LOCAL ADAPTATION AND GENETIC EFFECTS ON FITNESS: CALCULATIONS FOR EXPONENTIAL FAMILY MODELS WITH RANDOM EFFECTS

BY CHARLES J. GEYER[*], CAROLINE E. RIDLEY[†], ROBERT G. LATTA[‡], JULIE R. ETTERSON[§] AND RUTH G. SHAW[*]

University of Minnesota[*], US Environmental Protection Agency[†], Dalhousie University[‡] and University of Minnesota, Duluth[§]

Random effects are implemented for aster models using two approximations taken from Breslow and Clayton [*J. Amer. Statist. Assoc.* **88** (1993) 9–25]. Random effects are analytically integrated out of the Laplace approximation to the complete data log likelihood, giving a closed-form expression for an approximate missing data log likelihood. Third and higher derivatives of the complete data log likelihood with respect to the random effects are ignored, giving a closed-form expression for second derivatives of the approximate missing data log likelihood, hence approximate observed Fisher information. This method is applicable to any exponential family random effects model. It is implemented in the CRAN package aster (R Core Team [R: A Language and Environment for Statistical Computing (2012) R Foundation for Statistical Computing], Geyer [R package aster (2012) http://cran.r-project.org/package=aster]). Applications are analyses of local adaptation in the invasive California wild radish (*Raphanus sativus*) and the slender wild oat (*Avena barbata*) and of additive genetic variance for fitness in the partridge pea (*Chamaecrista fasciculata*).

1. Introduction. Aster models [Geyer, Wagenius and Shaw (2007), Shaw et al. (2008)] are a partial generalization of generalized linear models (GLM) that allow different components of the response vector to have different families (some Bernoulli, some Poisson, some zero-truncated Poisson, some normal) and also to be dependent, the dependence being specified by a simple graphical model. Because of the way they incorporate dependence among components of the response, aster models are not GLM nor like other regression models with which statisticians are familiar, but they are special cases of graphical models and of exponential families. Although aster models can be used whenever their assumptions hold [for which see Geyer, Wagenius and Shaw (2007)], they were designed particularly for life history analysis of plants and animals, which aims to model total lifetime reproductive output (observed Darwinian fitness), a random variable that fits no familiar distribution, often having a large atom at zero (individuals that died

before producing offspring) as well as multiple modes. Aster models can fit such data adequately by using data on other components of fitness (survival in each year, number of flowers in each year, number of seeds in each year, and number of seeds that germinate in each year for a plant and similar sorts of data for other organisms) and modeling all these data jointly. It often turns out that, although the marginal distribution of total lifetime reproductive output is intractable, the conditional distribution of each component of the response vector given some other component is tractable (e.g., number of seeds per flower is Poisson). Biologists had recognized for decades that no statistical methods before aster allowed valid statistical analysis of life history data [Shaw et al. (2008)], so aster models are becoming widely used.

Here we extend aster models to allow for random effects. Our applications illustrate three areas where random effects models are traditional. First, when one categorical predictor is nested within another, the effects for the nested predictor are commonly treated as random, especially when they are nuisance parameters. This is seen in both of our analyses of local adaptation. Second, when levels of a categorical predictor (such as years) are not interesting in themselves but only as representatives of a larger population, the corresponding effects are commonly treated as random. This is seen in one of our analyses of local adaptation. Third, Fisher (1918), a paper that was the forerunner of all random effects models, introduced the idea of random effects representing the cumulative effects of many genes. To obtain evolutionary predictions from life history analysis, random effects models are necessary. This is seen in our analysis of genetic variance for fitness. (Mapping the genes that contribute to variation in fitness is not feasible; the number of them is so large, and many are individually of such small effect, that it is unrealistic to generate a sufficiently large study population to detect an informative subset of them [Travisano and Shaw (2013)]. If there were only a few genes for fitness, then sequencing and "machine learning" would help, but there is no sparsity here.)

As with GLM with random effects (generalized linear mixed models, GLMM), aster models with random effects have analytically intractable likelihoods necessitating the use of Monte Carlo, numerical integration or approximate likelihood. Markov chain Monte Carlo likelihood inference has a rich literature [Booth and Hobert (1999), Geyer (1994), Geyer and Thompson (1992), Hummel, Hunter and Handcock (2012), Hunter et al. (2008), Okabayashi and Geyer (2011), Penttinen (1984), Shaw, Geyer and Shaw (2002), Shaw et al. (1999), Thompson and Guo (1991)], but we have avoided it because it is very computationally intensive and also very difficult for ordinary users to do correctly. Ordinary Monte Carlo [Sung and Geyer (2007)] has also been used, but is also very computationally intensive. Numerical integration [Crouch and Spiegelman (1990)] is useful when there is only one variance component but not otherwise [McCulloch (2003), Section 7.2], and we have avoided this too. Approximate integrated likelihood (AIL) is based on the idea that if the complete data log likelihood were quadratic in the random

effects, then the random effects could be integrated out analytically, and if the complete data log likelihood is only close to quadratic in the random effects, then this is a reasonable approximation, usually referred to as Laplace approximation [Breslow and Clayton (1993)]. For sufficiently large sample sizes and sufficiently few random effects, the log likelihood is asymptotically expected to be approximately quadratic [Le Cam and Yang (2000), Chapter 6; Geyer (2013)], so this approximation may work well.

