}(y - X\beta - Z\delta) + \delta'$$

This formulation suggests a simple iterative algorithm. We start with initial estimates $\Omega^{(0)}$ and $\Gamma^{(0)}$ satisfying the constraints. We then solve the MME using these estimates, to find first estimates $\beta^{(0)}$ and $\delta^{(0)}$. This is the same thing as minimizing $\Delta^{\star}(\Omega^{(0)},\Gamma^{(0)},\beta,\delta)$ or $\Delta^{\circ}(\Omega^{(0)},\Gamma^{(0)},\beta,\delta)$. We then plug these estimates of β and δ into the function and minimize these over Ω and Γ satisfying the constraints to find updates for those quantities. Thus $\Omega^{(1)}$ and $\Gamma^{(1)}$ minimize $\Delta^{\star}(\Omega,\Gamma,\beta^{(0)},\delta^{(0)})$ or $\Delta^{\circ}(\Omega,\Gamma,\beta^{(0)},\delta^{(0)})$ over $\Omega \in \mathcal{W}$ and $\Gamma \in \mathcal{C}$. General theorems on block relaxation de Leeuw [1993],Oberhofer and Kmenta [1974] can be used to prove convergence of this procedure.

From the computational point of view, solving the MME is straightforward. Engel summarizes much of the computational wisdom Engel [1990]. The second subproblem, which calls for minimization over the variance and covariance components, is less straightforward. To some extend the best strategy will depend on the precise form of the restrictions. We first use *majorization* to simplify the problem somewhat more.

Majorization Lemma: If $X > 0$ then

$$\log \mid X \mid = \min_{Y>0} \log \mid Y \mid + \operatorname{tr} Y^{-1}(X - Y),$$

with the unique minimum attained at $\hat{Y} = X$.

Proof: In stead of giving a proof, we simply translate the result to statistical terminology. Suppose X is the sample covariance matrix. The lemma then states that X is the maximum likelihood estimate of the population covariance matrix if we are sampling from a multinormal distribution. This is classical. **Q.E.D.**

Let us now apply the Majorization Lemma to FIML estimation. We find

$$\log \mid \Omega^{-1} + Z'\Gamma^{-1}Z \mid == \min_{\tilde{\Omega}>0,\tilde{\Gamma}>0} \log \mid \tilde{\Omega}^{-1} + Z'\tilde{\Gamma}^{-1}Z \mid + \mathrm{tr}(\tilde{\Omega}^{-1} + Z'\tilde{\Gamma}^{-1}Z)^{-1}(\Omega^{-1} + Z'\Gamma^{-1}Z) - p.$$

It follows that FIML estimates can be computed by minimizing

$$\Delta^\star(\Omega, \Gamma, \tilde{\Omega}, \tilde{\Gamma}, \beta, \delta) \stackrel{\Delta}{=} \{\log \mid \Gamma \mid + \log \mid \Omega \mid + + \log \mid \tilde{\Omega}^{-1} + Z'\tilde{\Gamma}^{-1}Z \mid + \mathrm{tr}(\tilde{\Omega}^{-1} + Z'\tilde{\Gamma}^{-1}Z)^{-1}(\Omega^{-1} + Z'\Gamma^{-1}Z)$$

Block-relaxation can be implemented in many ways if there are many blocks of variables. We distinguish three subproblems. minimize $\Delta^\star$ over (β, δ) for fixed (Ω, Γ) and $(\tilde{\Omega}, \tilde{\Gamma})$. This is just solving the MME. minimize $\Delta^\star$ over $(\tilde{\Omega}, \tilde{\Gamma})$ for fixed (Ω, Γ) and (β, δ). By the Majorization Lemma the solution is just (Ω, Γ). minimize $\Delta^\star$ over (Ω, Γ) for fixed $(\tilde{\Omega}, \tilde{\Gamma})$ and (β, δ). This subproblem may still be problematic. For the complete algorithm we cycle through the three subproblems. Again, there are various alternatives. We can alternate subproblems (2) and (3) a number of times before we solve another (1). We can even alternate (2) and (3) until convergence before we solve another (1). Usually, however, we solve each subproblem in turn.

