Statistica Sinica **11**(2001), 97-120

MAXIMUM POSTERIOR ESTIMATION OF RANDOM EFFECTS IN GENERALIZED LINEAR MIXED MODELS

Jiming Jiang, Haomiao Jia and Hegang Chen

Case Western Reserve University, University of Tennessee and University of Minnesota

Abstract: Given a vector of observations and a vector of dispersion parameters (variance components), the fixed and random effects in a generalized linear mixed model are estimated by maximizing the posterior density. Although such estimates of the fixed and random effects depend on the (unknown) vector of variance components, we demonstrate both numerically and theoretically that in certain large sample situations the consistency of a restricted version of these estimates is not affected by variance components at which they are computed. The method is applied to a problem of small area estimation using data from a sample survey.

Key words and phrases: Consistency, GLMM, maximum posterior, small area estimation.

1. Introduction

Recently inference about generalized linear mixed models (GLMM) has received considerable attention. These models take into account the fact that in many practical problems responses are both discrete and correlated. One of the early applications of GLMM was to salamander mating data, McCullagh and Nelder (1989, §14.5). For applications of GLMM in medical research, sample surveys and other fields, see Breslow and Clayton (1993), Lee and Nelder (1996), and Malec, Sedransk, Moriarity and LeClere (1997).

For years the major difficulty in inference about GLMM has been computational. Consider, for example, a logit model with crossed random effects: $p_{ij} = P(y_{ij} = 1|u, v)$,

$$\text{logit}(p_{ij}) = \mu + u_i + v_j \ , \tag{1.1}$$

$i = 1, \ldots, m_1$, $j = 1, \ldots, m_2$, where u_i's and v_j's are independent random variables such that $u_i \sim N(0, \sigma^2)$, $v_j \sim N(0, \tau^2)$. The log-likelihood for estimating μ, σ^2 and τ^2 has the form

$$\text{constant} - \frac{m_1}{2}\log\sigma^2 - \frac{m_2}{2}\log\tau^2 + \mu y_{..}$$

$$+\log\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\Big(\prod_{i=1}^{m_1}\prod_{j=1}^{m_2}(1+\exp(\mu+u_i+v_j))^{-1}\Big)$$
$$\cdot\exp\Big(\sum_{i=1}^{m_1}u_iy_{i\cdot}-\frac{1}{2\sigma^2}\sum_{i=1}^{m_1}u_i^2+\sum_{j=1}^{m_2}v_jy_{\cdot j}-\frac{1}{2\tau^2}\sum_{j=1}^{m_2}v_j^2\Big)\prod_{i=1}^{m_1}du_i\prod_{j=1}^{m_2}dv_j\ , \qquad (1.2)$$

where $y_{\cdot\cdot}=\sum_{i=1}^{m_1}\sum_{j=1}^{m_2}y_{ij}$, $y_{i\cdot}=\sum_{j=1}^{m_2}y_{ij}$, and $y_{\cdot j}=\sum_{i=1}^{m_1}y_{ij}$. If $m_1=m_2=40$, as in the salamander mating problem mentioned above, then the integral in (1.2) will be 80 dimensional. There is, of course, no analytic form for it, and it is not feasible to evaluate such a high-dimensional integral by numerical integration. Furthermore, the integrand in (1.2) involves a product of 1600 terms with each term less than one. Such a product would be numerically zero, creating difficulties for Monte-Carlo integration even with advanced techniques such as importance sampling.

To overcome such difficulties, alternative methods have been proposed. One approach is based on estimation of the random effects via maximization of the joint density of the observations and random effects. The original idea goes back to Henderson (1950) and is known as best linear unbiased prediction, or BLUP (see Robinson (1991)). Such an approach avoids integration altogether and therefore is computationally feasible. It has been used with different adjustments and approximations in making inference about GLMM (e.g., Schall (1991), Breslow and Clayton (1993), McGilchrist (1994), Kuk (1995), and Lin and Breslow (1996)). In a recent paper by Lee and Nelder (1996), the authors have given the name "maximum hierarchical likelihood" for the extension of Henderson's approach to nonlinear models. In the following we take another look at Henderson's approach, and hence give a different justification of the BLUP in nonlinear situations.