We use a second approximation, also introduced by Breslow and Clayton (1993), that is likewise an assumption that the log likelihood is close to quadratic in the random effects. If the complete data log likelihood were exactly quadratic in the random effects, then all derivatives higher than second would be zero, and we assume this. Since the AIL already involves second derivatives with respect to the random effects of the complete data log likelihood, second derivatives of the log AIL would involve fourth derivatives of the complete data log likelihood and would be computationally intractable. This approximation allows us to compute approximate second derivatives of the log AIL and hence approximate observed Fisher information.

2. Theory of approximate integrated likelihoods.
Although we are particularly interested in aster models, 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

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

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

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. We assume the matrix D is diagonal, so the random effects are independent random variables. The diagonal components of D are called *variance components*.

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. Typically each variance component corresponds to many random effects, so each component of ν occurs multiple times as a diagonal element of D.

In order to agree with the optimization literature, we prefer to minimize rather than maximize. Thus, we use minus log likelihoods. Minus the complete data log likelihood is

$$(4) \qquad -l(a + M\alpha + Zb) + \tfrac{1}{2}b^T D^{-1} b + \tfrac{1}{2} \log \det(D)$$

in case none of the variance components are zero. We deal with the case of zero variance components in Sections 3 and 4.

Let b^* denote the result of minimizing (4) considered as a function of b for fixed α and ν. 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 first term on the right-hand side of (4) is bounded below [Geyer (2009), Theorems 4 and 6] and the second term has bounded level sets. It follows that there is a 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.

We define minus the log AIL to be

$$(5) \qquad \begin{aligned} q(\alpha, \nu) &= -l(a + M\alpha + Zb^*) + \tfrac{1}{2}(b^*)^T D^{-1} b^* \\ &\quad + \tfrac{1}{2} \log \det[Z^T W(a + M\alpha + Zb^*) ZD + I], \end{aligned}$$

where I denotes the identity matrix of the appropriate dimension, where b^* is a function of α and ν and D is a function of ν, although this is not indicated by the notation. Our equation (5) is the negation of equation (5) in Breslow and Clayton (1993), who introduced the terminology *penalized quasi-likelihood* (PQL) for this approach. Minimizing (5) gives approximate maximum likelihood estimates of α and ν, and differentiating (5) twice gives an approximate observed Fisher information matrix.

However, (5) is not easy to differentiate because W is already the second derivative matrix of the cumulant function, so second derivatives of (5) 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 and the software [the R package aster, Geyer (2012)] does not compute them. 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]. Thus, we ignore derivatives higher than second, which is equivalent to assuming W is constant or that c and $-l$ are quadratic.

This leads to the following idea. Rather than basing inference on (5), we actually use

$$(6) \qquad q(\alpha, \nu) = -l(a + M\alpha + Zb^*) + \tfrac{1}{2}(b^*)^T D^{-1} b^* + \tfrac{1}{2} \log \det[Z^T \hat{W} ZD + I],$$

where $\hat{W}$ is a constant matrix (not a function of α and ν). This makes sense for any choice of $\hat{W}$ that is symmetric and positive semidefinite, but we will choose $\hat{W}$ that are close to $W(a + M\hat{\alpha} + Z\hat{b})$, where $\hat{\alpha}$ and $\hat{\nu}$ are the joint minimizers of (5) and $\hat{b} = b^*(\hat{\alpha}, \hat{\nu})$. Note that (6) is a redefinition of $q(\alpha, \nu)$. Hereafter we will no longer use the definition (5).

Introduce

$$(7) \qquad p(\alpha, b, \nu) = -l(a + M\alpha + Zb) + \tfrac{1}{2}b^T D^{-1} b + \tfrac{1}{2}\log\det[Z^T \hat{W} Z D + I],$$

where, as the left-hand side says, α, b and ν are all free variables and, as usual, D is a function of ν. Since the terms that contain b are the same in both (4) and (7), b^* can also be defined as the result of minimizing (7) considered as a function of b for fixed α and ν. Thus, (6) is a profile of (7) and $(\hat{\alpha}, \hat{b}, \hat{\nu})$ is the joint minimizer of (7).

We now switch notation for partial derivatives, using subscripts to indicate derivatives, explained in more detail in Section 1.6 of the accompanying technical report [Geyer et al. (2012)]. Then second derivatives of (6) can be written using the implicit function theorem and the fact that b^* minimizes (7) as

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

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

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

a particularly simple and symmetric form [for a detailed derivation see Sections 1.7 and 1.8 of Geyer et al. (2012)]. 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

$$(8) \qquad q_{\psi\psi}(\psi) = p_{\psi\psi}(\psi, b^*) - p_{\psi b}(\psi, b^*) p_{bb}(\psi, b^*)^{-1} p_{b\psi}(\psi, b^*).$$

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

$$(9) \qquad \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

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

$$(10) \qquad \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

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 $\text{var}(v^T X_1 \mid X_2) = 0$, then

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

almost surely, 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 (8).

