Design Document for Random Effects Aster Models

Charles J. Geyer

June 11, 2021

Abstract

This design document works out details of approximate maximum likelihood estimation for aster models with random effects. Fixed and random effects are estimated by penalized log likelihood. Variance components are estimated by integrating out the random effects in the Laplace approximation of the complete data likelihood (this can be done analytically) and maximizing the resulting approximate missing data likelihood. A further approximation treats the second derivative matrix of the cumulant function of the exponential family where it appears in the approximate missing data log likelihood as a constant (not a function of parameters). Then first and second derivatives of the approximate missing data log likelihood can be done analytically. Minus the second derivative matrix of the approximate missing data log likelihood is treated as approximate Fisher information and used to estimate standard errors.

1 Theory

Aster models (Geyer, Wagenius and Shaw, 2007; Shaw, Geyer, Wagenius, Hangelbroek, and Etterson, 2008) have attracted much recent attention. Several researchers have raised the issue of incorporating random effects in aster models, and we do so here.

1.1 Complete Data Log Likelihood

Although we are particularly interested in aster models (Geyer et al., 2007), our theory works for any exponential family model. The log likelihood can be written

$$l(\varphi) = y^T \varphi - c(\varphi),$$

where y is the canonical statistic vector, φ is the canonical parameter vector, and the cumulant function c satisfies

$$\mu(\varphi) = E_\varphi(y) = c'(\varphi) \tag{1}$$

$$W(\varphi) = \operatorname{var}_\varphi(y) = c''(\varphi) \tag{2}$$

where $c'(\varphi)$ denotes the vector of first partial derivatives and $c''(\varphi)$ denotes the matrix of second partial derivatives.

We assume a canonical affine submodel with random effects determined by

$$\varphi = a + M\alpha + Zb, \tag{3}$$

where a is a known vector, M and Z are known matrices, b is a normal random vector with mean vector zero and variance matrix D. The vector a is called the *offset vector* and the matrices M and Z are called the *model matrices* for fixed and random effects, respectively, in the terminology of the R function `glm`. (The vector a is called the *origin* in the terminology of the R function `aster`. *Design matrix* is alternative terminology for `model matrix`.) The matrix D is assumed to be diagonal, so the random effects are independent random variables. The diagonal components of D are called *variance components* in the classical terminology of random effects models (Searle et al., 1992). Typically the components of b are divided into blocks having the same variance (Searle et al., 1992, Section 6.1), so there are only a few variance components but many random effects, but nothing in this document uses this fact.

The unknown parameter vectors are α and ν, where ν is the vector of variance components. Thus D is a function of ν, although this is not indicated by the notation.

The "complete data log likelihood" (i. e., what the log likelihood would be if the random effect vector b were observed) is

$$l_c(\alpha, b, \nu) = l(a + M\alpha + Zb) - \tfrac{1}{2} b^T D^{-1} b - \tfrac{1}{2} \log \det(D) \tag{4}$$

in case none of the variance components are zero. We deal with the case of zero variance components in Sections 1.9, 1.10, and 1.11 below.

1.2 Missing Data Likelihood

Ideally, inference about the parameters should be based on the *missing data likelihood*, which is the complete data likelihood with random effects b integrated out

$$L_m(\alpha, \nu) = \int e^{l_c(\alpha, b, \nu)} \, db \tag{5}$$

Maximum likelihood estimates (MLE) of α and ν are the values that maximize (5). However MLE are hard to find. The integral in (5) cannot be done analytically, nor can it be done by numerical integration except in very simple cases. There does exist a large literature on doing such integrals by ordinary or Markov chain Monte Carlo (Thompson and Guo, 1991; Geyer and Thompson, 1992; Geyer, 1994; Shaw, Promislow, Tatar, Hughes, and Geyer, 1999; Shaw, Geyer and Shaw, 2002; Sung and Geyer, 2007), but these methods take a great deal of computing time and are difficult for ordinary users to apply. We wish to avoid that route if at all possible.

1.3 A Digression on Minimization

The theory of constrained optimization (Section 1.10 below) has a bias in favor of minimization rather than maximization. The explication below will be simpler if we switch now from maximization to minimization (minimizing minus the log likelihood) rather than switch later.

1.4 Laplace Approximation

Breslow and Clayton (1993) proposed to replace the integrand in (5) by its Laplace approximation, which is a normal probability density function so the random effects can be integrated out analytically. Let b^* denote the result of maximizing (4) considered as a function of b for fixed α and ν. Then $-\log L_m(\alpha, \nu)$ is approximated by

$$q(\alpha, \nu) = \tfrac{1}{2} \log \det[\kappa''(b^*)] + \kappa(b^*)$$

where

$$\begin{aligned} \kappa(b) &= -l_c(a + M\alpha + Zb) \\ \kappa'(b) &= -Z^T[y + \mu(a + M\alpha + Zb)] + D^{-1}b \\ \kappa''(b) &= Z^T W(a + M\alpha + Zb)Z + D^{-1} \end{aligned}$$

Hence

$$\begin{aligned} q(\alpha, \nu) &= -l_c(\alpha, b^*, \nu) + \tfrac{1}{2} \log \det\bigl[\kappa''(b^*)\bigr] \\ &= -l(a + M\alpha + Zb^*) + \tfrac{1}{2}(b^*)^T D^{-1} b^* + \tfrac{1}{2} \log \det(D) \\ &\quad + \tfrac{1}{2} \log \det\bigl[Z^T W(a + M\alpha + Zb^*)Z + D^{-1}\bigr] \qquad (6) \\ &= -l(a + M\alpha + Zb^*) + \tfrac{1}{2}(b^*)^T D^{-1} b^* \\ &\quad + \tfrac{1}{2} \log \det\bigl[Z^T W(a + M\alpha + Zb^*)ZD + I\bigr] \end{aligned}$$

where I denotes the identity matrix of the appropriate dimension (which must be the same as the dimension of D for the expression it appears in to make sense), where b^* is a function of α and ν and D is a function of ν, although this is not indicated by the notation, and where the last equality uses the rule sum of logs is log of product and the rule product of determinants is determinant of matrix product (Harville, 1997, Theorem 13.3.4).

Since minus the log likelihood of an exponential family is a convex function (Barndorff-Nielsen, 1978, Theorem 9.1) and the middle term on the right-hand side of (4) is a strictly convex function, it follows that (4) considered as a function of b for fixed α and ν is a strictly convex function. Moreover, this function has bounded level sets, because the middle term on the right-hand side of (4) does. It follows that there is unique global minimizer (Rockafellar and Wets, 2004, Theorems 1.9 and 2.6). Thus $b^*(\alpha, \nu)$ is well defined for all values of α and ν.

The key idea is to use (6) as if it were the log likelihood for the unknown parameters (α and ν), although it is only an approximation. However, this is also problematic. In doing likelihood inference using (6) we need first and second derivatives of it (to calculate Fisher information), but W is already the second derivative matrix of the cumulant function, so first derivatives of (6) would involve third derivatives of the cumulant function and second derivatives of (6) would involve fourth derivatives of the cumulant function. For aster models there are no published formulas for derivatives higher than second of the aster model cumulant function nor does software (the R package `aster`, Geyer, 2015) provide such — the derivatives do, of course, exist because every cumulant function of a regular exponential family is infinitely differentiable at every point of the canonical parameter space (Barndorff-Nielsen, 1978, Theorem 8.1) — they are just not readily available. Breslow and Clayton (1993) noted the same problem in the context of GLMM, and proceeded as if W were a constant function of its argument, so all derivatives of W were zero. This is not a bad approximation because "in asymptopia" the aster model log likelihood is exactly quadratic and W is a constant function, this being a general property of likelihoods (Geyer, 2013). Hence we adopt this idea too, more because we are forced to by the difficulty of differentiating W than by our belief that we are "in asymptopia."