Suppose y is a vector of observations, γ a vector of unobservable "random variables", and θ a vector of parameters. For example, γ may be a vector of random effects, or a vector of fixed effects and random effects. Correspondingly, θ may represent a vector of fixed effects and variance components, or simply variance components. Let $f(y,\gamma|\theta)$ be the joint density of y and γ given θ. Then

$$f(y,\gamma|\theta)=f(y|\theta)f(\gamma|y,\theta)\ . \qquad (1.3)$$

The first factor on the RHS of (1.3) is the likelihood obtained by integrating out γ, while the second factor represents the posterior density of γ given y and θ. Henderson's original approach was to find $\hat{\gamma}=\hat{\gamma}(y,\theta)$ to maximize $f(y,\gamma|\theta)$. From (1.3) we see this is equivalent to maximizing $f(\gamma|y,\theta)$. Thus the BLUP $\hat{\gamma}$ may be regarded as the vector that maximizes the posterior density of γ given y and θ. Note that, although under the linear mixed models $\hat{\gamma}$ corresponds to the

best linear unbiased predictor (e.g., Searle, Casella and McCulloch (1992, §7.4)), there is indeed no special reason to maintain the term BLUP in the nonlinear situation. Therefore we call $\hat{\gamma}$ the *maximum posterior* estimate (MPE) of γ.

It should be pointed out that in many cases the random effects are treated as nuisance parameters, while the fixed parameters, such as the variance components, are of main interest. However, there are also many cases in which the estimation or prediction of the random effects is of interest. Robinson (1991) gives an excellent review on the estimation of random effects, with examples and applications. In particular, the estimation of random effects is important to small-area estimation (see Ghosh and Rao (1994) for a review).

Given y and θ, $\hat{\gamma}$ is usually obtained by solving the system of equations

$$\frac{\partial}{\partial \gamma_i} \log f(y, \gamma|\theta) = 0, \quad i = 1, \ldots, n, \tag{1.4}$$

where n is the dimension of γ. Although in practice the number of fixed effects is often fairly small, the number of random effects can be quite large. For example, in the salamander mating problem mentioned earlier the number of random effects corresponding to the female and male animals is 80; in the problem of National Health Interview Survey discussed in Malec *et al* (1997) the number of random effects corresponding to small areas is about 600. It is well-known that standard methods of solving nonlinear systems, such as Newton-Raphson, may be inefficient and extremely slow when the dimension of the solution is high. Even in the linear case, directly solving the BLUP equation may involve inverting a large matrix and this may be computationally burdensome. Jiang (2000) proposed a Gauss-Seidel type recursive algorithm which effectively solves (1.4). It is shown that the algorithm converges in virtually all typical situations of GLMM.

Although the MPE has been widely used in making inference about GLMM, its theoretical properties are mostly unknown except in the linear case. In the case of linear mixed models, Jiang (1998a) considers asymptotic properties of the empirical best linear unbiased estimate (BLUE) and BLUP. As noted earlier, the MPEs considered in this paper are natural generalization of the BLUE and BLUP to nonlinear mixed models. Under the assumption that the unknown variance components are estimated by restricted maximum likelihood (REML) estimates (e.g., Searle, Casella and McCulloch (1992, §6)), Jiang (1998a) proves the convergence of the empirical distribution of the empirical BLUPs to the true distribution of the random effects. There is some discussion in Lee and Nelder (1996) about the asymptotic properties of the MPE for the fixed effects. It was questioned whether the asymptotics would apply when insufficient data was available for estimating the individual random effects (Breslow (1996), Clayton (1996)). In fact, even with "sufficient data" the problem is more complicated

than it appears. The MPE depends on the vector θ. In the literature, whenever $\hat{\gamma}$ is treated as an estimate, θ is either assumed known or replaced by a consistent estimate. For example in Jiang (1998a), θ is a vector of variance components, and is asymptotically correctly specified because the REML estimates are consistent (Jiang (1996)). Will the choice of θ affect the asymptotic behavior of $\hat{\gamma}$? The main goal of this paper is to answer the question from a consistency point of view. Such results are not available for nonlinear (mixed) models even with correctly specified (i.e., known) θ. See further discussion in the next section. We show that, given sufficient information about the random effects, a restricted version of the MPE is consistent no matter at which θ they are evaluated. This may sound surprising, but it obviously has practical impact. In practice, θ may consist of unknown variance components which (in nonlinear models) are difficult to estimate. Furthermore, the computational difficulty in estimating the variance components increases with the sample size. Now the good thing is that in some large sample situations one does not have to worry too much about θ if the main interest is to estimate γ, since it will make very little difference whether $\hat{\gamma}$ is evaluated at a consistent estimate of θ, or at any reasonable guess of it.

