Design Document for Random Effects Aster Models

Charles J. Geyer

May 18, 2012

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 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 θ, where D is a function of θ, although this is not indicated by the notation. Temporarily, we leave the choice of exactly what function D is of θ unspecified. That choice will be made in Section 1.6 below.

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

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

in case none of the variance components are zero. We deal with the case of zero variance components in Section 1.4 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

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

Maximum likelihood estimates (MLE) of α and θ 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 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 θ. Then $\log L_m(\alpha, \theta)$ is approximated by

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

where

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

Hence

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

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 θ and D is a function of θ, although this is not indicated by the notation, and where the last equality uses the rule sum of logs is log of product and and the rule product of determinants is determinant of matrix product (Harville, 1997, Theorem 13.3.4) and $D^{1/2}$ denotes the symmetric square root of D, which in this case is the diagonal matrix whose diagonal components are the square roots of the corresponding diagonal components of D. Our equation (6) is our analog of equation (5) in Breslow and Clayton (1993).

The key idea is to use (6) as if it were the log likelihood for the unknown parameters (α and θ), 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, 2012) provide such — the derivatives do, of course, exist because every cumulant function of a full 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, in press). 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

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

where $\widehat{W}$ is a constant matrix (not a function of α and θ). 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 $\hat{\theta}$ are the joint maximizers of (6) and $\hat{b} = b^*(\hat{\alpha}, \hat{\theta})$. Note that (7) is a redefinition of $q(\alpha, \theta)$. Hereafter we will no longer use the definition (6).

1.4 Zero Variance Components

When some variance components are zero, the corresponding diagonal components of D are zero, and the corresponding elements of b are zero almost surely. The order of the elements 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

$$l_c(\alpha, b, \theta) = l(a + M\alpha + Z_1 b_1) - \tfrac{1}{2} b_1^T D_1^{-1} b_1 - \tfrac{1}{2} \log \det(D_1) \tag{8}$$

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

$$\begin{aligned} q(\alpha, \theta) = l(a + M\alpha + Z_1 b_1^*) - \tfrac{1}{2} (b_1^*)^T D_1^{-1} b_1^* \\ - \tfrac{1}{2} \log \det \bigl[D_1^{1/2} Z_1^T \widehat{W} Z_1 D_1^{1/2} + I \bigr] \end{aligned} \tag{9}$$

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 (8) considered as a function of b_1 with α and θ fixed, so it is actually a function of α and θ although the notation does not indicate this. Since

$$\begin{aligned} D^{1/2} Z^T \widehat{W} Z D^{1/2} + I &= \begin{pmatrix} D_1^{1/2} Z_1^T \widehat{W} Z_1 D_1^{1/2} + I & D_1^{1/2} Z_1^T \widehat{W} Z_2 D_2^{1/2} \\ D_2^{1/2} Z_2^T \widehat{W} Z_1 D_1^{1/2} & D_2^{1/2} Z_2^T \widehat{W} Z_2 D_2^{1/2} + I \end{pmatrix} \\ &= \begin{pmatrix} D_1^{1/2} Z_1^T \widehat{W} Z_1 D_1^{1/2} + I & 0 \\ 0 & I \end{pmatrix} \end{aligned}$$

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, and 0 also denotes different zero matrices, the two appearances being transposes of each other,

$$\begin{aligned} \det(D^{1/2} Z^T \widehat{W} Z D^{1/2} + I) &= \det(D_1^{1/2} Z_1^T \widehat{W} Z_1 D_1^{1/2} + I) \det(I) \\ &= \det(D_1^{1/2} Z_1^T \widehat{W} Z_1 D_1^{1/2} + I) \end{aligned}$$

by the rule that the determinant of a block-diagonal 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,

$$\begin{aligned} q(\alpha, \theta) = l(a + M\alpha + Zb^*) &- \tfrac{1}{2} (b_1^*)^T D_1^{-1} b_1^* \\ &- \tfrac{1}{2} \log \det \bigl[D^{1/2} Z^T \widehat{W} Z D^{1/2} + I \bigr] \end{aligned} \tag{10}$$

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

$$\begin{aligned} q(\alpha, \theta) = l(a + M\alpha + Zb^*) &- \tfrac{1}{2} (b^*)^T D^{+} b^* \\ &- \tfrac{1}{2} \log \det \bigl[D^{1/2} Z^T \widehat{W} Z D^{1/2} + I \bigr] \end{aligned} \tag{11}$$

for the approximate log likelihood. Partitioned matrices are no longer needed when the Moore-Penrose pseudoinverse is used. The definition of b^* must change, however; it must now have the form

$$b^* = \begin{pmatrix} b_1^* \\ 0 \end{pmatrix}$$

so partitioning seems to be needed here. Alternatively, we can say that b^* is the *constrained* maximizer of

$$l(a + M\alpha + Zb) - \tfrac{1}{2} b^T D^+ b \tag{12}$$

considered as a function of b with α and θ fixed subject to the constraints that the elements of b corresponding to diagonal components of D that are zero are constrained to be zero. With this introduction of constrained optimization, partitioning is no longer necessary at all. This finishes our description of the method of Breslow and Clayton (1993, Section 2.3) for dealing with zero variance components.