The second derivatives of p we need for the second derivatives of q are

$$p_{\alpha\alpha}(\alpha, b, \nu) = M^T W(a + M\alpha + Zb)M,$$

$$p_{\alpha b}(\alpha, b, \nu) = M^T W(a + M\alpha + Zb)Z,$$

$$p_{bb}(\alpha, b, \nu) = Z^T W(a + M\alpha + Zb)Z + D^{-1},$$

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

$$p_{b \nu_k}(\alpha, b, \nu) = -D^{-1} E_k D^{-1} b,$$

$$p_{\nu_j \nu_k}(\alpha, b, \nu) = b^T D^{-1} E_j D^{-1} E_k D^{-1} b$$

$$- \tfrac{1}{2} \operatorname{tr}([Z^T \hat{W} ZD + I]^{-1} Z^T \hat{W} Z E_j$$

$$\times [Z^T \hat{W} ZD + I]^{-1} Z^T \hat{W} Z E_k),$$

where $E_k = D_{\nu_k}$ [for a detailed derivation see Section 1.8 of Geyer et al. (2012)]. 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 positive definite, because $W(a + M\alpha + Zb)$ is positive semidefinite by (2).

3. Square roots of variance components. It is part of the folklore of random effects models that introducing square roots of variance components avoids issues with zero variance components and with constrained optimization. Introduce new parameters by $\nu_j = \sigma_j^2$ and new random effects by $b = Ac$, where A is diagonal and $A^2 = D$. Then the objective function (7) becomes

$$(11) \quad \tilde{p}(\alpha, c, \sigma) = -l(a + M\alpha + ZAc) + \tfrac{1}{2} c^T c + \tfrac{1}{2} \log \det[Z^T \hat{W} Z A^2 + I].$$

There are now no constraints (the σ_j are allowed to be negative) and (11) is a continuous function of all variables (there is no discontinuity when $\sigma_j = 0$).

We find this change-of-parameter useful and use it to avoid constrained optimization [R package `aster`, Geyer (2012)]. However, it also causes problems.

First, it introduces spurious zeros of the first derivative of (11) that are not stationary points of (7). In fact, the partial of (11) with respect to σ_j is always zero when $\sigma_j = 0$ by symmetry. Thus, first derivatives of (11) cannot be used to test whether the minimum occurs when some variance component is zero. Since the issue of whether a variance component is zero is often of scientific interest, this is very problematic. We solve this problem by looking at first derivatives of (6) on

the original parameter scale (Section 4 below) and using the theory of constrained optimization.

Second, the formula (8) for observed Fisher information, although guaranteed to be positive definite if infinite precision arithmetic is used, is not so guaranteed if it is evaluated by the usual computer arithmetic (with 16 decimal digit precision). We found that the analog of (8) after the change of parameter from ν to σ was even more computationally unstable.

Thus, although (11) is useful for finding approximate maximum likelihood estimates, we find it problematic for calculating approximate observed Fisher information or for determining whether approximate maximum likelihood estimates of variance components are zero.

4. Theory of constrained optimization. In order to determine whether the minimizer of (7) occurs on the boundary of the parameter space where some variance component is zero, we need to use the theory of constrained optimization. Unfortunately, we cannot use the Karush–Kuhn–Tucker theory [Fletcher (1987), Section 9.1; Nocedal and Wright (1999), Section 12.2], which is familiar to some statisticians, because the constraint set is not determined by smooth inequality constraints. More advanced nonsmooth analysis [Rockafellar and Wets (2004)] does handle our problem, but is unfamiliar to most statisticians. Fortunately, for our analysis we can use a simplification of the latter theory based on the notion of directional derivatives. The technical report [Geyer et al. (2012)] uses the full theory from Rockafellar and Wets (2004), but the results are the same as those stated here in terms of directional derivatives.

4.1. *Incorporating constraints in the objective function.* The formula (7) makes sense when all variance components are positive (so D is invertible). Otherwise, it does not. As is common in nonsmooth analysis [Rockafellar and Wets (2004), Section 1A], we define the objective function to have the value $+\infty$ off of the constraint set. Since $+\infty$ can never minimize the objective function, this incorporates the constraints in the objective function. On the boundary of the constraint set (where some variance components are zero and the corresponding random effects are also zero) we extend the objective function by lower semicontinuity.

Since all but the middle term on the right-hand side of (7) 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. Define

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

Let $v_{k(i)}$ denote the variance of b_i, and let $\dim(b)$ denote the number of random effects. Then (7) can be rewritten

(13)

$$p(\alpha, b, v) = -l(a + M\alpha + Zb) + \frac{1}{2} \sum_{i=1}^{\dim(b)} h(b_i, v_{k(i)})$$

$$+ \frac{1}{2} \log \det[Z^T \hat{W} Z D + I],$$

where h is given by (12), provided all of the components of v are nonnegative. The proviso is necessary because the third term on the right-hand side is not defined for all values of v, only those such that the argument of the determinant is a positive definite matrix. Hence, we must separately define $p(\alpha, b, v) = +\infty$ whenever any component of v is negative.