The paper is organized as follows. In Section 2 we introduce GLMM and give some examples. The main result about asymptotic behavior of the MPE is stated and explained in Section 3, with proofs given in Appendix. More examples are considered in Section 4. In Section 5 we give some remarks about estimation of the variance components based on the MPE. Finally, in Section 6, we apply the method to a problem of small area estimation using data from the Behavioral Risk Factor Surveillance System (BRFSS).

2. Generalized Linear Mixed Models, Examples, and Notation

Suppose that given a vector α of random effects, the observations $y_1, \ldots, y_N$ are independent with conditional density

$$f(y_i|\alpha) = \exp\left\{\frac{y_i\eta_i - b_i(\eta_i)}{a_i(\phi)} + c_i(y_i, \phi)\right\}, \tag{2.1}$$

$i = 1, \ldots, N$, where $b_i(\cdot)$s and $c_i(\cdot, \cdot)$s are specific functions corresponding to the type of exponential family, ϕ is a vector of dispersion parameters, and the $a_i(\cdot)$s are some functions of weights. Quite often the functions $b_i(\cdot)$ do not depend on i. Furthermore, assume that the vector $\eta = (\eta_i)_{1\le i\le N}$ has the expression

$$\eta = X\beta + Z\alpha\,, \tag{2.2}$$

where $X = (x_{ij})_{1\le i\le N, 1\le j\le p}, Z = (z_{ik})_{1\le i\le N, 1\le k\le m}$ are known design matrices, and $\beta = (\beta_j)_{1\le j\le p}$ is a vector of unknown constants (the fixed effects). WLOG, we assume that $\text{rank}(X) = p$.

We assume that the random effects α satisfy

$$E(\alpha) = 0\ , \qquad \mathrm{Var}(\alpha) \ = \ D\ , \tag{2.3}$$

where $D = D(\theta)$ depends on a vector θ of variance components. In deriving the MPE we need,

$$\alpha \sim N(0, D)\ . \tag{2.4}$$

Under such an assumption, the log-joint density of y and α is given by

$$l_J = c(y;\theta,\phi) + \sum_{i=1}^{N} \frac{y_i\eta_i - b_i(\eta_i)}{a_i(\phi)} - \frac{1}{2}\alpha^t D^{-1}\alpha\ , \tag{2.5}$$

where $c(y;\theta,\phi) = -(m/2)\log 2\pi + \sum_{i=1}^{N} c_i(y_i,\phi) - (1/2)\log\det(D)$ ($\det(D)$ is the determinant of D). Let

$$l(\beta,\alpha) = l(\beta,\alpha|y,\theta,\phi) \ = \ \sum_{i=1}^{N} \frac{y_i\eta_i - b_i(\eta_i)}{a_i(\phi)} - \frac{1}{2}\alpha^t D^{-1}\alpha\ . \tag{2.6}$$

We consider the combined vector (β,α) as the vector of "random variables" γ in (1.3). Then maximizing l_J over (β,α) is equivalent to maximizing l over (β,α). If we let $\xi = (a_i^{-1}(\phi)(y_i - b_i'(\eta_i)))_{1\le i\le N}$, then the maximizer $(\hat\beta,\hat\alpha)$, where $\hat\beta = \hat\beta(y,\theta,\phi)$ and $\hat\alpha = \hat\alpha(y,\theta,\phi)$, satisfies

$$\frac{\partial l}{\partial \beta} = X^t\xi \ = \ 0, \tag{2.7}$$

$$\frac{\partial l}{\partial \alpha} = Z^t\xi - D^{-1}\alpha \ = \ 0. \tag{2.8}$$

According to properties of the exponential family, we have

$$E(y_i|\alpha) \ = \ b_i'(\eta_i)\ , \quad \mathrm{var}(y_i|\alpha) \ = \ a_i(\phi)b_i''(\eta_i)\ . \tag{2.9}$$

We assume the model is nondegenerate in the sense that $b_i''(\cdot) > 0,\ 1 \le i \le N$.