1.5 A Digression on Partial Derivatives

Let $f(\alpha, c, \theta)$ be a scalar-valued function of three vector variables. We write partial derivative vectors using subscripts: $f_\alpha(\alpha, c, \theta)$ 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 $f_c(\alpha, c, \theta)$. We also use this convention for partial derivatives with respect to single variables: $f_{\theta_k}(\alpha, c, \theta)$, 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: $f_{\alpha c}(\alpha, c, \theta)$ 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 c_j. Thus the row dimension of $f_{\alpha c}(\alpha, c, \theta)$ is the dimension of α, the column dimension is the dimension of c, and $f_{c\alpha}(\alpha, c, \theta)$ is the transpose of $f_{\alpha c}(\alpha, c, \theta)$.

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

We also use this convention of subscripts denoting partial derivatives with vector-valued functions. If $f(\alpha, c, \theta)$ is a column-vector-valued function of vector variables, then $f_\alpha(\alpha, c, \theta)$ denotes the matrix of partial derivatives having i, j component that is the partial derivative of the i-th component of $f_\alpha(\alpha, c, \theta)$ with respect to α_j. Thus the row dimension of $f_\alpha(\alpha, c, \theta)$ is the dimension of $f(\alpha, c, \theta)$ and the column dimension is the dimension of α.

1.6 Nearly Zero Variance Components

The method described in Section 1.4 for dealing with zero variance components (due to Breslow and Clayton, 1993, Section 2.3) does not help with variance components that are nearly zero but not exactly zero, in which case there may be huge components of D^+ making calculation of (12) problematic due to inexactness of computer arithmetic. Another way to look at the problem is that the Moore-Penrose pseudoinverse operation $D \mapsto D^+$ is not continuous: components of D^+ corresponding to components of D that are nearly but not exactly zero are huge, whereas components of D^+ corresponding to components of D that are exactly zero are themselves exactly zero.

So the question arises: is the approximate log likelihood derived by Laplace approximation a continuous function of the parameters α and θ, or is it discontinuous? And, whatever the answer to that question is, how do we deal with any discontinuity or apparent discontinuity so inexact computer arithmetic does not cause computational problems?

To start an attack on this problem, we start with the case where all variance components are nonzero, although perhaps nearly zero, and define

$$c = D^{-1/2} b, \tag{13}$$

where $D^{-1/2}$ is the diagonal matrix whose diagonal components are $1 / \sqrt{d_{ii}}$, where d_{ii} are the diagonal components of D, where in this paragraph only subscripts indicate components rather than partial derivatives, so

$$b = D^{1/2} c \tag{14}$$

where $D^{1/2}$ is the diagonal matrix whose diagonal components are $\sqrt{d_{ii}}$. We use the substitution (14) everywhere. Then the penalized log likelihood for estimating $c^* = D^{-1/2} b^*$ is

$$h(c) = l(a + M \alpha + Z D^{1/2} c) - \tfrac{1}{2} c^T c \tag{15}$$

Since (15) is a strictly concave function with bounded level sets, the maximizer c^* exists and is unique. In fact if we replace $D^{1/2}$ by an arbitrary matrix in (15), the maximizer still exists and is unique.

We now choose the parameter θ. We choose the components of θ to be real numbers whose squares are variance components. Thus we have one component of θ for each distinct variance component. We allow negative components of θ so our reparameterization is not identifiable, because the variance components do not depend on the signs of the components of θ only on their absolute values. Also define a diagonal matrix A having components of θ as its diagonal components and satisfying $A^2 = D$, so the diagonal components of A and D correspond. We can replace (15) with

$$h(c) = l(a + M \alpha + Z A c) - \tfrac{1}{2} c^T c \tag{16}$$

if we also replace (14) by

$$b = A c. \tag{17}$$

If all components of θ are nonnegative, then $A = D^{1/2}$, and (15) and (16) are the same as are (14) and (17). But since components of θ are allowed to be negative, (16) and (17) are more general. When components are negative, nothing bad happens. The components of c corresponding to negative components of θ are the negative of what they would be if sign of the corresponding component of θ were positive. By the argument given above, the maximizer c^* of (16) exists and is unique.

Doug Bates (personal communication) says the reparameterization (17) is part of the folklore and used in various R packages of which he is an author (`nlme`, `lme`, `lme4`, `lmer`).

The derivative of (16) is

$$h_c(c) = AZ^T \bigl[y - \mu(a + M\alpha + ZAc)\bigr] - c \tag{18}$$

This is equal to zero when c is equal to the maximizer c^* of (16), so

$$c^* = AZ^T \bigl[y - \mu(a + M\alpha + ZAc^*)\bigr] \tag{19}$$

from which we see that components of c^* corresponding to zero variance components are actually zero (because they are something finite multiplied by the corresponding component of θ, which is zero). This fact is also clear from the fact that components of c corresponding to components of A that are zero appear in (16) only in the second term, which is clearly maximized when they are zero.