4.2. *Directional derivatives.* A necessary condition for a local minimum of a smooth function is that the first derivative is zero (Fermat's rule). This works at points in the interior of the constraint set where (13) is differentiable. It does not work at points on the boundary. There we need what Rockafellar and Wets [(2004), Theorem 10.1] call *Fermat's rule, generalized:* a necessary condition for a local minimum is that all directional derivatives are nonnegative.

For any extended-real-valued function f on $\mathbb{R}^d$, the directional derivative of f at the point x in the direction w is defined by

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

At a point x where f is differentiable, we have $df(x)(w) = w^T f'(x)$, and the notion of directional derivatives gives no information that cannot be obtained from partial derivatives. It is only on the boundary where we need directional derivatives.

In the interior of the constraint set, where this function is smooth, ordinary calculus gives

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

where the notation on the left-hand side means the directional derivative of h at the point (b, v) in the direction (u, v). On the boundary of the constraint set, which consists of the single point $(0, 0)$, the directional derivatives are given by

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

4.3. *Applying the generalization of Fermat's rule.* This theory tells us nothing we did not already know about points in the interior of the constraint set. The only way we can have $df(x)(w) \geq 0$ for all vectors w is if $f'(x) = 0$. It is only at points on the boundary of the constraint set, where directional derivatives are the key.

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. At a local minimum we have

$$(14) \qquad\qquad\qquad p_\alpha(\alpha, b, v) = 0$$

and

$$(15) \qquad\qquad \begin{aligned} p_{v_j}(\alpha, b, v) &= 0, \qquad j \text{ such that } v_j > 0, \\ p_{b_i}(\alpha, b, v) &= 0, \qquad i \text{ such that } v_{k(i)} > 0 \end{aligned}$$

[Geyer et al. (2012), Section 1.10.4, gives details].

Thus, assuming that we are at a point (α, b, v) where (14) and (15) hold, and we do assume this throughout the rest of this section, $dp(\alpha, b, v)(s, u, v)$ actually involves only v_j and u_i such that $v_j = 0$ and $k(i) = j$. Define

$$(16) \qquad \bar{p}(\alpha, b, v) = -l(a + M\alpha + Zb) + \tfrac{1}{2} \log \det[Z^T \hat{W} ZD + I]$$

[the part of (13) consisting of the smooth terms]. Then

$$(17) \qquad \begin{aligned} &dp(\alpha, b, v)(s, u, v) \\ &\quad = \sum_{j \in J} \left[v_j \bar{p}_{v_j}(\alpha, b, v) + \sum_{i \in k^{-1}(j)} (u_i \bar{p}_{b_i}(\alpha, b, v) + h(u_i, v_j)) \right], \end{aligned}$$

where J is the set of j such that $v_j = 0$, where $k^{-1}(j)$ denotes the set of i such that $k(i) = j$, and where h is defined by (12). To check that we are at a local minimum, we need to show that (17) 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 (17) 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, v) will decrease the objective function (13).

So how do we find a descent direction? We want to minimize (17) considered as a function of u and v for fixed α, b and v. We can consider the terms of (17) for each j separately. If the minimum of

$$(18) \qquad v_j \bar{p}_{v_j}(\alpha, b, v) + \sum_{i \in k^{-1}(j)} (u_i \bar{p}_{b_i}(\alpha, b, v) + h(u_i, v_j))$$

over all vectors u and v is nonnegative, then the minimum is zero, because (18) has the value zero when $u = 0$ and $v = 0$. Thus, we can ignore this j in calculating the descent direction.

Since we are only interested in finding a descent direction, the length of the direction vector does not matter. Thus, we can do a constrained minimization of (18),

constraining (u, v) to lie in a ball. This is found by the well-known Karush–Kuhn–Tucker theory of constrained optimization [Fletcher (1987), Section 9.1; Nocedal and Wright (1999), Section 12.2] to be the minimum of the Lagrangian function

$$(19) \quad L(u, v) = \lambda v_j^2 + v_j \bar{p}_{v_j}(\alpha, b, v) + \sum_{i \in k^{-1}(j)} \left(\lambda u_i^2 + u_i \bar{p}_{b_i}(\alpha, b, v) + \frac{u_i^2}{v_j} \right),$$

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

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

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

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

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

plugging this back into the second gives

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

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

$$(20) \qquad \bar{p}_{v_j}(\alpha, b, v) - \frac{1}{4} \sum_{i \in k^{-1}(j)} \bar{p}_{b_i}(\alpha, b, v)^2.$$

So that gives us a test for a descent direction: we have a descent direction if and only if (20) is negative. Conversely, we appear to have $\hat{v}_j = 0$ if (20) is nonnegative.