Lemma 2.1. *If the solution to* (2.7) *and* (2.8) *exists, it is unique and is equal to* $(\hat\beta,\hat\alpha)$*, the MPE.*

This follows directly from the strict concavity of l (see Haberman (1977)).

One difficulty with inference about random effects is that the number of the random effects in a GLMM is typically increasing with the sample size. Large sample performance of estimates of the fixed parameters, not the random effects, has been considered. Jiang (1998b) proposed a method of simulated moments (MSM) approach to estimation of the fixed effects and variance components in

a GLMM, and proved consistency of the MSM estimates. However, the MSM cannot deal with random effects. In linear mixed models, Jiang (1998a) studied large sample properties of the empirical BLUP (see discussion in Section 1). In the case of (fixed effects) generalized linear models, Haberman (1977) considered large sample properties of the maximum likelihood estimates (MLE) when the number of parameters increases with the sample size. Similar problems were also studied by Portnoy (1988). It is important to note that in problems involving random effects, there is often insufficient information in the data for asymptotically consistent estimation of all individual random effects. For example, in our Example 4.2, it is not true that $n_i \to \infty$ for every i. In such cases, unlike Haberman (1977) and Portnoy (1988), one cannot expect the MPE of every individual random effect to be consistent (e.g., in Example 4.2, $\tilde{\alpha}_i - \alpha_i \xrightarrow{P} 0$ may not hold for every i). However, in many cases, it is true that $m/N \to 0$, i.e., the total number of random effects is small compared with the total sample size. For example, in Malec *et al* (1997), the number of small areas, m, is about 600, while the total sample size, N, is about $120,000$; in the BRFSS data considered in Section 6, the number of HSA's is 118, while the total sample size is $29,505$. Therefore, it is reasonable to expect the MPE of the fixed effects to be consistent, and the MPE of the random effects to be consistent in some overall sense. The large sample performance of the MPE will be studied in Section 3. First consider some simple examples to see what to expect.

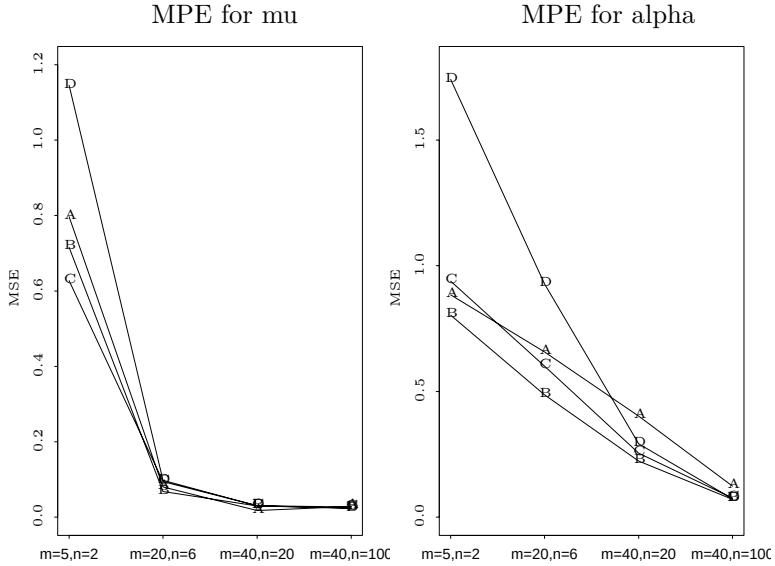


Figure 2.1.

Example 2.1. Consider the linear mixed model $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, where the α's are independent random effects with $E\alpha_i = 0$ and $\mathrm{var}(\alpha_i) = \sigma_\alpha^2$, the ϵ's are independent errors with $E\epsilon_{ij} = 0$ and $\mathrm{var}(\epsilon_{ij}) = \sigma_\epsilon^2$. The MPE for μ and α are the BLUE and BLUP, given respectively by

$$\hat{\mu} = \bar{y}_{\cdot\cdot} = \mu + \bar{\alpha}_{\cdot} + \bar{\epsilon}_{\cdot\cdot} ,$$

$$\hat{\alpha}_i = \frac{\lambda n}{1+\lambda n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}) = \frac{\lambda n}{1+\lambda n}(\alpha_i - \bar{\alpha}_{\cdot} + \bar{\epsilon}_{i\cdot} - \bar{\epsilon}_{\cdot\cdot}) , \quad i = 1, \ldots, m ,$$

where $\lambda = \sigma_\alpha^2/\sigma_\epsilon^2$. It is clear that for any $\lambda > 0$, as both m and $n \to \infty$, $\hat{\mu}$ and $\hat{\alpha}_i, 1 \le i \le m$, converge to the true μ and α_i.