Now consider minus the right-hand side of (18) considered as a function of all the variables

$$f(\alpha, c, \theta) = c - AZ^T \bigl[y - \mu(a + M\alpha + ZAc)\bigr] \tag{20}$$

which has derivatives

$$\begin{aligned} f_\alpha(\alpha, c, \theta) &= AZ^T W(a + M\alpha + ZAc)M && (21) \\ f_c(\alpha, c, \theta) &= AZ^T W(a + M\alpha + ZAc)ZA + I && (22) \\ f_{\theta_k}(\alpha, c, \theta) &= -E_k Z^T \bigl[y - \mu(a + M\alpha + ZAc)\bigr] \\ &\qquad + AZ^T W(a + M\alpha + ZAc)ZE_k c && (23) \end{aligned}$$

where

$$E_k = A_{\theta_k}(\theta) \tag{24}$$

is the diagonal matrix whose components are equal to one if the corresponding components of A are equal to θ_k by definition (rather than by accident when some other component of θ also has the same value) and whose components are otherwise zero. For now the only important point is that $f_c(\alpha, c, \theta)$ is strictly positive definite because $W(a+M\alpha+ZAc)$ is positive semidefinite, being a variance matrix by (2).

The implicit function theorem (Browder, 1996, Theorem 8.29) says that $c^*(\alpha, \theta)$ is locally well defined and differentiable. Our previous analysis shows that it is actually globally well defined, so the implicit function theorem only adds the knowledge that it is differentiable. Derivatives are given by

$$c^*_\alpha(\alpha, \theta) = -f_c\bigl(\alpha, c^*(\alpha, \theta), \theta\bigr)^{-1} f_\alpha\bigl(\alpha, c^*(\alpha, \theta), \theta\bigr) \tag{25}$$

$$c^*_{\theta_k}(\alpha, \theta) = -f_c\bigl(\alpha, c^*(\alpha, \theta), \theta\bigr)^{-1} f_{\theta_k}\bigl(\alpha, c^*(\alpha, \theta), \theta\bigr) \tag{26}$$

We now know that c^* is a differentiable function of α and θ. Hence the approximate log likelihood

$$\begin{aligned} q(\alpha, \theta) = l(a + M\alpha + ZAc^*) - \tfrac{1}{2}(c^*)^T c^* \\ - \tfrac{1}{2} \log \det\bigl[AZ^T \widehat{W} ZA + I\bigr] \end{aligned} \tag{27}$$

is a differentiable function.

There is, however, a question about whether (27) is correct. The middle term should be $\frac{1}{2}(c^*)^T A D^+ A c^*$ rather than what it is. Are these the same? This is the sum of terms $(c^*_i)^2 a^2_{ii} d^+_{ii}$ because of A and D^+ being diagonal matrices, where in this paragraph only subscripts indicate components rather than partial derivatives. For i such that $a_{ii} \neq 0$, we have $d^+_{ii} = 1/d_{ii}$ and the term is $\frac{1}{2}(c^*_i)^2$, which is the same as the corresponding term of $\frac{1}{2}(c^*)^T c^*$. For i such that $a_{ii} = 0$, we have $d^+_{ii} = 0$ and the term is zero, which is the same as the corresponding term of $\frac{1}{2}(c^*)^T c^*$, because $c^*_i = 0$. Thus we see that (27) is a correct approximate log likelihood and is differentiable for all α and θ.

1.7 Approximate Maximum Likelihood Estimates

As Breslow and Clayton (1993) also note, an immediate consequence of treating W as a constant function of its argument is that then only the first term of (27) contains α. Let $\tilde{\alpha}$ denote the point at which the maximum of (27) considered as a function of α for fixed θ is achieved. Since the first term of (27) is a strictly concave function of α, the maximizer is unique if it exists. Then define $\tilde{c}(\theta) = c^*\bigl(\tilde{\alpha}(\theta), \theta\bigr)$.

Clearly, we get the same thing if we jointly maximize

$$g(\alpha, c) = l(a + M\alpha + ZAc) - \tfrac{1}{2} c^T c \tag{28}$$

which is the right-hand side of (16) considered as a function of α and c for fixed θ, that is, the maximizer is $(\tilde{\alpha}, \tilde{c})$.

The latter process is much easier, because (28) is a strictly concave function (Barndorff-Nielsen, 1978, Theorem 9.1) and hence every local maximizer is the unique global maximizer (local maximizers need not exist, but typically they will; if they do not the `summary` function applied to the aster model fit for the model containing only fixed effects and having model matrix M will complain about "possible directions of recession").

Thus we arrive at

$$\begin{aligned} r(\theta) &= q\bigl(\tilde{\alpha}(\theta), \theta\bigr) \\ &= l(a + M\tilde{\alpha} + ZA\tilde{c}) - \tfrac{1}{2}\tilde{c}^T\tilde{c} - \tfrac{1}{2}\log\det\bigl[AZ^T\widehat{W}ZA + I\bigr] \end{aligned} \tag{29}$$

as an (approximate) profile log likelihood for θ, where on the right-hand side $\tilde{\alpha}$, $\tilde{c}$, and A are functions of θ, although this is not indicated by the notation.

Maximizing (29) gives an estimate $\hat{\theta}$ of θ. Then

$$\begin{aligned} \hat{\alpha} &= \tilde{\alpha}(\hat{\theta}) \\ \hat{b} &= A(\hat{\theta})\tilde{c}(\hat{\theta}) \end{aligned}$$

are "estimates" of the corresponding quantities. Since α is a parameter vector, $\hat{\alpha}$ is a (vector) parameter estimate. Since the random effect vector b is not a parameter, $\hat{b}$ is not a parameter estimate. It is a competitor of BLUP predictions of random effects in normal-normal (normal response and normal random effects) random effects models.