5. *Raphanus sativus* example. We illustrate the use of this work with three examples, beginning with a study of the invasive California wild radish (*Raphanus sativus*) described by Ridley and Ellstrand (2010). For each individual, three re-

sponse variables are observed, connected by the following graphical model:

$$1 \xrightarrow{\text{Ber}} y_1 \xrightarrow{\text{0-Poi}} y_2 \xrightarrow{\text{Poi}} y_3$$

with y_1 being an indicator of whether any flowers were produced, y_2 being the count of the number of flowers produced, y_3 being the count of the number of fruits produced, the unconditional distribution of y_1 being Bernoulli, the conditional distribution of y_2 given y_1 being zero-truncated Poisson, and the conditional distribution of y_3 given y_2 being Poisson. (The combination of a Bernoulli arrow followed by a zero-truncated Poisson arrow gives a combined zero-inflated Poisson distribution, that is, the unconditional distribution of y_2 is zero-inflated Poisson.)

These data are found in the data set radish in the R package aster. They come from a designed experiment started with seeds collected from three large wild populations of northern, coastal California wild radish and three populations of southern, inland California wild radish. Thus, we have populations nested within region.

Plants were grown at two experimental sites, one northern, coastal California field site located at Point Reyes National Seashore and one southern, inland site located at the University of California Riverside Agricultural Experiment Station. Thus, we have blocks nested within site.

The issue of main scientific interest is the interaction of region and site, which is indicative of local adaptation when the pattern of mean values shows that each population has higher fitness in its home environment than in other environments. Testing significance of this interaction is complicated by the nesting of populations within region and blocks within site and the goal of scientists to account for variation due to these nested factors in evaluating effects of the higher factors.

The best surrogate of fitness in these data is the number of fruits produced. Thus, we form the "interaction" with the indicator of this component and all scientifically interesting predictors [see Section 5 of Geyer, Wagenius and Shaw (2007) or Section 4 of Geyer et al. (2012)].

The traditional way to deal with a situation like this is to treat the population effects as random (within region) and the block effects as random (within site). When we fit this model [see the technical report Geyer et al. (2012) for details], we obtained positive and statistically significantly greater than zero estimates of both variance components and an estimate 0.499 with standard error 0.012 for the fixed effect that is the scientifically important site-region interaction parameter.

Ridley and Ellstrand (2010) did not do a random effects aster analysis because it had not yet been invented. Nevertheless, the conclusions from their fixed effect aster analysis hold up. The main conclusion of interest is that there is evidence of local adaptation. This is indicated by the statistical significance of the fixed effect for region-site interaction together with the pattern of mean values for the different populations in the two sites, showing that populations growing near to their sampling locations had higher fitness than in the other location as found by Ridley and Ellstrand (2010).

The fact that random effects analysis and fixed effects analysis agree qualitatively on this one example does not, of course, imply that they would agree on all examples. In these data the region-site interaction is very large and almost any sensible statistical analysis would show it. When the interaction is not so large, the analysis done will make a difference.

The analysis reported above is based on the approximations derived in the theory section. We are using the log approximate integrated likelihood and its Hessian matrix to do likelihood inference. But what if its approximations are not valid? Section 6 of Geyer et al. (2012) does a parametric bootstrap of this analysis. It turns out that a 95% confidence interval for the parameter of interest (the region-site interaction) does not change much, but other aspects of the parametric bootstrap are interesting. Sampling distributions of the estimates of the variance components (as simulated by the parametric bootstrap) turn out to be highly nonnormal, and these estimators have bias that is a significant fraction of their standard errors.

6. *Avena barbata* example.   We use data on the slender wild oat (*Avena barbata*) described by Latta (2009) and contained in the data set `oats` in the R contributed package `aster`. For each individual, two response variables are observed, connected by the following graphical model:

$$1 \xrightarrow{\text{Ber}} y_1 \xrightarrow{\text{0-Poi}} y_2$$

with y_1 being an indicator of survival and y_2 being the count of the number of spikelets (compound flowers) produced, the unconditional distribution of y_1 being Bernoulli, and the conditional distribution of y_2 given y_1 being zero-truncated Poisson.

These data come from a designed experiment started with seeds collected in the 1980s in northern California of the xeric (found in drier regions) and mesic (found in less dry regions) ecotypes. The variable `Gen` is the ecotype ("X" or "M"). The variable `Fam` is the accession (nested within `Gen`). The variable `Site` is the site. The variable `Year` is the year (2003 to 2007). The experimental sites were at the Sierra Foothills Research and Extension Center (`Site == "SF"`), which is northeast of Sacramento on the east side of the Central Valley of California, and at the Hopland Research and Extension center (`Site = "Hop"`), which is in the California Coastal Ranges north of San Francisco. Hopland receives 30% more rainfall and has a less severe summer drought than the Sierra foothills. The best surrogate of fitness in these data is the number of spikelets produced. Thus, we form the "interaction" with the indicator of this component and all scientifically interesting predictors.

In the previous analysis [Latta (2009)] a linear mixed model was used, despite the response being highly nonnormal, because no better tool was available. Here we reanalyze these data using the same random effects structure in an aster model.