This leads to the following idea. Rather than basing inference on (6), we actually use

$$q(\alpha, \nu) = -l(a + M\alpha + Zb^*) + \tfrac{1}{2}(b^*)^T D^{-1} b^* + \tfrac{1}{2} \log \det\bigl[Z^T \widehat{W} Z D + I\bigr] \quad (7)$$

where $\widehat{W}$ is a constant matrix (not a function of α and ν). This makes sense for any choice of $\widehat{W}$ that is symmetric and positive semidefinite, but we will

choose $\widehat{W}$ that are close to $W(a + M\hat{\alpha} + Z\hat{b})$, where $\hat{\alpha}$ and $\hat{\nu}$ are the joint minimizers of (6) and $\hat{b} = b^*(\hat{\alpha}, \hat{\nu})$. Note that (7) is a redefinition of $q(\alpha, \nu)$. Hereafter we will no longer use the definition (6).

1.5 A Key Concept

Introduce

$$p(\alpha, b, \nu) = -l(a + M\alpha + Zb) + \tfrac{1}{2} b^T D^{-1} b + \tfrac{1}{2} \log\det\left[Z^T \widehat{W} Z D + I\right] \quad (8)$$

where, as the left-hand side says, α, b, and ν are all free variables and, as usual, D is a function of ν, although the notation does not indicate this. Since the terms that contain b are the same in both (4) and (8), b^* can also be defined at the result of minimizing (8) considered as a function of b for fixed α and ν. Thus (7) is a profile of (8) and $(\hat{\alpha}, \hat{b}, \hat{\nu})$ is the joint minimizer of (8).

Since $p(\alpha, b, \nu)$ is a much simpler function than $q(\alpha, \nu)$, the latter having no closed form expression and requiring an optimization as part of each evaluation, it is much simpler to find $(\hat{\alpha}, \hat{b}, \hat{\nu})$ by minimizing the former rather than the latter.

1.6 A Digression on Partial Derivatives

Let $f(\alpha, b, \nu)$ be a scalar-valued function of three vector variables. We write partial derivative vectors using subscripts: $f_\alpha(\alpha, b, \nu)$ denotes the vector of partial derivatives with respect to components of α. Our convention is that we take this to be a column vector. Similarly for $f_b(\alpha, b, \nu)$. We also use this convention for partial derivatives with respect to single variables: $f_{\nu_k}(\alpha, b, \nu)$, which are, of course, scalars. We use this convention for any scalar-valued function of any number of vector variables.

We continue this convention for second partial derivatives: $f_{\alpha b}(\alpha, b, \nu)$ denotes the matrix of partial derivatives having i, j component that is the (mixed) second partial derivative of f with respect to α_i and b_j. Thus the row dimension of $f_{\alpha b}(\alpha, b, \nu)$ is the dimension of α, the column dimension is the dimension of b, and $f_{b\alpha}(\alpha, b, \nu)$ is the transpose of $f_{\alpha b}(\alpha, b, \nu)$.

This convention allows easy indication of points at which partial derivatives are evaluated. For example, $f_{\alpha b}(\alpha, b^*, \nu)$ indicates that b^* is plugged in for b in the expression for $f_{\alpha b}(\alpha, b, \nu)$.

We also use this convention of subscripts denoting partial derivatives with vector-valued functions. If $f(\alpha, b, \nu)$ is a column-vector-valued function of vector variables, then $f_\alpha(\alpha, b, \nu)$ denotes the matrix of partial derivatives

having i, j component that is the partial derivative of the i-th component of $f_\alpha(\alpha, b, \nu)$ with respect to α_j. Thus the row dimension of $f_\alpha(\alpha, b, \nu)$ is the dimension of $f(\alpha, b, \nu)$ and the column dimension is the dimension of α.

1.7 First Derivatives

Start with (8). Its derivatives are

$$p_\alpha(\alpha, b, \nu) = -M^T \big[y - \mu(a + M\alpha + Zb) \big] \tag{9}$$

$$p_b(\alpha, b, \nu) = -Z^T \big[y - \mu(a + M\alpha + Zb) \big] + D^{-1} b \tag{10}$$

and

$$p_{\nu_k}(\alpha, b, \nu) = -\tfrac{1}{2} b^T D^{-1} E_k D^{-1} b + \tfrac{1}{2} \operatorname{tr}\Big(\big[Z^T \widehat{W} Z D + I \big]^{-1} Z^T \widehat{W} Z E_k \Big) \tag{11}$$

where

$$E_k = A_{\nu_k}(\nu) \tag{12}$$

is the diagonal matrix whose components are equal to one if the corresponding components of D are equal to ν_k by definition (rather than by accident when some other component of ν also has the same value) and whose components are otherwise zero. The formula for the derivative of a matrix inverse comes from Harville (1997, Chapter 15, Equation 8.15). The formula for the derivative of the log of a determinant comes from Harville (1997, Chapter 15, Equation 8.6).

The estimating equation for b^* can be written

$$p_b\big(\alpha, b^*, \nu\big) = 0 \tag{13}$$

and by the multivariate chain rule (Browder, 1996, Theorem 8.15) we have

$$\begin{aligned} q_\alpha(\alpha, \nu) &= p_\alpha(\alpha, b^*, \nu) + b^*_\alpha(\alpha, \nu)^T p_b(\alpha, b^*, \nu) \\ &= p_\alpha(\alpha, b^*, \nu) \end{aligned} \tag{14}$$

by (13), and

$$\begin{aligned} q_{\nu_k}(\alpha, \nu) &= b^*_{\nu_k}(\alpha, \nu)^T p_b(\alpha, b^*, \nu) + p_{\nu_k}(\alpha, b^*, \nu) \\ &= p_{\nu_k}(\alpha, b^*, \nu) \end{aligned} \tag{15}$$

again by (13).

1.8 Second Derivatives

We will proceed in the opposite direction from the preceding section, calculating abstract derivatives before particular formulas for random effects aster models, because we need to see what work needs to be done before doing it (we may not need all second derivatives).

By the multivariate chain rule (Browder, 1996, Theorem 8.15)

$$\begin{aligned}
q_{\alpha\alpha}(\alpha, \nu) &= p_{\alpha\alpha}(\alpha, b^*, \nu) + p_{\alpha b}(\alpha, b^*, \nu) b^*_\alpha(\alpha, \nu) \\
q_{\alpha\nu}(\alpha, \nu) &= p_{\alpha\nu}(\alpha, b^*, \nu) + p_{\alpha b}(\alpha, b^*, \nu) b^*_\nu(\alpha, \nu) \\
q_{\nu\nu}(\alpha, \nu) &= p_{\nu\nu}(\alpha, b^*, \nu) + p_{\nu b}(\alpha, b^*, \nu) b^*_\nu(\alpha, \nu)
\end{aligned}$$

The estimating equation (13) defines b^* implicitly. Thus derivatives of b^* are computed using the implicit function theorem (Browder, 1996, Theorem 8.29)

$$b^*_\alpha(\alpha, \nu) = -p_{bb}(\alpha, b^*, \nu)^{-1} p_{b\alpha}(\alpha, b^*, \nu) \tag{16}$$

$$b^*_\nu(\alpha, \nu) = -p_{bb}(\alpha, b^*, \nu)^{-1} p_{b\nu}(\alpha, b^*, \nu) \tag{17}$$