Example 2.2. Suppose that y_{ij} is binary with $\mathrm{logit}(P(y_{ij} = 1|\alpha)) = \mu + \alpha_i$, $1 \le i \le m$, $1 \le j \le n$, where the α's are independent and $\sim N(0, \sigma^2)$. Figure 2.1 shows the simulated MSE, i.e., $E(\hat{\mu} - \mu_0)^2$ and $E(\frac{1}{m}\sum_{i=1}^m (\hat{\alpha}_i - \alpha_{0i})^2)$ under different sample sizes, where $\mu_0 = 0$, and the true random effects α_{0i}'s are generated from $N(0, \sigma_0^2)$ with $\sigma_0^2 = 1$. The MPE $\hat{\mu}$ and $\hat{\alpha}_i$'s are computed at $\sigma^2 =$ 0.2 (A), 1.0 (B), 4.0 (C), and 9.0 (D). Each MSE is based on 100 simulations. Note that as long as m and n are reasonably large, there is not much difference whether $\hat{\mu}$ and $\hat{\alpha}$ are computed at the right (B) or wrong (A, C, D) σ^2! It should be pointed out that for the MPE in the above examples to be consistent it is necessary that both m and n go to infinity. For example, in Example 2.1 if $m \to \infty$ but n remains bounded, the BLUP will not be consistent even if evaluated at the true variance components. It will be seen that worse things happen in Example 2.2: if $m \to \infty$ but n remains bounded, the MPE of μ is not consistent even if evaluated at the true σ^2 (see Example 4.3).

Notation

Let $v_1, \ldots, v_n$ be column vectors and $G_1, \ldots, G_n$ be matrices. We use the symbol $(v_1, \ldots, v_n)$ for the vector $(v_1^t \cdots v_n^t)^t$. To avoid confusion, a row vector will be written as $(\lambda_1 \cdots \lambda_n)$, i.e., without commas between the components. Let $\mathrm{diag}(G_1, \ldots, G_n)$ be the block-diagonal matrix with G_i being its ith diagonal block, $1 \le i \le n$. We use I_n (1_n) to represent the n-dimensional identity matrix (vector of 1's). Let $v = (v_l)_{1 \le l \le n}$ be a vector and $G = (g_{ij})_{1 \le i \le m, 1 \le j \le n}$ be a matrix. Define $|v| = (\sum_{l=1}^n v_l^2)^{1/2}$, $\|v\| = \max_{1 \le l \le n} |v_l|$; $\lambda_{\max}(G)$ $(\lambda_{\min}(G))$ = the largest (smallest) eigenvalue of G, $|G| = (\mathrm{tr}(G^t G))^{1/2}$, $\|G\| = (\lambda_{\max}(G^t G))^{1/2}$. Let $\mathcal{V}$ be a vector space. Define $\lambda_{\min}(G)|_{\mathcal{V}} = \inf_{v \in \mathcal{V} \setminus \{0\}} \{v^t G v / v^t v\}$.

Define P_1^*: $\alpha \to \alpha^*$ by $\alpha^* = \mathrm{argmin}\{a^t D^{-1} a : Za = Xb + Z\alpha \text{ for some } b\}$, and P^*: $(\beta, \alpha) \to (\beta^*, \alpha^*)$ by $\alpha^* = P_1^* \alpha$ and $\beta^* = \beta + (X^t X)^{-1} X^t Z(\alpha - \alpha^*)$.

The following notation will be used throughout: $H = (X\ Z)^t (X\ Z)$; $X_j =$ the jth column of X $(1 \le j \le p)$ and $Z_k =$ the kth column of Z $(1 \le k \le m)$;

$W = \text{diag}(|X_1|, \ldots, |X_p|, |Z_1|, \ldots, |Z_m|)$. Let $S = P^* R^{p+m}$, the range of P^*, and $WS = \{W(\beta, \alpha) : (\beta, \alpha) \in S\}$. Denote $r = \text{rank}(X\ Z)$, and $s = p + m - r$.