1.8 Halftime Summary

All of this is quite confusing. So we recap. The key quantity is

$$p(\alpha, c, \theta) = l(a + M\alpha + ZAc) - \tfrac{1}{2}c^Tc - \tfrac{1}{2}\log\det\bigl[AZ^T\widehat{W}ZA + I\bigr] \tag{30}$$

where, as the left-hand side says, α, c, and θ are all free variables and, as usual, A is a function of θ, although the notation does not indicate this.

1.8.1 Joint and Profiles

If we maximize (30) considered as a function of c for fixed α and θ we get c^*, which is a function of α and θ. If we maximize (30) considered as a function of α and c for fixed θ we get $\tilde{\alpha}$ and $\tilde{c}$, both of which are functions of θ (only). Then we have

$$q(\alpha, \theta) = p\bigl(\alpha, c^*(\alpha, \theta), \theta\bigr) \tag{31}$$

$$r(\theta) = p\bigl(\tilde{\alpha}(\theta), \tilde{c}(\theta), \theta\bigr) \tag{32}$$

which repeat (27) and (29). Both of these are of interest. In fact, since we know that maximizing a profile is the same as maximizing the original function, if we let $(\hat{\alpha}, \hat{c}, \hat{\theta})$ denote the joint maximizer of (30), then $\hat{\theta}$ is the maximizer of (32) and $(\hat{\alpha}, \hat{\theta})$ is the joint maximizer of (31) and $\hat{\alpha} = \tilde{\alpha}(\hat{\theta})$ and $\hat{c} = c^*(\hat{\alpha}, \hat{\theta}) = \tilde{c}(\hat{\theta})$. These are just different ways of describing the joint optimizer of the key quantity (30).

1.8.2 Treating W as Nonconstant

In computing estimators, the easiest and simplest method will be to just optimize (30) directly with no optimization of functions themselves computed by optimization (profiles), or at least it would be easiest and simplest if we did not ever want to adjust $\widehat{W}$. Since we do want to adjust $\widehat{W}$, we start with a procedure that is none of these, optimizing

$$\begin{aligned} s(\theta) = l(a + M\tilde{\alpha} + ZA\tilde{c}) - \tfrac{1}{2}\tilde{c}^T\tilde{c} \\ - \tfrac{1}{2}\log\det\bigl[AZ^T W(a + M\tilde{\alpha} + ZA\tilde{c})ZA + I\bigr] \end{aligned} \tag{33}$$

where on the right-hand side $\tilde{\alpha}$ and $\tilde{c}$ and A are all functions of θ although the notation does not indicate this. Of course, (33) is what we do not know how to differentiate because we do not know derivatives of $W(\,\cdot\,)$. Thus we will have to use a no-derivative method of optimization to get close to the solution. Then we can switch to optimizing (30), for which we do have first and second derivatives.

1.8.3 Fisher Information

In doing inference, neither (30) nor (32) are helpful because they are not log likelihoods nor even approximate log likelihoods. Their derivatives do not give anything even approximating Fisher information (the derivative of a profile likelihood does not give Fisher information). Thus only (31) is helpful in deriving approximate standard errors. We are treating (31) as an approximate log likelihood, so minus its second derivative matrix is approximate observed Fisher information.

1.9 First Derivatives

Start with (30). Its derivatives are

$$p_\alpha(\alpha, c, \theta) = M^T\bigl[y - \mu(a + M\alpha + ZAc)\bigr] \tag{34}$$

$$p_c(\alpha, c, \theta) = AZ^T\bigl[y - \mu(a + M\alpha + ZAc)\bigr] - c \tag{35}$$

and

$$
\begin{aligned}
p_{\theta_k}(\alpha, c, \theta) &= c^T E_k Z^T \bigl[y - \mu(a + M\alpha + ZAc) \bigr] \\
&\qquad - \tfrac{1}{2} \operatorname{tr} \Bigl(\bigl[A Z^T \widehat{W} Z A + I \bigr]^{-1} \bigl[E_k Z^T \widehat{W} Z A + A Z^T \widehat{W} Z E_k \bigr] \Bigr) \\
&= c^T E_k Z^T \bigl[y - \mu(a + M\alpha + ZAc) \bigr] \\
&\qquad - \operatorname{tr} \Bigl(\bigl[A Z^T \widehat{W} Z A + I \bigr]^{-1} A Z^T \widehat{W} Z E_k \Bigr)
\end{aligned}
\tag{36}
$$

where E_k is given by (24). The formula for the derivative of the log of a determinant of a symmetric matrix comes from Searle et al. (1992, Appendix M, Section 7.f). The simplification of the trace in (36) is the fact that for any square matrices U and V we have $\operatorname{tr}(U+V) = \operatorname{tr}(U) + \operatorname{tr}(V)$ and $\operatorname{tr}(UV) = \operatorname{tr}(VU) = \operatorname{tr}(V^T U^T)$ (Harville, 1997, Sections 5.1 and 5.2). Some of this repeats work done in Section 1.6, but (36) is new.

The estimating equation for c^* can be written

$$
p_c\bigl(\alpha, c^*(\alpha, \theta), \theta\bigr) = 0 \tag{37}
$$

which repeats (19). And the estimating equations for $\tilde{\alpha}$ and $\tilde{c}$ can be written

$$
p_\alpha\bigl(\tilde{\alpha}(\theta), \tilde{c}(\theta), \theta\bigr) = 0 \tag{38}
$$

$$
p_c\bigl(\tilde{\alpha}(\theta), \tilde{c}(\theta), \theta\bigr) = 0 \tag{39}
$$

Actually, (39) is a consequence of (37), because of $\tilde{c}(\theta) = c^*\bigl(\tilde{\alpha}(\theta), \theta\bigr)$. All of these are useful in simplifying expressions for derivatives.