This theorem requires that $p_{bb}(\alpha, b^*, \nu)$ be invertible, and we shall see below that it is. Then the second derivatives above can be rewritten

$$\begin{aligned}
q_{\alpha\alpha}(\alpha, \nu) &= p_{\alpha\alpha}(\alpha, b^*, \nu) - p_{\alpha b}(\alpha, b^*, \nu) p_{bb}(\alpha, b^*, \nu)^{-1} p_{b\alpha}(\alpha, b^*, \nu) \\
q_{\alpha\nu}(\alpha, \nu) &= p_{\alpha\nu}(\alpha, b^*, \nu) - p_{\alpha b}(\alpha, b^*, \nu) p_{bb}(\alpha, b^*, \nu)^{-1} p_{b\nu}(\alpha, b^*, \nu) \\
q_{\nu\nu}(\alpha, \nu) &= p_{\nu\nu}(\alpha, b^*, \nu) - p_{\nu b}(\alpha, b^*, \nu) p_{bb}(\alpha, b^*, \nu)^{-1} p_{b\nu}(\alpha, b^*, \nu)
\end{aligned}$$

a particularly simple and symmetric form. If we combine all the parameters in one vector $\psi = (\alpha, \nu)$ and write $p(\psi, b)$ instead of $p(\alpha, b, \nu)$ we have

$$q_{\psi\psi}(\psi) = p_{\psi\psi}(\psi, b^*) - p_{\psi b}\big(\psi, b^*\big) p_{bb}\big(\psi, b^*\big)^{-1} p_{b\psi}\big(\psi, b^*\big) \tag{18}$$

This form is familiar from the conditional variance formula for normal distributions if

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{19}$$

is the partitioned variance matrix of a partitioned normal random vector with components X_1 and X_2, then the variance matrix of the conditional distribution of X_1 given X_2 is

$$\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \tag{20}$$

assuming that X_2 is nondegenerate (Anderson, 2003, Theorem 2.5.1). Moreover, if the conditional distribution is degenerate, that is, if there exists a nonrandom vector v such that $\operatorname{var}(v^T X_1 \mid X_2) = 0$, then

$$v^T X_1 = v^T \Sigma_{12} \Sigma_{22}^{-1} X_2$$

with probability one, assuming X_1 and X_2 have mean zero (also by Anderson, 2003, Theorem 2.5.1), and the joint distribution of X_1 and X_2 is also degenerate. Thus we conclude that if the (joint) Hessian matrix of p is nonsingular, then so is the (joint) Hessian matrix of q given by (18).

The remaining work for this section is deriving the second derivatives of p that we need (it has turned out that we need all of them)

$$\begin{aligned}
p_{\alpha\alpha}(\alpha, b, \nu) &= M^T W (a + M\alpha + Zb) M \\
p_{\alpha b}(\alpha, b, \nu) &= M^T W (a + M\alpha + Zb) Z \\
p_{bb}(\alpha, b, \nu) &= Z^T W (a + M\alpha + Zb) Z + D^{-1} \\
p_{\alpha\nu_k}(\alpha, b, \nu) &= 0 \\
p_{b\nu_k}(\alpha, b, \nu) &= -D^{-1} E_k D^{-1} b \\
p_{\nu_j\nu_k}(\alpha, b, \nu) &= b^T D^{-1} E_j D^{-1} E_k D^{-1} b \\
&\qquad - \tfrac{1}{2} \operatorname{tr}\Big(\left[Z^T \widehat{W} Z D + I \right]^{-1} Z^T \widehat{W} Z E_j \\
&\qquad\qquad \left[Z^T \widehat{W} Z D + I \right]^{-1} Z^T \widehat{W} Z E_k \Big)
\end{aligned}$$

This finishes the derivation of all the derivatives we need. Recall that in our use of the implicit function theorem we needed $p_{bb}(\alpha, b^*, \nu)$ to be invertible. From the explicit form given above we see that it is actually negative definite, because $W(a + M\alpha + Zb)$ is positive semidefinite by (2).

1.9 Zero Variance Components

When some variance components are zero, the corresponding diagonal components of D are zero, and the corresponding components of b are zero almost surely. The order of the components of b does not matter, so long as the rows of Z and the rows and columns of D are reordered in the same way. So suppose these objects are partitioned as

$$b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} \qquad Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \qquad D = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

where $D_2 = 0$ and the diagonal components of D_1 are all strictly positive, so the components of b_2 are all zero almost surely and the components of

b_1 are all nonzero almost surely. Since $Zb = Z_1 b_1$ almost surely, the value of Z_2 is irrelevant. In the expression for D we are using the convention that 0 denotes the zero matrix of the dimension needed for the expression it appears in to make sense, so the two appearances of 0 in the expression for D as a partitioned matrix denote different submatrices having all components zero (they are transposes of each other).

Then the correct expression for the complete data log likelihood is

$$l_c(\alpha, b, \nu) = l(a + M\alpha + Z_1 b_1) - \tfrac{1}{2} b_1^T D_1^{-1} b_1 - \tfrac{1}{2} \log \det(D_1) \tag{21}$$

that is, the same as (4) except with subscripts 1 on b, Z, and D. And this leads to the correct expression for the approximate log likelihood

$$\begin{aligned} q(\alpha, \nu) = &-l(a + M\alpha + Z_1 b_1^*) + \tfrac{1}{2} (b_1^*)^T D_1^{-1} b_1^* \\ &+ \tfrac{1}{2} \log \det\bigl[Z_1^T \widehat{W} Z_1 D_1 + I\bigr] \end{aligned} \tag{22}$$

where again I denotes the identity matrix of the appropriate dimension (which now must be the dimension of D_1 for the expression it appears in to make sense) and where b_1^* denotes the maximizer of (21) considered as a function of b_1 with α and ν fixed, so it is actually a function of α and ν although the notation does not indicate this. Since

$$\begin{aligned} Z^T \widehat{W} Z D + I &= \begin{pmatrix} Z_1^T \widehat{W} Z_1 D_1 + I & Z_1^T \widehat{W} Z_2 D_2 \\ Z_2^T \widehat{W} Z_1 D_1 & Z_2^T \widehat{W} Z_2 D_2 + I \end{pmatrix} \\ &= \begin{pmatrix} Z_1^T \widehat{W} Z_1 D_1 + I & 0 \\ Z_2^T \widehat{W} Z_1 D_1 & I \end{pmatrix} \end{aligned}$$

where again we are using the convention that I denotes the identity matrix of the appropriate dimension and 0 denotes the zero matrix of the appropriate dimension, so I denotes different identity matrices in different parts of this equation, having the dimension of D on the left-hand side, the dimension of D_1 in the first column of both partitioned matrices, and the dimension of D_2 in the second column of both partitioned matrices,

$$\begin{aligned} \det(Z^T \widehat{W} Z D + I) &= \det(Z_1^T \widehat{W} Z_1 D_1 + I) \det(I) \\ &= \det(Z_1^T \widehat{W} Z_1 D_1 + I) \end{aligned}$$

by the rule that the determinant of a blockwise lower triangular partitioned matrix is the product of the determinants of the blocks on the diagonal (Harville, 1997, Theorem 13.3.1). And since $Z_1 b_1 = Zb$ almost surely,

$$\begin{aligned} q(\alpha, \nu) = &-l(a + M\alpha + Zb^*) + \tfrac{1}{2} (b_1^*)^T D_1^{-1} b_1^* \\ &+ \tfrac{1}{2} \log \det\bigl[Z^T \widehat{W} Z D + I\bigr] \end{aligned} \tag{23}$$

that is, the subscripts 1 are only needed in the term where the matrix inverse appears and are necessary there because D^{-1} does not exist. Breslow and Clayton (1993, Section 2.3) suggest using the Moore-Penrose pseudoinverse (Harville, 1997, Chapter 20)

$$D^+ = \begin{pmatrix} D_1^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

which gives

$$q(\alpha, \nu) = -l(a + M\alpha + Zb^*) + \tfrac{1}{2}(b^*)^T D^+ b^* + \tfrac{1}{2} \log \det\left[Z^T \widehat{W} Z D + I\right] \quad (24)$$

for the approximate log likelihood. This hides but does not eliminate the partitioning. Although there is no explicit partitioning in (24), it is still there in the definition of b^*.