Let Θ and Φ be the parameter space for θ and ϕ, respectively; $\theta_0, \beta_0 = (\beta_{0j})_{1 \le j \le p}$, and $\alpha_0 = (\alpha_{0k})_{1 \le k \le m}$ be the true vectors θ, β, and α, respectively. Write $D_0 = D(\theta_0), \eta_0 = (\eta_{0i})_{1 \le i \le N} = X\beta_0 + Z\alpha_0$. Let $\{a(N)\}$ and $\{b(N)\}$ be sequences of numbers such that $a(N)$, $b(N) \to \infty$, and

$$R_N = \{(\beta, \alpha) : |\eta_i| \le M_{N,i}, 1 \le i \le N\}\ , \tag{2.10}$$

where $M_{N,i} = b(N) \sum_{j=1}^{p} |x_{ij}| + a(N) \sum_{k=1}^{m} |z_{ik}|$. Let $\Theta_0 \subset \{\theta \in \Theta : \liminf \lambda_{\min}(D) > 0\ ,\ \limsup \lambda_{\max}(D) < \infty\}$, $\Phi_0 \subset \{\phi \in \Phi : \limsup \max_{1 \le i \le N} \{a_i^{-1}(\phi) Eb_i''(\eta_{0i})\} < \infty\}$, and

$$\rho_N = \min_{1 \le i \le N} \Big\{ \inf_{|\lambda| \le M_{N,i}} b_i''(\lambda)/a_i(\phi) \Big\}\ . \tag{2.11}$$

3. Large Sample Performance of the MPE

In this section, we specify the conditions under which one can expect the MPE to be asymptotically accurate, and interpret these conditions. For the most part, the assumptions needed for the asymptotics may be classified into three groups.

Assumption 1. The design matrix Z for the random effects satisfies

$$\min_{1 \le k \le m} |Z_k| \ge c_0\ , \tag{3.1}$$

where c_0 is a positive constant.

Note that this condition is satisfied if the model has *standard design matrices* for the random effects in the following sense: η in (2.2) can be expressed as

$$\eta = X\beta + Z_{(1)}\alpha_{(1)} + \cdots + Z_{(q)}\alpha_{(q)}$$

for some fixed number q, where each $Z_{(u)}$, $1 \le u \le q$, consists only of 0's and 1's and there is exactly one 1 in each row and at least one 1 in each column. Let Z_{uv} be the vth column of $Z_{(u)}$ and $n_{uv} = |Z_{uv}|^2$, the number of appearances of the vth component of $\alpha_{(u)}$, $1 \le v \le m_u$, $1 \le u \le q$. Note that in this case, $\alpha = (\alpha_{(1)}, \ldots, \alpha_{(q)})$. It follows that, under standard design, $|Z_k| \ge 1$.

Assumption 2. *Asymptotic identifiability.* This means that the following three conditions hold:

$$\liminf \lambda_{\min}(W^{-1} H W^{-1})|_{WS}\ >\ 0\ , \tag{3.2}$$

$$(s/N)\|Z\|^2 \longrightarrow 0\ , \tag{3.3}$$

$$0 \ < \ \liminf \lambda_{\min}(X^tX)/N \ \leq \ \limsup \lambda_{\max}(X^tX)/N \ < \ \infty\ . \tag{3.4}$$

We now give some interpretations of these conditions. The basic idea of our proof for the asymptotic accuracy of the MPE seems quite natural. Let $\lambda(\eta)$ be the first term on the RHS of (2.6), and (β_*, α_*) be some point close to (β_0, α_0). Consider

$$d = l(\beta,\alpha) - l(\beta_*,\alpha_*) = \lambda(\eta) - \lambda(\eta_*) + \frac{1}{2}(\alpha_*^t D^{-1}\alpha_* - \alpha^t D^{-1}\alpha) = d_1 + d_2. \tag{3.5}$$