It will be clear that the corresponding majorization function for REML is

$$\Delta^\circ(\Omega, \Gamma, \tilde{\Omega}, \tilde{\Gamma}, \beta, \delta) \stackrel{\Delta}{=} \{\log \mid K'\Gamma K \mid + \log \mid \Omega \mid + + \log \mid \tilde{\Omega}^{-1} + Z'\tilde{W}Z \mid + + \mathrm{tr}(\tilde{\Omega}^{-1} + Z'\tilde{W}Z)^{-1}(\Omega^{-1} + Z'W$$

7. Multilevel Models

In a multilevel model we have G groups of observations. The model for group g is

$$\underline{\eta}_g = Z_g \underline{\gamma}_g + \underline{\epsilon}_g . indmod$$

But we also have a group-level model, which is

$$\underline{\gamma}_g = U_g \beta + \underline{\delta}_g . grpmod$$

We assume that the $\underline{\delta}_g$ and $\underline{\epsilon}_g$ are all uncorrelated, and we set

$$\mathbf{V}(\underline{\epsilon}_g) = \sigma_g^2 \mathcal{I}_g, \mathbf{V}(\underline{\delta}_g) = \Xi.$$

By substituting (**??**) in (**??**) we find

$$\underline{\eta}_g = Z_g U_g \beta + Z_g \underline{\delta}_g + \underline{\epsilon}_g. sinmod$$

Let

$$\underline{\eta} = \Big(\underline{\eta}_1 \underline{\eta}_2 \vdots \underline{\eta}_G, \underline{\delta} = \Big(\underline{\delta}_1 \underline{\delta}_2 \vdots \underline{\delta}_G, X = \Big(Z_1 U_1 Z_2 U_2 \vdots Z_G U_G.$$

Also, we use $\oplus$ for the *direct sum,* i.e. the direct sum of a number of matrices is a supermatrix with these matrices as the diagonal submatrices, and zeroes everywhere else. Let

$$Z = Z_1 \oplus Z_2 \oplus \cdots \oplus Z_G.$$

Then

$$\underline{\eta} = X\beta + Z\underline{\delta} + \underline{\epsilon},$$

which is the mixed model we started with. But of course it has restrictions on the dispersion matrices. They have to be of the form

$$\Gamma = \sigma_1^2 \mathcal{I}_1 \oplus \sigma_2^2 \mathcal{I}_2 \oplus \cdots \oplus \sigma_G^2 \mathcal{I}_G, \Omega = \underbrace{\Xi \oplus \Xi \oplus \cdots \oplus \Xi}_{\text{G times}}.$$

Also, X is in the space spanned by the columns of Z. In fact, if

$$U = \Big(U_1 U_2 \vdots U_g,$$

then $X = ZU$. This has important computational consequences.

Appendix A. Matrix Results

In this appendix we collect some of the main matrix results needed in the body of the paper. None of the results are entirely new. We systematically use the *singular value decomposition* in our proofs, which makes it easy to keep track of the singularities of the matrices, and which gives results that may be somewhat more general than known results.

Suppose $\Lambda = Z\Omega Z' + \Gamma$, as before, with $\Gamma > 0$ and Z an $n \times p$ matrix. Let $r \triangleq \mathrm{rank}(Z)$ and $\tilde{Z} \triangleq \Gamma^{-1/2} Z \Omega^{1/2}$, and $r \triangleq \mathrm{rank}(\tilde{Z})$. To avoid trivial notational complications we assume that $r > 0$. Also suppose $\tilde{Z} = K\Phi L'$ is the *singular value decomposition* of $\tilde{Z}$. Thus K is $n \times r$ with $K'K = \mathcal{I}_r$, L is $p \times r$ with $L'L = \mathcal{I}_r$, and Φ is $r \times r$, diagonal and positive definite. Let

$$P \triangleq \tilde{Z}(\tilde{Z}'\tilde{Z})^{+}\tilde{Z}' = KK', Q \triangleq \mathcal{I}_n - P.$$

Theorem Determinant: If $\Lambda = Z\Omega Z' + \Gamma$, with $\Gamma > 0$ and $\Omega > 0$, then

$$\log \mid \Lambda \mid = \log \mid \Gamma \mid + \log \mid \Omega \mid + \log \mid \Omega^{-1} + Z'\Gamma^{-1} Z \mid .$$