Although this proposal (Breslow and Clayton, 1993, Section 2.3) does deal with the situation where the zero variance components are somehow known, it does not adequately deal with estimating which variance components are zero. That is the subject of the following two sections.

1.10 The Theory of Constrained Optimization

1.10.1 Incorporating Constraints in the Objective Function

When zero variance components arise, optimization of (8) puts us in the realm of constrained optimization. The theory of constrained optimization (Rockafellar and Wets, 2004) has a notational bias towards minimization (Rockafellar and Wets, 2004, p. 5). One can, of course, straightforwardly translate every result in Rockafellar and Wets (2004) from the context of minimization to the context of maximization, because for any objective function f, maximizing f is the same as minimizing $-f$, and Rockafellar and Wets give infrequent hints and discussions of alternative terminology in aid of this. But since the theory of constrained optimization is strange to most statisticians, especially the abstract theory that is needed here (Karush-Kuhn-Tucker theory is not helpful here, as we shall see), it is much simpler to switch from maximization to minimization so we can use all of the theory in Rockafellar and Wets (2004) without modification. And we have done so.

The theory of constrained optimization incorporates constraints in the objective function by the simple device of defining the objective function (for a minimization problem) to have the value $+\infty$ off the constraint set (Rockafellar and Wets, 2004, Section 1A). Since no point where the objective function has the value $+\infty$ can minimize it, unless the the objective function

has the value $+\infty$ everywhere, which is not the case in any application, the unconstrained minimizer of this sort of objective function is the same as the constrained minimizer.

Thus we need to impose constraints on our key function (8), requiring that each component of ν be nonnegative and when any component of ν is zero the corresponding components of b are also zero. However, the formula (8) does not make sense when components of ν are zero, so we will have to proceed differently.

1.10.2 Lower Semicontinuous Regularization

Since all but the middle term on the right-hand side of (8) are actually defined on some neighborhood of each point of the constraint set and differentiable at each point of the constraint set, we only need to deal with the middle term. It is the sum of terms of the form b_i^2/ν_k, where ν_k is the variance of b_i. Thus we investigate functions of this form

$$h(b, \nu) = b^2/\nu \tag{25}$$

where, temporarily, b and ν are scalars rather than vectors (representing single components of the vectors). In case $\nu > 0$ we have derivatives

$$\begin{aligned} h_b(b, \nu) &= 2b/\nu \\ h_\nu(b, \nu) &= -b^2/\nu^2 \\ h_{bb}(b, \nu) &= 2/\nu \\ h_{b\nu}(b, \nu) &= -2b/\nu^2 \\ h_{\nu\nu}(b, \nu) &= 2b^2/\nu^3 \end{aligned}$$

The Hessian matrix

$$h''(b, \nu) = \begin{pmatrix} 2/\nu & -2b/\nu^2 \\ -2b/\nu^2 & 2b^2/\nu^3 \end{pmatrix}$$

has nonnegative determinants of its principal submatrices, since the diagonal components are positive and $\det\bigl(h''(b, \nu)\bigr)$ is zero. Thus the Hessian matrix is nonnegative definite (Harville, 1997, Theorem 14.9.11), which implies that h itself is convex (Rockafellar and Wets, 2004, Theorem 2.14) on the set where $\nu > 0$.

We then extend h to the whole of the constraint set (this just adds the origin to the points already considered) in two steps. First we define it to

have the value $+\infty$ at all points not yet considered (those where any component of ν is nonpositive). This gives us an extended-real-valued convex function defined on all of $\mathbb{R}^2$. Second we take it to be the lower semicontinuous (LSC) regularization (Rockafellar and Wets, 2004, p. 14) of the function just defined. The LSC regularization of a convex function is convex (Rockafellar and Wets, 2004, Proposition 2.32). For any sequences $b_n \to b \neq 0$ and $\nu_n \searrow 0$ we have $h(b_n, \nu_n) \to \infty$. Thus the LSC regularization has the value $+\infty$ for $\nu = 0$ but $b \neq 0$. If $b_n = 0$ and $\nu_n \searrow 0$ we have $h(b_n, \nu_n) = 0$ for all n. Since $h(b, \nu) \geq 0$ for all b and $\nu \geq 0$, we conclude

$$\liminf_{\substack{b \to 0 \\ \nu \searrow 0}} h(b, \nu) = 0$$

Thus the LSC regularization has the value 0 for $b = \nu = 0$. In summary

$$h(b, \nu) = \begin{cases} b^2/\nu, & \nu > 0 \\ 0, & \nu = 0 \text{ and } b = 0 \\ +\infty, & \text{otherwise} \end{cases} \tag{26}$$

is an LSC convex function, which agrees with our original definition in case $\nu > 0$. Note that $h(b, 0)$ considered as a function of b is minimized at $b = 0$ because that is the only point where this function is finite.

Let k denote the map from indices for b to indices for ν that gives corresponding components: $\nu_{k(i)}$ is the variance of b_i. Let $\dim(b)$ denote the number of random effects. Then our objective function can be written

$$p(\alpha, b, \nu) = -l(a + M\alpha + Zb) + \tfrac{1}{2} \sum_{i=1}^{\dim(b)} h(b_i, \nu_{k(i)}) + \tfrac{1}{2} \log \det\left[Z^T \widehat{W} Z D + I\right] \tag{27}$$

where h is given by (26), provided all of the components of ν are nonnegative. The proviso is necessary because the third term on the right-hand side is not defined for all values of ν, only those such that the argument of the determinant is a positive definite matrix. Hence, we must separately define $p(\alpha, b, \nu) = +\infty$ whenever any component of ν is negative.

1.10.3 Subderivatives

In calculus we learn that the first derivative is zero at a local minimum and, therefore, to check points where the first derivative is zero. This is called Fermat's rule. This rule no longer works for nonsmooth functions, including

those that incorporate constraints, such as (27). It does, of course, still work at points in the interior of the constraint set where (27) is differentiable. It does not work to check points on the boundary. There we need what Rockafellar and Wets (2004, Theorem 10.1) call *Fermat's rule, generalized:* at a local minimum the *subderivative function* is nonnegative.