Effect	Type
Site	Fixed
Year	Random
Gen	Fixed
Fam	Random
Gen $*$ Site	Fixed
Gen $*$ Year	Random
Site $*$ Fam	Random
Year $*$ Fam	Random

We have only three fixed effects parameters because there are only two levels of Site and two levels of Gen. There are five variance components, one for each row of the table having random type.

All variance components are estimated to be significantly different from zero except for the Fam random effect, which is estimated to be exactly zero.

The results of the reanalysis agree qualitatively with the original analysis. Local adaptation, which would have been shown by a statistically significant site-ecotype (Gen $*$ Site) interaction, was not found in either analysis (for this interaction, the aster random effects analysis obtained the point estimate 0.091 and standard error 0.143). Moreover, the pattern of mean values was not consistent with local adaptation. Latta (2009) found that the mesic ecotype had higher fitness (survived and reproduced better) in all environments. This means that even if the site-ecotype interaction had been statistically significant, it would not have indicated local adaptation.

7. *Chamaecrista fasciculata* data. We use data on the partridge pea (*Chamaecrista fasciculata*) described by Etterson (2004a, 2004b) and Etterson and Shaw (2001) and contained in the data set chamae3 in the R contributed package aster. *C. fasciculata* grows in the Great Plains of North America from southern Minnesota to Mexico. Three focal populations were sampled in the following locations:

1. Kellog-Weaver Dunes, Wabasha County, Minnesota;
2. Konza Prairie, Riley County, Kansas;
3. Pontotoc Ridge, Pontotoc County, Oklahoma.

These sites are progressively more arid from north to south and also differ in other characteristics. Seed pods were collected from 200 plants in each of these three natural populations. From these, plants were grown and crosses were done; parent plants are indicated by the variables SIRE and DAM in the data set. The resulting seeds were germinated and established as seedlings in the greenhouse and then planted using a randomized block design [Etterson (2004b)] in three field sites:

"O" Robert S. Kerr Environmental Research Center, Ada, Oklahoma;
"K" Konza Prairie Research Natural Area, Manhatten, Kansas;
"M" University of Minnesota, St. Paul Minnesota.

The Oklahoma field site was 30 km northwest of the Oklahoma natural population; the Kansas field site was 5 km from the Kansas natural population; the Minnesota field site was 110 km northwest of the Minnesota natural population. For each individual, two response variables are observed, connected by the following graphical model:

$$1 \xrightarrow{\text{Ber}} y_1 \xrightarrow{\text{0-Poi}} y_2$$

with y_1 being an indicator of whether any fruits were produced, y_2 being the count of the number of fruits produced, the unconditional distribution of y_1 being Bernoulli, and the conditional distribution of y_2 given y_1 being zero-truncated Poisson.

We here consider a subset of data previously analyzed by nonaster methods by Etterson (2004a, 2004b) and Etterson and Shaw (2001) and by aster without allowing for random (genetic) effects by Shaw et al. (2008). Though seed counts were also observed, the complexity of the seed count data makes analysis difficult [Shaw et al. (2008)], so it does not serve as a good example. Thus, here we analyze only the pod number data, which does have straightforward aster analysis and serves as a better example, even though this makes our reanalysis not really comparable with the analysis in Etterson (2004b) which does use the seed counts. To aid design of future experiments, Shaw et al. (2008), page E43, explain two alternative experimental designs that permit straightforward aster analysis (including random effects aster models). Stanton-Geddes, Shaw and Tiffin (2012) used one of these designs.

Individuals descended from all three natural populations were planted in all three field sites, so these data can address local adaptation and previous analyses [Etterson (2004b), Discussion] did find local adaptation. But local adaptation is not the main point of interest for our analysis here. Instead we investigate sire and dam effects, which we treat as random effects, as did the previous conventional quantitative genetics analysis [Etterson (2004b)]. We focus on sire effects because in this experimental design sire effects are expected to correspond closely to pure breeding values (additive genetic effects) but dam effects confound additive with maternal and dominance effects.

Because the biology that leads to fitness may differ at different sites and in different populations, we did nine separate analyses, one for each population-site combination.

We found that the sire variance components for the Minnesota and Oklahoma natural population are not close to statistically significant at the Minnesota field site. All the other sire variance components are at least borderline statistically significant.

Our analysis produces not only estimates of the variance components but also estimates of the random effects (these are the penalized quasi-likelihood estimates b^* described in the theory sections). As a matter of purely statistical interest, we examined the Gaussianity of the random effects. They seemed to be normal (or at least not statistically significantly nonnormal by a Shapiro–Wilk test). We conjectured that this apparent normality was due to the estimation procedure, but this turned out not to be the case, since when we redid penalized quasi-likelihood estimates of the random effects using much smaller penalties than the maximum likelihood penalty, the random effects still seemed normal.