If, in addition, Z is of full column rank, then

$$\log \mid \Lambda \mid = \log \mid \Gamma \mid + \log \mid Z'\Gamma^{-1} Z \mid + \log \mid \Omega + (Z'\Gamma^{-1} Z)^{-1} \mid .$$

Proof: Clearly

$$\Gamma^{-1/2}\Lambda\Gamma^{-1/2} = \tilde{Z}\tilde{Z}' + \mathcal{I}_n.a1$$

Also,

$$\mid \tilde{Z}\tilde{Z}' + \mathcal{I}_n \mid = \mid \tilde{Z}'\tilde{Z} + \mathcal{I}_p \mid,$$

and thus

$$\log \mid \Lambda \mid = \log \mid \Gamma \mid + \log \mid \mathcal{I}_p + \Omega^{1/2} Z'\Gamma^{-1} Z \Omega^{1/2} \mid .$$

If Ω is non-singular, this can be written as stated in the theorem.

On the other hand

$$\Gamma^{-1/2}\Lambda\Gamma^{-1/2} = K\Phi^2 K' + \mathcal{I}_n = K(\Phi^2 + \mathcal{I}_r)K' + Q,$$

which shows that

$$\log \mid \Lambda \mid = \log \mid \Gamma \mid + \log \mid \Phi^2 + \mathcal{I}_r \mid = \log \mid \Gamma \mid + \log \mid \Phi^2 \mid + \log \mid \mathcal{I}_p + (\tilde{Z}'\tilde{Z})^{+} \mid .$$

If Ω is non-singular, then this simplifies to

$$\log \mid \Lambda \mid = \log \mid \Gamma \mid + \log \mid \Phi^2 \mid - \log \mid \Omega \mid + \log \mid \Omega + (Z'\Gamma^{-1}Z)^+ \mid .$$

If Z is of full column rank then this simplifies to the second statement in the theorem. **Q.E.D.**

Theorem Inverse: If $\Lambda = Z\Omega Z' + \Gamma$, with $\Gamma > 0, \Omega > 0$, and Z of full column-rank, then

$$\Lambda^{-1} == \Gamma^{-1} - \Gamma^{-1}Z(\Omega^{-1} + Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1} == (Z')^+\{\Omega + (Z'\Gamma^{-1}Z)^{-1}\}^+Z^+ + \{\Gamma^{-1} - \Gamma^{-1}Z(Z'\Gamma^{-1}Z)^{-1}Z'\Gamma$$

Proof: In the first place,

$$\Gamma^{1/2}\Lambda^{-1}\Gamma^{1/2} = (K\Phi^2K' + \mathcal{I}_n)^{-1}.$$

We can easily verify, by direct multiplication, that

$$(K\Phi^2K' + \mathcal{I}_n)^{-1} = \mathcal{I}_n - K\Phi(\mathcal{I}_r + \Phi^2)^{-1}\Phi K'.$$

Now use

$$\tilde{Z} = \Gamma^{-1/2}Z\Omega^{1/2} = K\Phi L',$$

to find

$$\Lambda^{-1} = \Gamma^{-1} - \Gamma^{-1}Z\Omega^{1/2}L(\mathcal{I}_r + \Phi^2)^{-1}L'\Omega^{1/2}Z'\Gamma^{-1} == \Gamma^{-1} - \Gamma^{-1}Z\Omega^{1/2}(LL' + \Omega^{1/2}Z'\Gamma^{-1}Z\Omega^{1/2})^+\Omega^{1/2}Z'\Gamma^{-}$$

If $\Omega > 0$ this becomes

$$\Lambda^{-1} = \Gamma^{-1} - \Gamma^{-1}Z(\Omega^{-1/2}LL'\Omega^{-1/2} + Z'\Gamma^{-1}Z)^+Z'\Gamma^{-1}.$$

and if in addition $\tilde{Z}$ is of full column rank, then

$$\Lambda^{-1} = \Gamma^{-1} - \Gamma^{-1}Z(\Omega^{-1} + Z'\Gamma^{-1}Z)^{-1}Z'\Gamma^{-1}.$$