For any extended-real-valued function f on $\mathbb{R}^d$, the *subderivative function*, denoted $df(x)$ is also an extended-real-valued function on $\mathbb{R}^d$ defined by

$$df(x)(\bar{w}) = \liminf_{\substack{\tau \searrow 0 \\ w \to \bar{w}}} \frac{f(x + \tau w) - f(x)}{\tau}$$

(Rockafellar and Wets, 2004, Definition 8.1). The notation on the left-hand side is read the subderivative of f at the point x in the direction $\bar{w}$. Fortunately, we do not have to use this definition to calculate subderivatives we want, because the calculus of subderivatives allows us to use simpler formulas in special cases. Firstly, there is the notion of subdifferential regularity (Rockafellar and Wets, 2004, Definition 7.25), which we can use without knowing the definition. The sum of regular functions is regular and the subderivative of a sum is the sum of the subderivatives (Rockafellar and Wets, 2004, Corollary 10.9). A smooth function is regular and the subderivative is given by

$$df(x)(w) = w^T f'(x), \tag{28}$$

where, as in Sections 1.1 and 1.4 above, $f'(x)$ denotes the gradient vector (the vector of partial derivatives) of f at the point x (Rockafellar and Wets, 2004, Exercise 8.20). Every LSC convex function is regular (Rockafellar and Wets, 2004, Example 7.27). Thus in computing subderivatives of (27) we may compute them term by term, and for the first and last terms, they are given in terms of the partial derivatives already computed by (28). For an LSC convex function f, we have the following characterization of the subderivative (Rockafellar and Wets, 2004, Proposition 8.21). At any point x where $f(x)$ is finite, the limit

$$g(w) = \lim_{\tau \searrow 0} \frac{f(x + \tau w) - f(x)}{\tau}$$

exists and defines a sublinear function g, and then $df(x)$ is the LSC regularization of g. An extended-real-valued function g is *sublinear* if $g(0) = 0$ and

$$g(a_1 x_1 + a_2 x_2) \le a_1 g(x_1) + a_2 g(x_2)$$

for all vectors x_1 and x_2 and positive scalars a_1 and a_2 (Rockafellar and Wets, 2004, Definition 3.18). The subderivative function of every regular LSC function is sublinear (Rockafellar and Wets, 2004, Theorem 7.26).

So let us proceed to calculate the subderivative of (26). In the interior of the constraint set, where this function is smooth, we can use the partial derivatives already calculated

$$dh(b, \nu)(u, v) = \frac{2bu}{\nu} - \frac{b^2 v}{\nu^2}$$

where the notation on the left-hand side means the subderivative of h at the point (b, ν) in the direction (u, v). On the boundary of the constraint set, which consists of the single point $(0, 0)$, we take limits. In case $v > 0$, we have

$$\lim_{\tau \searrow 0} \frac{h(\tau u, \tau v) - h(0, 0)}{\tau} = \lim_{\tau \searrow 0} \frac{\tau^2 u^2 / (\tau v)}{\tau} = \lim_{\tau \searrow 0} \frac{u^2}{v} = \frac{u^2}{v}$$

In case $v \le 0$ and $u \ne 0$, we have

$$\lim_{\tau \searrow 0} \frac{h(\tau u, \tau v) - h(0, 0)}{\tau} = \lim_{\tau \searrow 0} (+\infty) = +\infty$$

In case $v = 0$ and $u = 0$, we have

$$\lim_{\tau \searrow 0} \frac{h(\tau u, \tau v) - h(0, 0)}{\tau} = 0$$

Thus if we define

$$g(u, v) = \begin{cases} u^2 / v, & v > 0 \\ 0, & u = v = 0 \\ +\infty, & \text{otherwise} \end{cases}$$

The theorem says $dh(0, 0)$ is the LSC regularization of g. But we recognize $g = h$, so g is already LSC, and we have

$$dh(0, 0)(u, v) = h(u, v)$$

1.10.4 Applying the Generalization of Fermat's Rule

The theory of constrained optimization tells us nothing we did not already know (from Fermat's rule) about smooth functions. The only way we can have $df(x)(w) = w^T f'(x) \ge 0$ for all vectors w is if $f'(x) = 0$. It is only at points where the function is nonsmooth, in the cases of interest

to us, points on the boundary of the constraint set, where the theory of constrained optimization tells us things we did not know and need to know.

Even on the boundary, the conclusions of the theory about components of the state that are not on the boundary agree with what we already knew. We have

$$dp(\alpha, b, \nu)(s, u, v) = s^T p_\alpha(\alpha, b, \nu) + \text{terms not containing } s$$

and the only way this can be nonnegative for all s is if

$$p_\alpha(\alpha, b, \nu) = 0 \tag{29}$$

in which case $dp(\alpha, b, \nu)(s, u, v)$ is a constant function of s, or, what is the same thing in other words, the terms of $dp(\alpha, b, \nu)(s, u, v)$ that appear to involve s are all zero (and so do not actually involve s).

Similarly, $dp(\alpha, b, \nu)(s, u, v) \geq 0$ for all u_i and v_j such that $\nu_j > 0$ and $k(i) = j$ only if

$$\begin{aligned} p_{\nu_j}(\alpha, b, \nu) &= 0, \qquad j \text{ such that } \nu_j > 0 \\ p_{b_i}(\alpha, b, \nu) &= 0, \qquad i \text{ such that } \nu_{k(i)} > 0 \end{aligned} \tag{30}$$

in which case we conclude that $dp(\alpha, b, \nu)(s, u, v)$ is a constant function of such u_i and v_j.

Thus, assuming that we are at a point (α, b, ν) where (29) and (30) hold, and we do assume this throughout the rest of this section, $dp(\alpha, b, \nu)(s, u, v)$ actually involves only v_j and u_i such that $\nu_j = 0$ and $k(i) = j$. Define

$$\bar{p}(\alpha, b, \nu) = -l(a + M\alpha + Zb) + \tfrac{1}{2} \log \det \left[Z^T \widehat{W} Z D + I \right] \tag{31}$$

(the part of (27) consisting of the smooth terms). Then

$$\begin{aligned} dp(\alpha, b, \nu)(s, u, v) = \sum_{j \in J} \Bigg[& v_j \bar{p}_{\nu_j}(\alpha, b, \nu) \\ & + \sum_{i \in k^{-1}(j)} \Big(u_i \bar{p}_{b_i}(\alpha, b, \nu) + h(u_i, v_j) \Big) \Bigg] \end{aligned} \tag{32}$$

where J is the set of j such that $\nu_j = 0$, where $k^{-1}(j)$ denotes the set of i such that $k(i) = j$, and where h is defined by (26). Fermat's rule generalized says we must consider all of the terms of (32) together. We cannot consider partial derivatives, because the partial derivatives do not exist. To check

that we are at a local minimum we need to show that (32) is nonnegative for all vectors u and v. Conversely, to verify that we are not at a local minimum, we need to find one pair of vectors u and v such that (32) is negative. Such a pair (u, v) we call a *descent direction*. Since Fermat's rule generalized is a necessary but not sufficient condition (like the ordinary Fermat's rule), the check that we are at a local minimum is not definitive, but the check that we are not is. If a descent direction is found, then moving in that direction away from the current value of (α, b, ν) will decrease the objective function (27).

So how do we find a descent direction? We want to minimize (32) considered as a function of u and v for fixed α, b, and ν. On further consideration, we can consider the terms of (32) for each j separately. If the minimum of

$$v_j \bar{p}_{\nu_j}(\alpha, b, \nu) + \sum_{i \in k^{-1}(j)} \left(u_i \bar{p}_{b_i}(\alpha, b, \nu) + h(u_i, v_j) \right) \tag{33}$$

over all vectors u and v is nonnegative, then the minimum is zero, because (33) has the value zero when $u = 0$ and $v = 0$. Thus we can ignore this j in calculating the descent direction.