Now we have

$$
\begin{aligned}
q_\alpha(\alpha, \theta) &= p_\alpha(\alpha, c^*, \theta) + c^*_\alpha(\alpha, \theta)^T p_c(\alpha, c^*, \theta) \\
&= p_\alpha(\alpha, c^*, \theta)
\end{aligned}
\tag{40}
$$

by (37), and

$$
\begin{aligned}
q_{\theta_k}(\alpha, \theta) &= c^*_{\theta_k}(\alpha, \theta)^T p_c(\alpha, c^*, \theta) + p_{\theta_k}(\alpha, c^*, \theta) \\
&= p_{\theta_k}(\alpha, c^*, \theta)
\end{aligned}
\tag{41}
$$

again by (37), and

$$
\begin{aligned}
r_{\theta_k}(\theta) &= \tilde{\alpha}_{\theta_k}(\theta)^T p_\alpha(\tilde{\alpha}, \tilde{c}, \theta) + \tilde{c}_{\theta_k}(\theta)^T p_c(\tilde{\alpha}, \tilde{c}, \theta) + p_{\theta_k}(\tilde{\alpha}, \tilde{c}, \theta) \\
&= p_{\theta_k}(\tilde{\alpha}, \tilde{c}, \theta)
\end{aligned}
\tag{42}
$$

by (38) and (39).

Now by definition of $\hat{\theta}$ we have (42) equal to zero at $\theta = \hat{\theta}$, that is,

$$r_{\theta_k}(\hat{\theta}) = p_{\theta_k}(\hat{\alpha}, \hat{c}, \hat{\theta}) = 0 \tag{43}$$

and this implies

$$q_{\theta_k}(\hat{\alpha}, \hat{\theta}) = p_{\theta_k}(\hat{\alpha}, \hat{c}, \hat{\theta}) = 0$$

and we already have

$$q_\alpha\bigl(\tilde{\alpha}(\theta), \theta\bigr) = p_\alpha\bigl(\tilde{\alpha}(\theta), \tilde{c}(\theta), \theta\bigr) = 0$$

holding for all θ and, in particular, for $\theta = \hat{\theta}$ by (38). Thus derivatives of the (approximate) log likelihood and profile log likelihood are zero at the (approximate) maximum likelihood estimators. This must, of course, have been the case, but it is good that everything checks.

1.10 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). We do only second derivatives of q, ignoring r, because it is not yet clear that second derivatives of r are useful. (It turns out that r is not used at all in the function `reaster` that calculates random effects aster models in the R package `aster` (Geyer, 2012). See Section 2 below.)

Now by the multivariate chain rule (Browder, 1996, Theorem 8.15)

$$\begin{aligned}
q_{\alpha\alpha}(\alpha, \theta) &= p_{\alpha\alpha}(\alpha, c^*, \theta) + p_{\alpha c}(\alpha, c^*, \theta) c^*_\alpha(\alpha, \theta) \\
q_{\alpha\theta_k}(\alpha, \theta) &= p_{\alpha\theta_k}(\alpha, c^*, \theta) + p_{\alpha c}(\alpha, c^*, \theta) c^*_{\theta_k}(\alpha, \theta) \\
q_{\theta_j\theta_k}(\alpha, \theta) &= p_{\theta_j\theta_k}(\alpha, c^*, \theta) + p_{\theta_j c}(\alpha, c^*, \theta) c^*_{\theta_k}(\alpha, \theta)
\end{aligned}$$

The derivatives of c^* needed here have already been derived in (25) and (26), but we note that

$$f(\alpha, c, \theta) = -p_c(\alpha, c, \theta)$$

so (25) and (26) can also be written

$$c^*_\alpha(\alpha, \theta) = -p_{cc}(\alpha, c^*, \theta)^{-1} p_{c\alpha}(\alpha, c^*, \theta) \tag{44}$$

$$c^*_{\theta_k}(\alpha, \theta) = -p_{cc}(\alpha, c^*, \theta)^{-1} p_{c\theta_k}(\alpha, c^*, \theta) \tag{45}$$

and the second derivatives above can be rewritten

$$\begin{aligned}
q_{\alpha\alpha}(\alpha, \theta) &= p_{\alpha\alpha}(\alpha, c^*, \theta) - p_{\alpha c}(\alpha, c^*, \theta) p_{cc}(\alpha, c^*, \theta)^{-1} p_{c\alpha}(\alpha, c^*, \theta) \\
q_{\alpha\theta_k}(\alpha, \theta) &= p_{\alpha\theta_k}(\alpha, c^*, \theta) - p_{\alpha c}(\alpha, c^*, \theta) p_{cc}(\alpha, c^*, \theta)^{-1} p_{c\theta_k}(\alpha, c^*, \theta) \\
q_{\theta_j\theta_k}(\alpha, \theta) &= p_{\theta_j\theta_k}(\alpha, c^*, \theta) - p_{\theta_j c}(\alpha, c^*, \theta) p_{cc}(\alpha, c^*, \theta)^{-1} p_{c\theta_k}(\alpha, c^*, \theta)
\end{aligned}$$

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

$$q_{\psi\psi}(\psi) = p_{\psi\psi}(\psi, c^*) - p_{\psi c}\bigl(\psi, c^*\bigr) p_{cc}\bigl(\psi, c^*\bigr)^{-1} p_{c\psi}\bigl(\psi, c^*\bigr) \tag{46}$$

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

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \tag{47}$$

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

$$\Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \tag{48}$$

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 (46).