We can also proceed along a different route. It starts with

$$\Lambda^{-1} = \Gamma^{-1/2}\{K(\Phi^2 + \mathcal{I}_r)K' + Q\}^{-1}\Gamma^{-1/2} == \Gamma^{-1/2}K(\Phi^2 + \mathcal{I}_r)^{-1}K'\Gamma^{-1/2} + \Gamma^{-1/2}Q\Gamma^{-1/2} == (\Omega^{1/2}Z')^+\{LL$$

This simplifies if $\Omega > 0$ to

$$\Lambda^{-1} = (Z')^+\{\Omega^{1/2}LL'\Omega^{1/2} + (Z'\Gamma^{-1}Z)^+\}^+Z^+ + \{\Gamma^{-1} - \Gamma^{-1}Z(Z'\Gamma^{-1}Z)^+Z'\Gamma^{-1}\}.$$

If you do not like the LL' in these formulas, it can be replaced by

$$LL' = \Omega^{1/2}Z'(Z\Omega Z')^+Z\Omega^{1/2}.$$

If $\tilde{Z}$ is of full column rank, then finally we have the second result in the theorem. **Q.E.D.**

The first identity in Theorem Inverse is a classical result of Duncan Duncan [1944], recently discussed in the variance components context by De Hoog, Speed, and Williams de Hoog et al. [1990]. The second identity generalizes a result used heavily by Swamy Swamy [1971] in random coefficient modeling.

Lemma A2: If $A = B + TCT'$, with $B > 0$ and $C > 0$, then

$$y'A^{-1}y = \min_x \{(y - Tx)'B^{-1}(y - Tx) + x'C^{-1}x\},$$

Proof: We actually proof a slightly more general result. We solve

$$\min_x \{(y - Tx)'D(y - Tx) + x'Ex\},$$

with $D \geq 0$ and $E \geq 0$. Clearly the minimum at attained for any x that solves $(T'DT + E)x = T'Dy$. The minimum is equal to

$$y'[D - DT(T'DT + E)^{+}TD]y.$$

Thus we must have $A = [D - DT(T'DT + E)^{+}TD]^{-1}$.

References

F.R. de Hoog, T.P. Speed, and E.R. Williams. On a matrix identity associated with generalized least squares. *Linear Algebra and its Applications*, 127:449–456, 1990.

J. de Leeuw. Majorization methods in statistical computation. Preprint ??, UCLA Statistics, Los Angeles, CA, 1993.

W.J. Duncan. Some devices for the solution of large sets of simultaneous linear equations (with an appendix on the reciprocation of partitioned matrices). *Philosophical Magazine (7)*, 35:660–670, 1944.

B. Engel. The analysis of unbalanced linear models with variance components. *Statistica Neerlandica*, 44:195–219, 1990.

A.S. Goldberger. Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, 57: 369–375, 1962.

D.A. Harville. Maximum likelihood approaches to variance component estimation and related problems. *Journal of the American Statistical Association*, 72:320–340, 1977.

J. Hemelrijk. Underlining random variables. *Statistica Neerlandica*, 20: 1–7, 1966.

C. R. Henderson. ANOVA, MIVQUE, REML, and ML algorithms for estimation of variances and covariances. In H.A. David and H.T. David, editors, *Statistics: an appraisal*. Iowa State University Press, Ames, Iowa, 1984.

C.R. Henderson. Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics*, 21:309–310, 1950.

W. Oberhofer and J. Kmenta. A general procedure for obtaining maximum likelihood estimates in generalized regression models. *Econometrica*, 42: 579–590, 1974.

H.D. Patterson and R. Thomson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554, 1971.

C.R. Rao and J. Kleffe. *Estimation of Variance Components and applications*. North Holland Publishing Company, Amsterdam, The Netherlands, 1988.

G.K. Robinson. That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science*, 6:15–51, 1991.

S.R. Searle. Topics in variance component estimation. *Biometrics*, 27:1–76, 1971a.

S.R. Searle. *Linear Models*. Wiley, New York, NY, 1971b.

S.R. Searle, G. Casella, and C.E. McCulloch. *Variance Components*. Wiley, New York, NY, 1992.

P.A.V.B Swamy. *Statistical inference in random coefficients regression models*. Springer Verlag, New York, NY, 1971.

Department of Statistics, University of California, Los Angeles