On the other hand, if the minimum is negative, then the minimum does not occur at $v = 0$ and the minimum is actually $-\infty$ by the sublinearity of the subderivative, one consequence of sublinearity being positive homogeneity

$$df(x)(\tau w) = \tau df(x)(w), \qquad \tau \geq 0$$

which holds for for any subderivative. Thus (as our terminology hints) we are only trying to find a descent *direction*, the length of the vector (u, v) does not matter, only its direction. Thus to get a finite minimum we can do a constrained minimization of (33), constraining (u, v) to lie in a ball. This is found by the well-known Karush-Kuhn-Tucker theory of constrained optimization to be the minimum of the Lagrangian function

$$L(u, v) = \lambda v_j^2 + v_j \bar{p}_{\nu_j}(\alpha, b, \nu) + \sum_{i \in k^{-1}(j)} \left(\lambda u_i^2 + u_i \bar{p}_{b_i}(\alpha, b, \nu) + \frac{u_i^2}{v_j} \right) \tag{34}$$

where $\lambda > 0$ is the Lagrange multiplier, which would have to be adjusted if we were interested in constraining (u, v) to lie in a particular ball. Since we do not care about the length of (u, v) we can use any λ. We have replaced $h(u_i, v_i)$ by u_i^2 / v_j because we know that if we are finding an actual descent

direction, then we will have $v_j > 0$. Now

$$L_{u_i}(u, v) = 2\lambda u_i + \bar{p}_{b_i}(\alpha, b, \nu) + \frac{2u_i}{v_j}, \qquad i \in k^{-1}(j)$$

$$L_{v_j}(u, v) = 2\lambda v_j + \bar{p}_{\nu_j}(\alpha, b, \nu) - \sum_{i \in k^{-1}(j)} \frac{u_i^2}{v_j^2}$$

The minimum occurs where these are zero. Setting the first equal to zero and solving for u_i gives

$$\hat{u}_i(v_j) = -\frac{\bar{p}_{b_i}(\alpha, b, \nu)}{2(\lambda + 1/v_j)}$$

plugging this back into the second gives

$$L_{v_j}\big(\hat{u}(v), v\big) = 2\lambda v_j + \bar{p}_{\nu_j}(\alpha, b, \nu) - \frac{1}{4(\lambda v_j + 1)^2} \sum_{i \in k^{-1}(j)} \bar{p}_{b_i}(\alpha, b, \nu)^2$$

and we seek zeros of this. The right-hand is clearly an increasing function of v_j so it is negative somewhere only if it is negative when $v_j = 0$ where it has the value

$$\bar{p}_{\nu_j}(\alpha, b, \nu) - \frac{1}{4} \sum_{i \in k^{-1}(j)} \bar{p}_{b_i}(\alpha, b, \nu)^2 \tag{35}$$

So that gives us a test for a descent direction: we have a descent direction if and only if (35) is negative. Conversely, we appear to have $\hat{\nu}_j = 0$ if (35) is nonnegative.

That finishes our treatment of the theory of constrained optimization. We have to ask is all of this complication really necessary? It turns out that it is and it isn't. We can partially avoid it by a change of variables. But the cure is worse than the disease in some ways. This is presented in the following section.

1.11 Square Roots

We can avoid constrained optimization by the following change of parameter. Introduce new parameter variables by

$$\nu_j = \sigma_j^2$$
$$b = Ac$$

where A is diagonal and $A^2 = D$, so the i-th diagonal component of A is $\sigma_{k(i)}$. Then the objective function (8) becomes

$$\tilde{p}(\alpha, c, \sigma) = -l(a + M\alpha + ZAc) + \tfrac{1}{2} c^T c + \tfrac{1}{2} \log \det\left[Z^T \widehat{W} Z A^2 + I \right] \quad (36)$$

There are now no constraints and (36) is a continuous function of all variables.

The drawback is that by symmetry we must have $\tilde{p}_{\sigma_j}(\alpha, c, \sigma)$ equal to zero when $\sigma_j = 0$. Thus first derivatives become useless for checking for descent directions, and second derivative information is necessary. However, that is not the way unconstrained optimizers like the R functions `optim` and `nlminb` work. They do not expect such pathological behavior and do not deal with it correctly. If we want to use such optimizers to find local minima of (36), then we must provide starting points that have no component of ν equal to zero, and hope that the optimizer will never get any component of ν close to zero unless zero actually is a solution. But this is only a hope. The theory that guided the design of these optimizers does not provide any guarantees for this kind of objective function.

Moreover, optimizer algorithms stop when close to but not exactly at a solution, a consequence of inexactness of computer arithmetic. Thus when the optimizer stops and declares convergence with one or more components of ν close to zero, how do we know whether the true solution is exactly zero or not? We don't unless we return to the original parameterization and apply the theory of the preceding section. The question of whether the MLE of the variance components are exactly zero or not is of scientific interest, so it seems that the device of this section does not entirely avoid the theory of constrained optimization. We must change back to the original parameters and use (35) to determine whether or not we have $\nu_j = 0$.

Finally, there is another issue with this "square root" parameterization. The analogs of the second derivative formulas derived in Section 1.8 above, for this new parameterization are extraordinarily ill-behaved. The Hessian matrices are badly conditioned and sometimes turn out to be not positive definite when calculated by the computer's arithmetic (which is inexact) even though theory says they must be positive definite. We know this because at one point we thought that this "square root" parameterization was the answer to everything and tried to use it everywhere. Months of frustration ensued where it mostly worked, but failed on a few problems. It took us a long time to see that it is fundamentally wrong-headed. As we said above, the cure is worse than the disease.

Thus we concluded that, while we may use this "square root" parameterization to do unconstrained rather than constrained minimization, we

should only use it only for that. The test (35) should be used to determine whether variance components are exactly zero or not, and the formulas in Section 1.8 should be used to derive Fisher information.

1.11.1 First Derivatives

Some of R's optimization routines can use first derivative information, thus we derive first derivatives in this parameterization.

$$\tilde{p}_\alpha(\alpha, c, \sigma) = -M^T[y - \mu(a + M\alpha + ZAc)] \tag{37}$$

$$\tilde{p}_c(\alpha, c, \sigma) = -AZ^T[y - \mu(a + M\alpha + ZAc)] + c \tag{38}$$

$$\begin{aligned} \tilde{p}_{\sigma_j}(\alpha, c, \sigma) = &-c^T E_j Z^T[y - \mu(a + M\alpha + ZAc)] \\ &+ \operatorname{tr}\left([Z^T\widehat{W}ZA^2 + I]^{-1} Z^T\widehat{W}ZAE_j\right) \end{aligned} \tag{39}$$

where E_j is given by (12).

1.12 Fisher Information

The observed Fisher information matrix is minus the second derivative matrix of the log likelihood. As we said above, we want to do this in the original parameterization.

Assembling stuff derived in preceding sections and introducing

$$\begin{aligned} \mu^* &= \mu\big(a + M\alpha + Zb^*(\alpha, \nu)\big) \\ W^* &= W\big(a + M\alpha + Zb^*(\alpha, \nu)\big) \\ H^* &= Z^T W^* Z + D^{-1} \\ \widehat{H} &= Z^T\widehat{W}ZD + I \end{aligned}$$

we obtain

$$\begin{aligned} q_{\alpha\alpha}(\alpha, \nu) &= M^T W^* M - M^T W^* Z (H^*)^{-1} Z^T W^* M \\ q_{\alpha\nu_j}(\alpha, \nu) &= M^T W^* Z (H^*)^{-1} D^{-1} E_j D^{-1} b^* \\ q_{\nu_j\nu_k}(\alpha, \nu) &= (b^*)^T D^{-1} E_j D^{-1} E_k D^{-1} b^* \\ &\quad - \tfrac{1}{2} \operatorname{tr}\left(\widehat{H}^{-1} Z^T\widehat{W}ZE_j\widehat{H}^{-1} Z^T\widehat{W}ZE_k\right) \\ &\quad - (b^*)^T D^{-1} E_j D^{-1} (H^*)^{-1} D^{-1} E_k D^{-1} b^* \end{aligned}$$

In all of these b^*, μ^*, W^*, and H^* are functions of α and ν even though the notation does not indicate this.