For interpretability, biologists want random effects mapped to the mean value parameter scale (rather than the canonical parameter scale where they originally are). To illustrate this, we mapped the sire effects for two population-site pairs to the mean value parameter scale, setting the dam effects to be zero (the middling value) and setting the block effect to be block 1 (each site was divided into blocks). Figure 1 shows these plots; for details of how they were done see Section 8.6 of Geyer et al. (2012). This figure was made using the default smoothing parameter
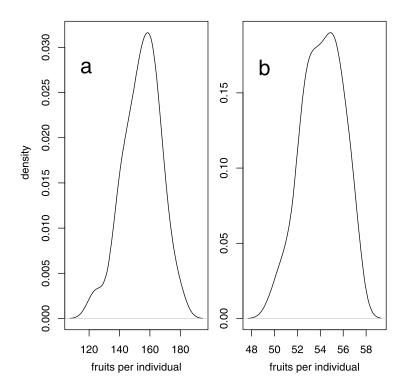

FIG. 1.   *Density plot of sire effects on the mean value parameter scale for an individual in block* 1 *having the various sire effects in the data. Panel* (a) *is the Kansas population in the Kansas field site. Panel* (b) *is the Kansas population in the Oklahoma field site.*

selection of the R function `density`. The apparent non-Gaussianity is not statistically significant [Shapiro–Wilk test, Geyer et al. (2012), Section 8.6].

Thus, the aster model can include random effects for parents and permit quantitative genetic inference for fitness variation.

8. Discussion. Our methods are founded on two approximations taken from Breslow and Clayton (1993). Our technical innovations are that we provide derivatives of the log approximate integrated likelihood (Section 2) and the test (20) for when variance components are zero, which is based on the theory of constrained optimization. Our methods work well when there are multiple variance components. Two examples had two variance components and one had five variance components. However, problems arise when there are thousands of random effects, and especially when there is one random effect per individual. Since quantitative genetics traditionally does have individual random effects as well as parental random effects, our methods are not fully comparable to traditional quantitative genetics.

Rutter et al. (2012) have already used aster models with random effects for an analysis of the effect of known spontaneous mutations on fitness in *Arabidopsis thaliana* grown in different environments.

Past experience [Sung and Geyer (2007)] with examples taken from the literature shows that log integrated likelihoods are often far from quadratic, even when no approximations are done, in which case asymptotics based on Fisher information are inaccurate. Thus, we recommend the parametric bootstrap here, as we do whenever there is doubt about the validity of asymptotics for parametric inference. We illustrate the parametric bootstrap for one of our examples. This need for doing a parametric bootstrap is another reason for preferring computationally efficient methods. In particular, it is incredibly time consuming to bootstrap Monte Carlo calculations if Monte Carlo run lengths are long enough for accurate calculation.

Breslow and Clayton (1993) introduced yet another approximation that is supposed to be analogous to restricted maximum likelihood (REML), but we did not use this. First, the analogy to REML is weak, and this method has no provable mathematical properties. Second, we do not see how this method extends to general exponential family models, such as aster models. Third, even in conventional linear mixed models, REML does not seem to be appropriate when the parameters of interest are fixed effects, which is often the case in biology and is the case in some of our examples.

REFERENCES

ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. Wiley, Hoboken, NJ. MR1990662

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families*. Wiley, Chichester. MR0489333

BOOTH, J. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **61** 265–285.

BRESLOW, N. E. and CLAYTON, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88** 9–25.

CROUCH, E. A. C. and SPIEGELMAN, D. (1990). The evaluation of integrals of the form $\int_{-\infty}^{+\infty} f(t) \exp(-t^2)\, dt$: Application to logistic-normal models. *J. Amer. Statist. Assoc.* **85** 464–469. MR1141749

ETTERSON, J. R. (2004a). Evolutionary potential of Chamaecrista fasciculata in relation to climate change. I. Clinal patterns of selection along an environmental gradient in the great plains. *Evolution* **58** 1446–1458.

ETTERSON, J. R. (2004b). Evolutionary potential of Chamaecrista fasciculata in relation to climate change. II. Genetic architecture of three populations reciprocally planted along an environmental gradient in the great plains. *Evolution* **58** 1459–1471.

ETTERSON, J. R. and SHAW, R. G. (2001). Constraint to adaptive evolution in response to global warming. *Science* **294** 151–154.

FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinburgh* **52** 399–433.

FLETCHER, R. (1987). *Practical Methods of Optimization*, 2nd ed. Wiley, Chichester. MR0955799

GEYER, C. J. (1994). On the convergence of Monte Carlo maximum likelihood calculations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **56** 261–274. MR1257812

GEYER, C. J. (2009). Likelihood inference in exponential families and directions of recession. *Electron. J. Stat.* **3** 259–289. MR2495839

GEYER, C. J. (2012). R package aster, version 0.8-19. Available at http://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 *Multivariate Statistics in Modern Statistical Analysis*: *A Festschrift for Morris L. Eaton* (G. Jones and X. Shen, eds.) **10** 1–24. IMS, Hayward, CA.