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)

$$\begin{aligned}
p_{\alpha\alpha}(\alpha, c, \theta) &= -M^T W(a + M\alpha + ZAc) M \\
p_{\alpha c}(\alpha, c, \theta) &= -M^T W(a + M\alpha + ZAc) ZA \\
p_{cc}(\alpha, c, \theta) &= -AZ^T W(a + M\alpha + ZAc) ZA - I \\
p_{\alpha\theta_k}(\alpha, c, \theta) &= -M^T W(a + M\alpha + ZAc) ZE_k c \\
p_{\theta_k c}(\alpha, c, \theta) &= \bigl[y - \mu(a + M\alpha + ZAc)\bigr]^T ZE_k \\
&\qquad - c^T E_k Z^T W(a + M\alpha + ZAc) ZA \\
p_{\theta_j \theta_k}(\alpha, c, \theta) &= -c^T E_j Z^T W(a + M\alpha + ZAc) ZE_k c \\
&\qquad - \operatorname{tr}\Bigl(\bigl[AZ^T \widehat{W} ZA + I\bigr]^{-1} E_k Z^T \widehat{W} ZE_j\Bigr) \\
&\qquad + \operatorname{tr}\Bigl(\bigl[AZ^T \widehat{W} ZA + I\bigr]^{-1} \\
&\qquad\qquad \bigl[E_k Z^T \widehat{W} ZA + AZ^T \widehat{W} ZE_k\bigr] \\
&\qquad\qquad \bigl[AZ^T \widehat{W} ZA + I\bigr]^{-1} AZ^T \widehat{W} ZE_j\Bigr)
\end{aligned}$$

where the formula for the derivative of a matrix inverse comes from Searle et al. (1992, Appendix M, Section 7.e). This finishes the derivation of all the derivatives we need.

1.11 Fisher Information

The observed Fisher information matrix is minus the second derivative matrix of the log likelihood.

Assembling stuff derived in preceding sections and introducing

$$\begin{aligned}
\mu^* &= \mu\bigl(a + M\alpha + ZAc^*(\alpha, \theta)\bigr) \\
W^* &= W\bigl(a + M\alpha + ZAc^*(\alpha, \theta)\bigr) \\
H^* &= AZ^T W^* ZA + I \\
\widehat{H} &= AZ^T \widehat{W} ZA + I
\end{aligned}$$

we obtain for the α, α block of the observed Fisher information matrix

$$\begin{aligned}
-q_{\alpha\alpha}(\alpha, \theta) &= M^T W^* M - M^T W^* ZA f_c\bigl(\alpha, c^*, \theta\bigr)^{-1} f_\alpha\bigl(\alpha, c^*, \theta\bigr) \\
&= M^T W^* M - M^T W^* ZA (H^*)^{-1} AZ^T W^* M
\end{aligned}$$

for the α, θ_k block of the observed Fisher information matrix

$$\begin{aligned}
-q_{\alpha \theta_k}(\alpha, \theta) &= M^T W^* Z E_k c^* - M^T W^* Z A f_c\bigl(\alpha, c^*, \theta\bigr)^{-1} f_{\theta_k}\bigl(\alpha, c^*, \theta\bigr) \\
&= M^T W^* Z E_k c^* \\
&\qquad - M^T W^* Z A (H^*)^{-1} \bigl[A Z^T W^* Z E_k c^* - E_k Z^T (y - \mu^*) \bigr]
\end{aligned}$$

and for the θ_j, θ_k block of the observed Fisher information matrix

$$\begin{aligned}
-q_{\theta_j \theta_k}(\alpha, \theta) &= (c^*)^T E_j Z^T W^* Z E_k c^* \\
&\qquad + \operatorname{tr}\bigl(\widehat{H}^{-1} E_k Z^T \widehat{W} Z E_j\bigr) \\
&\qquad - \operatorname{tr}\bigl(\widehat{H}^{-1} A Z^T \widehat{W} Z E_k \widehat{H}^{-1} A Z^T \widehat{W} Z E_j\bigr) \\
&\qquad - \operatorname{tr}\bigl(\widehat{H}^{-1} E_k Z^T \widehat{W} Z A \widehat{H}^{-1} A Z^T \widehat{W} Z E_j\bigr) \\
&\qquad - \bigl[(c^*)^T E_j Z^T W^* Z A - (y - \mu^*)^T Z E_j \bigr] \\
&\qquad\qquad f_c\bigl(\alpha, c^*, \theta\bigr)^{-1} f_{\theta_k}\bigl(\alpha, c^*, \theta\bigr) \\
&= (c^*)^T E_j Z^T W^* Z E_k c^* \\
&\qquad + \operatorname{tr}\bigl(\widehat{H}^{-1} E_k Z^T \widehat{W} Z E_j\bigr) \\
&\qquad - \operatorname{tr}\bigl(\widehat{H}^{-1} A Z^T \widehat{W} Z E_k \widehat{H}^{-1} A Z^T \widehat{W} Z E_j\bigr) \\
&\qquad - \operatorname{tr}\bigl(\widehat{H}^{-1} E_k Z^T \widehat{W} Z A \widehat{H}^{-1} A Z^T \widehat{W} Z E_j\bigr) \\
&\qquad - \bigl[A Z^T W^* Z E_j c^* - E_j Z^T (y - \mu^*) \bigr]^T \\
&\qquad\qquad (H^*)^{-1} \bigl[A Z^T W^* Z E_k c^* - E_k Z^T (y - \mu^*) \bigr]
\end{aligned}$$