It is tempting to think expected Fisher information simplifies things because we "know" $E(y) = \mu$ and $\operatorname{var}(y) = W$, except we don't know that! What we do know is

$$E(y \mid b) = \mu(a + M\alpha + Zb)$$

but we don't know how to take the expectation of the right hand side (and similarly for the variance). Rather than introduce further approximations of dubious validity, it seems best to just use (approximate) observed Fisher information.

1.13 Standard Errors for Random Effects

Suppose that the approximate Fisher information derived in Section 1.12 can be used to give an approximate asymptotic variance for the parameter vector $\psi = (\alpha, \nu)$. This estimate of the asymptotic variance is $q_{\psi\psi}(\hat{\psi})^{-1}$, where $q_{\psi\psi}(\psi)$ is given by (18) and $\hat{\psi} = (\hat{\alpha}, \hat{\nu})$.

To apply the delta method to get asymptotic standard errors for $\hat{b}$ we need the derivatives (16) and (17). Stacking these we obtain

$$b^*_{\psi}(\hat{\psi}) = \begin{pmatrix} -p_{bb}(\hat{\alpha}, \hat{b}, \hat{\nu})^{-1} p_{b\alpha}(\hat{\alpha}, \hat{b}, \hat{\nu}) \\ -p_{bb}(\hat{\alpha}, \hat{b}, \hat{\nu})^{-1} p_{b\nu}(\hat{\alpha}, \hat{b}, \hat{\nu}) \end{pmatrix}$$

and the delta method gives

$$b^*_{\psi}(\hat{\psi})^T q_{\psi,\psi}(\hat{\psi})^{-1} b^*_{\psi}(\hat{\psi}) \tag{40}$$

for the asymptotic variance of the estimator $\hat{b}$.

It must be conceded that in this section we are living what true believers in random effects models would consider a state of sin. The random effects vector b is not a parameter, yet $b^*(\hat{\psi})$ treats it as a function of parameters (which is thus a parameter) and the "asymptotic variance" (40) is derived by considering $\hat{b}$ just such a parameter estimate. So (40) is correct in what it does, so long as we buy the assumption that $q_{\psi\psi}(\hat{\psi})$ is approximate Fisher information for ψ, but it fails to treat random effects as actually random. Since any attempt to actually treat random effects as random would lead us to integrals that we cannot do, we leave the subject at this point. The asymptotic variance (40) may be philosophically incorrect in some circles, but it seems to be the best we can do.

1.14 REML?

Breslow and Clayton (1993) do not maximize the approximate log likelihood (6), but make further approximations to give estimators motivated by REML (restricted maximum likelihood) estimators for linear mixed models (LMM). Breslow and Clayton (1993) concede that the argument that justifies REML estimators for LMM does not carry over to their REML-like estimators for generalized linear mixed models (GLMM). Hence these REML-like estimators have no mathematical justification. Even in LMM the widely used procedure of following REML estimates of the variance components with so-called BLUE estimates of fixed effects and BLUP estimates of random effects, which are actually only BLUE and BLUP if the variance components are assumed known rather than estimated, is obviously wrong: ignoring the fact that the variance components are estimated cannot be justified (and Breslow and Clayton say this in their discussion section). Hence REML is not justified even in LMM when fixed effects are the parameters of interest. In aster models, because components of the response vector are dependent and have distributions in different families, it is very unclear what REML-like estimators in the style of Breslow and Clayton (1993) might be. The analogy just breaks down. Hence, we do not pursue this REML analogy and stick with what we have described above.

2 Practice

Our goal is to minimize (6). We replace (6) with (7) in some steps because of our inability to differentiate (6), but our whole procedure must minimize (6).

2.1 Step 1

To get close to $(\hat{\alpha}, \hat{c}, \hat{\sigma})$ starting from far away we minimize

$$\begin{aligned} r(\sigma) = {} & -l(a + M\tilde{\alpha} + ZA\tilde{c}) + \tfrac{1}{2}\tilde{c}^T\tilde{c} \\ & + \tfrac{1}{2}\log\det\bigl[Z^T W(a + M\tilde{\alpha} + ZA\tilde{c})ZA^2 + I\bigr] \end{aligned} \tag{41}$$

where $\tilde{\alpha}$ and $\tilde{c}$ are the joint minimizers of (36) considered as a function of α and c for fixed σ. In (41), $\tilde{\alpha}$, $\tilde{c}$, and A are all functions of σ although the notation does not indicate this.

Because we cannot calculate derivatives of (41) we minimize using by the R function `optim` with `method = "Nelder-Mead"`, the so-called Nelder-

Mead simplex algorithm, a no-derivative method nonlinear optimization, not to be confused with the simplex algorithm for linear programming.

2.2 Step 2

Having found α, c, and σ close to the MLE values via the preceding step, we then switch to minimization of (36) for which we have the derivative formulas (37), (38), and (39). In this step we can use one of R's optimization functions that uses first derivative information: `nlm` or `nlminb` or `optim` with optional argument `method = "BFGS"` or `method = "CG"` or `method = "L-BFGS-B"`.

To define (36) we also need a $\widehat{W}$, and we take the value at the current values of α, c, and σ. Because W is typically a very large matrix ($n \times n$, where n is the number of nodes in complete aster graph, the number of nodes in the subgraph for a single individual times the number of individuals), we actually store $Z^T \widehat{W} Z$, which is only $r \times r$, where r is the number of random effects. We set

$$Z^T \widehat{W} Z = Z^T W (a + M\alpha + ZAc) Z \tag{42}$$

where α, c, and $A = A(\sigma)$ are the current values before we start minimizing $\tilde{p}(\alpha, c, \sigma)$ and this value of $Z^T \widehat{W} Z$ is fixed throughout the minimization, as is required by the definition of $\tilde{p}(\alpha, c, \sigma)$.

Having minimized $\tilde{p}(\alpha, c, \sigma)$ we are still not done, because now (42) is wrong. We held it fixed at the values of α, c, and σ we had before the minimization, and now those values have changed. Thus we should re-evaluate (42) and re-minimize, and continue doing this until convergence.

We terminate this iteration when σ values do not change (to within some prespecified tolerance) because the α and c values are, in theory, determined by σ, because $\tilde{p}$ considered as a function of α and c for fixed σ is convex and hence has at most one local minimizer, so we do not need to worry about them converging.

When this iteration terminates we are done with this step, and we have our point estimates $\hat{\alpha}$, $\hat{c}$, and $\hat{\sigma}$. We also have our point estimates $\hat{b}$ of the random effects on the original scale given by $A(\hat{\nu})\hat{c}$ and our point estimates $\nu_j = \sigma_j^2$ of the variance components.

2.3 Step 3

Having converted back to the original parameters, if any of the ν_j are close to zero we use the check (35) to determine whether or not they are exactly zero.

2.4 To Do

A few issues that have not been settled. Points 1 and 2 in the following list are not specific to random effects models. They arise in fixed effect aster models too, even in generalized linear models and log-linear models in categorical data analysis.