GEYER, C. J. and THOMPSON, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **54** 657–699. MR1185217

GEYER, C. J., WAGENIUS, S. and SHAW, R. G. (2007). Aster models for life history analysis. *Biometrika* **94** 415–426. MR2380569

GEYER, C. J., RIDLEY, C. E., LATTA, R. G., ETTERSON, J. R. and SHAW, R. G. (2012). Aster models with random effects via penalized likelihood. Technical Report 692, Univ. Minnesota School of Statistics. Available at http://purl.umn.edu/135870.

HUMMEL, R. M., HUNTER, D. R. and HANDCOCK, M. S. (2012). Improving simulation-based algorithms for fitting ERGMs. *J. Comput. Graph. Statist.* **21** 920–939. MR3005804

HUNTER, D. R., HANDCOCK, M. S., BUTTS, C. T., GOODREAU, S. M. and MORRIS, M. (2008). ergm: A Package to fit, simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* **24** nihpa54860.

LATTA, R. G. (2009). Testing for local adaptation in *Avena barbata*, a classic example of ecotypic divergence. *Molecular Ecology* **18** 3781–3791.

LE CAM, L. and YANG, G. L. (2000). *Asymptotics in Statistics*: *Some Basic Concepts*, 2nd ed. Springer, New York. MR1784901

McCULLOCH, C. E. (2003). *Generalized Linear Mixed Models*. *NSF-CBMS Regional Conference Series in Probability and Statistics* **7**. IMS, Beachwood, OH. MR1993816

NOCEDAL, J. and WRIGHT, S. J. (1999). *Numerical Optimization*. Springer, New York. MR1713114

OKABAYASHI, S. and GEYER, C. J. (2011). Gradient-based search for maximum likelihood in exponential families. *Electron. J. Stat.* **6** 123–147. MR2879674

PENTTINEN, A. (1984). Modelling interations in spatial point patterns: Parameter estimation by the maximum likelihood method. Jyväskylä Studies in Computer Science, Economics and Statistics No. 7, Univ. Jyväskylä.

RIDLEY, C. E. and ELLSTRAND, N. C. (2010). Rapid evolution of morphology and adaptive life history in the invasive California wild radish (*Raphanus sativus*) and the implications for management. *Evolutionary Applications* **3** 64–76.

ROCKAFELLAR, R. T. and WETS, R. J. B. (2004). *Variational Analysis*, 2nd corrected printing. Springer, Berlin. MR1491362

RUTTER, M. T., ROLES, A., CONNER, J. K., SHAW, R. G., SHAW, F. H., SCHNEEBERGER, K., OSSOWSKI, S., WEIGEL, D. and FENSTER, C. B. (2012). Fitness of Arabidopsis thaliana mutation accumulation lines whose spontaneous mutations are known. *Evolution* **66** 2335–2339.

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, 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, R. G., GEYER, C. J., WAGENIUS, S., HANGELBROEK, H. H. and ETTERSON, J. R. (2008). Unifying life-history analyses for inference of fitness and population growth. *Am. Nat.* **172** E35–E47.

STANTON-GEDDES, J., SHAW, R. G. and TIFFIN, P. (2012). Interactions between soil habitat and geographic range location affect plant fitness. *PLoS ONE* **7** e36015.

SUNG, Y. J. and GEYER, C. J. (2007). Monte Carlo likelihood inference for missing data models. *Ann. Statist.* **35** 990–1011. MR2341695

R CORE TEAM (2012). R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. Available at http://www.R-project.org/.

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.

TRAVISANO, M. and SHAW, R. G. (2013). Lost in the map. *Evolution* **67** 305–314.

C. J. GEYER
SCHOOL OF STATISTICS
UNIVERSITY OF MINNESOTA
313 FORD HALL
224 CHURCH ST. SE
MINNEAPOLIS, MINNESOTA 55455
USA
E-MAIL: geyer@umn.edu

R. G. LATTA
DEPARTMENT OF BIOLOGY
DALHOUSIE UNIVERSITY
LIFE SCIENCE CENTRE
1355 OXFORD STREET
PO BOX 15000
HALIFAX, NOVA SCOTIA B3H 4R2
CANADA
E-MAIL: robert.latta@dal.ca

C. E. RIDLEY
US ENVIRONMENTAL PROTECTION AGENCY
1200 PENNSYLVANIA AVENUE, NW
WILLIAM JEFFERSON CLINTON
    FEDERAL BUILDING, MAIL CODE: 8623P
WASHINGTON, DISTRICT OF COLUMBIA 20460
USA
E-MAIL: ridley.caroline@epa.gov

J. R. ETTERSON
DEPARTMENT OF BIOLOGY
UNIVERSITY OF MINNESOTA DULUTH
153B SWENSON SCIENCE BUILDING
DULUTH, MINNESOTA 55812
USA
E-MAIL: jetterso@d.umn.edu

R. G. SHAW
DEPARTMENT OF ECOLOGY, EVOLUTION
    AND BEHAVIOR
UNIVERSITY OF MINNESOTA
100 ECOLOGY
1987 UPPER BUFORD CIRCLE
ST. PAUL, MINNESOTA 55108
USA
E-MAIL: shawx016@umn.edu