In all of these c^*, μ^*, W^*, and H^* are functions of α and θ even though the notation does not indicate this and A is a function of θ 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

$$E(y \mid c) = \mu(a + M\alpha + ZAc)$$

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.

1.12 Penalized Likelihood Calculation

For penalized likelihood calculation of either c^* or (jointly) of $\tilde{\alpha}$ and $\tilde{c}$ we need first and second derivatives of the objective function, which is given by (15) in the first case and by (28) in the second case.

In the first case we have derivatives

$$\begin{aligned} h_c(c) &= p_c(\alpha, c, \theta) \\ h_{cc}(c) &= p_{cc}(\alpha, c, \theta) \end{aligned}$$

the first of these agreeing with (18). In the second case we have derivatives

$$\begin{aligned} g_\alpha(c) &= p_\alpha(\alpha, c, \theta) \\ g_c(c) &= p_c(\alpha, c, \theta) \\ g_{\alpha,\alpha}(c) &= p_{\alpha\alpha}(\alpha, c, \theta) \\ g_{\alpha c}(c) &= p_{\alpha c}(\alpha, c, \theta) \\ g_{cc}(c) &= p_{cc}(\alpha, c, \theta) \end{aligned}$$

1.13 Standard Errors for Random Effects

Suppose that the approximate Fisher information derived in Section 1.11 can be used to give an approximate asymptotic variance for the parameter vector $\psi = (\alpha, \theta)$. This would be $q_{\psi\psi}(\hat{\psi}, \hat{c})^{-1}$, where $q_{\psi\psi}(\psi, c^*)$ is given by (46) and $\hat{\psi} = (\hat{\alpha}, \hat{\theta})$ and $\hat{c} = c^*(\hat{\alpha}, \hat{\theta})$.

We would like standard errors for the point estimates of the random effects

$$\hat{b} = \hat{A}\hat{c} = u(\hat{\alpha}, \hat{\theta}) \tag{49}$$

where

$$u(\alpha, \theta) = A(\theta) c^*(\alpha, \theta)$$

To apply the delta method to get asymptotic standard errors for $\hat{b}$ we need the derivatives

$$\begin{aligned} u_\alpha(\alpha, \theta) &= A(\theta) c^*_\alpha(\alpha, \theta) \\ &= -A(\theta) c^*_\alpha(\alpha, \theta) p_{cc}(\alpha, c^*, \theta)^{-1} p_{c\alpha}(\alpha, c^*, \theta) \\ u_{\theta_k}(\alpha, \theta) &= E_k c^*(\alpha, \theta) + A(\theta) c^*_\alpha(\alpha, \theta) \\ &= E_k c^*(\alpha, \theta) - A(\theta) p_{cc}(\alpha, c^*, \theta)^{-1} p_{c\theta_k}(\alpha, c^*, \theta) \end{aligned}$$

which use (24), (44), and (45). Stacking these we obtain

$$u_\psi(\hat{\psi}) = \begin{pmatrix} -A(\hat{\theta}) p_{cc}(\hat{\alpha}, \hat{c}, \hat{\theta})^{-1} p_{c\alpha}(\hat{\alpha}, \hat{c}, \hat{\theta}) \\ E_k\hat{c} - A(\hat{\theta}) p_{cc}(\hat{\alpha}, \hat{c}, \hat{\theta})^{-1} p_{c\theta}(\hat{\alpha}, \hat{c}, \hat{\theta}) \end{pmatrix}$$

and the delta method gives

$$u_\psi(\hat{\psi})^T q_{\psi,\psi}(\hat{\psi}) u_\psi(\hat{\psi}) \tag{50}$$

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

It must be conceded that we are living what true believers in random effects models would consider a state of sin in this section. The random effects vector b is not a parameter, yet (49) treats it as a function of parameters (which is thus a parameter) and the "asymptotic variance" (50) is derived by considering $\hat{b}$ just such a parameter estimate. So (50) 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 (50) 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 either the approximate log likelihood (27) or the approximate profile log likelihood (29), 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. 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

2.1 Step 1

To get close to $\hat{\theta}$ starting from far away we minimize the function $s(\theta)$ defined by (33). For starting points of this optimization we first set α equal to the MLE for the fixed effects model (leaving out random effects) and set all components of θ to one. Then we evaluate $c^*(\alpha, \theta)$ at this starting point and then set θ_k to be the empirical standard deviation of the components of $c^*(\alpha, \theta)$ that have θ_k as their standard deviation. We then use this θ as the starting point for the minimization of $s(\theta)$

Because we cannot calculate derivatives of $s(\theta)$, we optimize 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.