1. Verify no directions of recession of fixed-effects-only model.

2. Verify supposedly nested models are actually nested.

3. How about constrained optimization and hypothesis tests of variance components being zero? How does the software automagically or educationally do the right thing? That is, do we just do the Right Thing or somehow explain to lusers what the Right Thing is?

A Cholesky

How do we calculate log determinants and derivatives thereof? R has a function `determinant` that calculates the log determinant. It uses LU decomposition.

An alternative method is to use Cholesky decomposition, but that only works when the given matrix is symmetric. This may be better because there is a sparse version (the `chol` function in the `Matrix` package) that may enable us to do much larger problems (perhaps after some other issues getting in the way of scaling are also fixed).

We need to calculate the log determinant that appears in (8) or (36), but the matrix is not symmetric. It can, however, be rewritten so as to be symmetric. Assuming A is invertible

$$\begin{aligned} \det\bigl(Z^T \widehat{W} Z A^2 + I\bigr) &= \det\bigl(Z^T \widehat{W} Z A + A^{-1}\bigr) \det\bigl(A\bigr) \\ &= \det\bigl(A Z^T \widehat{W} Z A + I\bigr) \end{aligned}$$

If A is singular, we can see by continuity that the two sides must agree there too. That takes care of (36). The same trick works for (8); just replace A by $D^{1/2}$, which is the diagonal matrix whose diagonal components are the nonnegative square roots of the corresponding diagonal components of D.

Cholesky can also be used to efficiently calculate matrix inverses (done by the `chol2inv` function in the `Matrix` package). So we investigate whether we can use Cholesky to calculate derivatives.

A.1 First Derivatives

For the trace in the formula (39) for $\tilde{p}_{\sigma_j}(\alpha, c, \sigma)$ we have in case A is invertible

$$\begin{aligned}
&\operatorname{tr}\left(\left[Z^T\widehat{W}ZA^2 + I\right]^{-1} Z^T\widehat{W}ZAE_j\right) \\
&\qquad = \operatorname{tr}\left(\left[A^{-1}(AZ^T\widehat{W}ZA + I)A\right]^{-1} Z^T\widehat{W}ZAE_j\right) \\
&\qquad = \operatorname{tr}\left(A^{-1}\left[AZ^T\widehat{W}ZA + I\right]^{-1} AZ^T\widehat{W}ZAE_j\right) \\
&\qquad = \operatorname{tr}\left(\left[AZ^T\widehat{W}ZA + I\right]^{-1} AZ^T\widehat{W}ZAE_jA^{-1}\right) \\
&\qquad\qquad = \operatorname{tr}\left(\left[AZ^T\widehat{W}ZA + I\right]^{-1} AZ^T\widehat{W}ZE_j\right)
\end{aligned}$$

the next-to-last equality being $\operatorname{tr}(AB) = \operatorname{tr}(BA)$ and the last equality using the fact that A, E_j, and A^{-1} are all diagonal so they commute. Again we see that we get the same identity of the first and last expressions even when A is singular by continuity.

For the trace in the formula (11) for $p_{\nu_k}(\alpha, b, \nu)$ we have in case D is invertible

$$\begin{aligned}
&\operatorname{tr}\Big(\left[Z^T\widehat{W}ZD + I\right]^{-1} Z^T\widehat{W}ZE_k\Big) \\
&\qquad = \operatorname{tr}\Big(D^{-1/2}\left[D^{1/2}Z^T\widehat{W}ZD^{1/2} + I\right]^{-1} D^{1/2}Z^T\widehat{W}ZE_k\Big) \\
&\qquad\qquad = \operatorname{tr}\Big(\left[D^{1/2}Z^T\widehat{W}ZD^{1/2} + I\right]^{-1} D^{1/2}Z^T\widehat{W}ZD^{-1/2}E_k\Big)
\end{aligned}$$

This, of course, does not work when D is singular. We already knew we cannot differentiate $p(\alpha, b, \nu)$ on the boundary of the constraint set.

A.2 Second Derivatives

For the trace in the formula in Section 1.8 for $p_{\nu_j\nu_k}(\alpha, b, \nu)$ we have in case D is invertible

$$\begin{aligned}
&\operatorname{tr}\Big(\left[Z^T\widehat{W}ZD + I\right]^{-1} Z^T\widehat{W}ZE_j\left[Z^T\widehat{W}ZD + I\right]^{-1} Z^T\widehat{W}ZE_k\Big) \\
&\qquad = \operatorname{tr}\Big(D^{-1/2}\left[D^{1/2}Z^T\widehat{W}ZD^{1/2} + I\right]^{-1} D^{1/2}Z^T\widehat{W}ZE_j \\
&\qquad\quad D^{-1/2}\left[D^{1/2}Z^T\widehat{W}ZD^{1/2} + I\right]^{-1} D^{1/2}Z^T\widehat{W}ZE_k\Big) \\
&\qquad = \operatorname{tr}\Big(\left[D^{1/2}Z^T\widehat{W}ZD^{1/2} + I\right]^{-1} D^{1/2}Z^T\widehat{W}ZE_jD^{-1/2} \\
&\qquad\qquad\qquad \left[D^{1/2}Z^T\widehat{W}ZD^{1/2} + I\right]^{-1} D^{1/2}Z^T\widehat{W}ZE_kD^{-1/2}\Big)
\end{aligned}$$

Again, this does not work when D is singular.

The same trace occurs in the expression for $q_{\nu_j \nu_k}(\alpha, \nu)$ given in Section 1.12 and can be calculated the same way.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Hoboken: John Wiley & Sons.

Barndorff-Nielsen, O. (1978). *Information and Exponential Families*. Chichester: John Wiley & Sons.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Browder, A. (1996). *Mathematical Analysis: An Introduction*. New York: Springer-Verlag.

Geyer, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society, Series B*, **56**, 261–274.

Geyer, C. J. (2015). R package `aster` (Aster Models), version 0.8-31. `http://www.stat.umn.edu/geyer/aster/` and `https://cran.r-project.org/package=aster`

Geyer, C. J. (2013). Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, G. L. Jones and X. Shen eds. IMS Collections, Vol. 10, pp. 1–24. Institute of Mathematical Statistics: Hayward, CA.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data, (with discussion). *Journal of the Royal Statistical Society, Series B*, **54**, 657–699.

Geyer, C. J., Wagenius, S. and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, **94**, 415–426.

Harville, D. A. (1997). Matrix Algebra From a Statistician's Perspective. New York: Springer.

Rockafellar, R. T. and Wets, R. J.-B. (2004). *Variational Analysis*, corr. 2nd printing. Berlin: Springer-Verlag.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley.

Shaw, F. H., Promislow, D. E. L., Tatar, M., Hughes, K. A. and Geyer, C. J. (1999). Towards reconciling inferences concerning genetic variation in senescence. *Genetics*, **152**, 553–566.

Shaw, F. H., Geyer, C. J. and Shaw, R. G. (2002). A Comprehensive Model of Mutations Affecting Fitness and Inferences for *Arabidopsis thaliana* *Evolution*, **56**, 453–463.

Shaw, R. G., Geyer, C. J., Wagenius, S., Hangelbroek, H. H., and Etterson, J. R. (2008). Unifying life history analysis for inference of fitness and population growth. *American Naturalist* **172**, E35–E47.

Sung, Y. J. and Geyer, C. J. (2007). Monte Carlo likelihood inference for missing data models. *Annals of Statistics*, **35**, 990–1011.

Thompson, E. A. and Guo, S. W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA J. Math. Appl. Med. Biol.*, **8**, 149–169.