Note that $d_1 = \lambda(\eta) - \lambda(\eta_*)$ does not depend on θ. If, for a certain large sample, d_1 is the dominant factor for maximizing l, the asymptotic behavior of the MPE would not be affected by θ. However it is not true, even within a bounded range of α, that d_1 will dominate d_2. The reason for this is simple: the vector (β, α) may not be identifiable by η, there may be many vectors such that $\eta = X\beta + Z\alpha = X\beta_* + Z\alpha_* = \eta_*$. To solve this problem, we recall notation. Since α^* is the minimizer of the norm $\|a\|_D = \{a^tD^{-1}a\}^{1/2}$ in the hyperplane $\alpha + \{a : Za = Xb \text{ for some } b\}$, P_1^* is an orthogonal projection in the norm $\|a\|_D$ and $D^{-1/2}\alpha^*$ is the projection of $D^{-1/2}\alpha$ in the Euclidean norm $|a|$. Also, since X is of full rank, $X\beta^* + Z\alpha^* = X\beta + Z\alpha$ and (β^*, α^*) minimizes $\|a\|_D$ among all (b, a) satisfying $Xb+Za = X\beta+Z\alpha$, and we have $\text{rank}(P^*) = \text{rank}(X, Z) = r$ and $\text{rank}(P_1^*) = r - p$. It follows that $l(\beta^*, \alpha^*) \geq l(\beta, \alpha)$ with equality iff $(\beta, \alpha) \in S$. We thus conclude the following.

Lemma 3.1. *The MPE* $(\hat\beta, \hat\alpha) \in S$, *therefore* $\sup l(\beta, \alpha) = \sup_{(\beta,\alpha)\in S} l(\beta, \alpha)$.

From Lemma 3.1 we see that maximizing l is equivalent to maximizing l over S and, restricted on S, (β, α) is uniquely determined by η. This means $\lambda_{\min}(H)|_S > 0$. Asymptotically, it is more appropriate to consider a normalized limiting version of the above, and a natural set of normalizing constants are the diagonal elements of H. Therefore, we assume (3.2).

Note 1. (3.2) indicates that the eigenvalues of H jump from zero to a large number. This is typical when considering asymptotics in a mixed model. For example, consider Example 2.2. It is easy to show that

$$H = n\begin{pmatrix} m & 1_m^t \\ 1_m & I_m \end{pmatrix},$$

thus the eigenvalues of H are $0, \underbrace{n, \ldots, n}_{m-1}, n(m+1)$. If both m and $n \to \infty$, the eigenvalues of H jump from zero to a large number. Even if n is fixed, say, $n = 2$,

but $m \to \infty$, the eigenvalues of H still jump from $n = 2$ to $2(m + 1)$. Therefore, one needs to normalize H, and this is (3.2).

Note 2. We consider consistency of the MPE as $N \to \infty$, and such a result holds if it holds for each sequence with N strictly monotone. Therefore, WLOG, we regard the matrices X and Z as depending on N, and the numbers p and m as functions of N. Limiting processes are understood as asymptotic in N.

To see what (3.3) means, suppose that the model has standard design matrices. Note that $Z_{(u)}^t Z_{(u)} = \text{diag}(n_{uv},\ 1 \le v \le m_u)$. It follows that $\|Z\|^2 \le \sum_{u=1}^q \lambda_{\max}(Z_{(u)} Z_{(u)}^t) \le q \max_{1\le u\le q} \max_{1\le v\le m_u} n_{uv}$. Thus, (3.3) is satisfied provided $(s/N) \max_{1\le u\le q} \max_{1\le v\le m_u} n_{uv} \to 0$. Furthermore, suppose that $Z_{(u)}$'s are *balanced* in the sense that $n_{u1} = \cdots = n_{um_u}$. Then (3.3) is satisfied provided $s/\min_{1\le u\le q} m_u \to 0$. This indicates that the matrix $(X\ Z)$ is asymptotically nearly of full rank. Note that quite often s does not grow with N.