Evaluation of $s(\theta)$ requires an inner optimization to evaluate $\tilde{\alpha}(\theta)$ and $\tilde{c}(\theta)$. Since we have first and second derivatives of the objective function $g(\alpha, c)$ whose joint optimizer is $(\tilde{\alpha}, \tilde{c})$ we can use a method of optimization that uses these derivatives (given in Section 1.12). The current version of the `reaster` function in the `aster` package (Geyer, 2012) uses the optimizer `trust` in the `trust` package to do this inner optimization that occurs inside each evaluation of $s(\theta)$.

A minor technical detail is the starting value for the inner optimization. We need a value of (α, c) to start at, and it would be silly to use the same starting point for every inner optimization. Even if we had a good starting point at the beginning, it would rapidly become bad as θ changes throughout the minimization. Since every inner optimization maximizes the function $g(\alpha, c)$ which is encoded as an R function `penmlogl`, which just calculates $g(\alpha, c)$ and its derivatives, there is no way to return the "good" values of (α, θ) found in each inner optimization for use in the next inner optimization. We could, of course, just write these in the R global environment, using it as a scratchpad, but it is better programming practice to not define things in the R global environment that users will wonder where they came from (and might even clobber user defined objects if there was a name collision), thus we write it down in a special environment, which is an argument `cache` to the function `pickle` that evaluates $s(\theta)$. Each evaluation of $s(\theta)$ "remembers" (in the environment `cache`) the value of (α, c) which was the optimal value of $g(\alpha, c)$ in this evaluation. Then that value is used as the starting point for the next inner optimization.

2.2 Step 2

Having found a θ close to $\hat{\theta}$ via the preceding step, we then set α and c to $\tilde{\alpha}(\theta)$ and $\tilde{c}(\theta)$, respectively, to obtain a starting point for the minimization of $p(\alpha, c, \theta)$ given by (30), which is computed by the R function `pickle3` in the `aster` package. To define (30) we also need a $\widehat{W}$ and we take the value at the current values of α, c, and θ. 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

$$Z^T \widehat{W} Z = Z^T W(a + M\alpha + ZAc)Z \tag{51}$$

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

Because we have first and second derivatives of $p(\alpha, c, \theta)$ we can use these derivatives in the optimization. The current version of the `reaster` function in the `aster` package uses the optimizer `trust` in the `trust` package.

Having minimized $p(\alpha, c, \theta)$ we are still not done, because now (51) is wrong. We held it fixed at the values of α, c, and θ we had before the minimization, and now those values have changed. Thus we should re-evaluate (51) and re-minimize, and continue doing this until convergence. We terminate this iteration when θ values do not change (to within some prespecified tolerance) because the α and c values are, in theory, determined by θ, as $\tilde{\alpha}(\theta)$ and $\tilde{c}(\theta)$, respectively, 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 $\hat{\theta}$. We also have our points estimates $\hat{b}$ given by (49) of the random effects on the original scale.

2.3 Step 3

Point estimation is now done. All that remains is computation of standard errors. Now we come to an issue that came as a shock to your humble author (though in hindsight, it shouldn't have). The beautiful formula (46) does not work in a computer because of inexactness of computer arithmetic. When the Hessian matrix of $p(\alpha, c, \theta)$

$$\begin{pmatrix} p_{\alpha\alpha}(\alpha, c, \theta) & p_{\alpha c}(\alpha, c, \theta) & p_{\alpha\theta}(\alpha, c, \theta) \\ p_{c\alpha}(\alpha, c, \theta) & p_{cc}(\alpha, c, \theta) & p_{c\theta}(\alpha, c, \theta) \\ p_{\theta\alpha}(\alpha, c, \theta) & p_{\theta c}(\alpha, c, \theta) & p_{\theta\theta}(\alpha, c, \theta) \end{pmatrix}$$

is ill conditioned (when the ratio of its largest and smallest eigenvalues is large) there is no reason to expect (46) to yield a positive definite matrix.

The minus sign in (46) is the culprit. Subtracting two large positive quantities that are supposed to have a small positive difference doesn't always work in inexact computer arithmetic. This is called "catastrophic cancellation." Worse we see no way to rearrange the computation to avoid catastrophic cancellation.

There is, however, another way to get at $q_{\psi\psi}(\psi)$. We can compute its Hessian matrix by finite-difference approximation. We hand the function $q(\alpha, \theta)$, which is computed by the function `pickle2` in the `aster` package, to the R function `optim` with `method = "BFGS"`, a quasi-Newton optimizer, and option `hessian = TRUE`, asking it to compute the Hessian by finite differences. Because we start at the (already found) solution $(\hat{\alpha}, \hat{\theta})$ the only work that needs to be done is calculating the Hessian. Because we provide first derivatives, given by (34), (35), and (36), the calculation of the Hessian is as efficient as possible (given that we have no useful analytic formula). This gives a good estimate of $q_{\psi\psi}(\hat{\psi})$ that can be used to calculate standard errors of the parameters α and θ. These standard errors are the square roots of the diagonal elements of $q_{\psi\psi}(\hat{\psi})^{-1}$.

Then we use (50) to calculate standard errors of $\hat{b}$. There is no catastrophic cancellation here.

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?

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. (2012). aster: Aster Models. R package version 0.8-11. `http://www.stat.umn.edu/geyer/aster/`.

Geyer, C. J. (in press). Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. To appear in *Festschrift for Morris L. Eaton*, G. Jones and X. Shen eds. 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.

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.