Finally we require that the fixed effects β be asymptotically identifiable, which is necessary for the consistency of the MPE. For fixed sample size, the identifiability of β is equivalent to $\lambda_{\min}(X^t X) > 0$. Note that even though we have assumed $\text{rank}(X) = p$, the matrix X may still be "asymptotically not of full rank". For example, consider $X^t = \begin{pmatrix} 1 & 1 \cdots 1 \\ 1 - 1/N & 1 \cdots 1 \end{pmatrix}_{2\times N}$. In this case $\beta = (\beta_1, \beta_2)$ is asymptotically not identifiable. A further observation shows in this example that $\lambda_{\min}(X^t X)/N \to 0$. This suggests, as before, that one should consider a normalized limiting version of $\lambda_{\min}(X^t X) > 0$. If the fixed effects include an intercept, then the first column of X is 1_N and hence the first diagonal element of $X^t X$ is N. Therefore, Assumption (3.4) is a natural requirement for β to be asymptotically identifiable.

Note 3. In linear regression identifiability in the weakest sense means that $\lambda_{\min}(X^t X) \to \infty$ (e.g., Lai and Wei (1982)). The reason we require stronger conditions is that one has to consider identifiability of both fixed and random effects (while X only corresponds to the fixed part), and the random effects cannot be treated the same way. Furthermore, in a mixed model there may be other quantities that go to infinity in addition to the sample size N. For example m, the number of random effects, is assumed to go to ∞. Thus, one needs to specify the rates at which different quantities go to ∞. In many cases, (3.4) gives the right rate at which $\lambda_{\min}(X^t X) \to \infty$. See the discussion above and Example 4.2 in the sequel.

Assumption 3. *The number of effects grows at a slower rate than the sample size.* This means that

$$(p + m)/N = o(\rho_N^2)\ , \tag{3.6}$$

and $\rho_N \to 0$ as $N \to \infty$, where ρ_N is as in (2.11).

Note that we do not assume that p, the number of the fixed effects, is fixed or bounded. Of course the number m typically converges to infinity as $N \to \infty$. Such conditions as (3.6) are considered necessary when the number of parameters to be estimated is growing with the sample size (e.g., Portnoy (1988)). To see what (3.6) means, suppose first that the fixed and random effects are bounded, and $b_i(\cdot) = b(\cdot)$. Then there is a positive lower bound for the $b_i''(\eta_i)$'s. Therefore, assuming the $a_i(\phi)$'s are bounded, $\rho_N \geq \rho > 0$, thus (3.6) simply means that $(p+m)/N \to 0$. It follows that the coefficients of the quadratic terms in the Talor expansion of $l(\beta, \alpha)$ (see the proof of Theorem 3.1) are bounded away from 0. Thus, in a neighborhood near the true (β, α), $l(\beta, \alpha)$ is uniformly strictly concave. This is a key condition for the consistency of the estimates. Now, suppose the effects are not bounded but $\|\beta\| \leq b(N)$, $\|\alpha\| \leq a(N)$, where $a(N)$, $b(N) \to \infty$. Then $|\eta_i| \leq M_{N,i}$ (see below (2.10)), but now ρ_N may approach 0. In this case, (3.6) simply requires that $(p+m)/N \to 0$ at a rate faster than ρ_N^2 to overcome the decay of ρ_N. As will be seen (e.g., in Example 4.1), this only adds a minor restriction to the way the sample size increases.

Finally, it should be pointed out that the conditions given here by no mean are the weakest. However, going for the most generality is not the main goal of this paper. We also would like to keep our conditions easy to interpret, and to be associated with typical situations of GLMM.

We now define the estimate

$$(\tilde{\beta}, \tilde{\alpha}) = \text{the maximizer of } l \text{ over } R_N \tag{3.7}$$

(see (2.10)). Note that $(\tilde{\beta}, \tilde{\alpha})$ may be regarded as a restricted version of the MPE (see Remark 3 below). However, the following lemma states the relationship between $(\tilde{\beta}, \tilde{\alpha})$ and $(\hat{\beta}, \hat{\alpha})$, the MPE.

Lemma 3.2. *If* $\|\tilde{\beta}\| < b(N)$ *and* $\|\tilde{\alpha}\| < a(N)$, *then* $(\tilde{\beta}, \tilde{\alpha}) = (\hat{\beta}, \hat{\alpha})$.

This follows from Lemma 2.1. Note that $\{(\beta, \alpha) : \|\beta\| < b(N), \|\alpha\| < a(N)\} \subset$ the interior of R_N.

Also, we note that for suitably chosen $b(N)$ and $a(N)$ the vector (β_0, α_0) belongs to R_N with probability $\to 1$. For example, if p is fixed, then $\|\beta_0\|/b(N) \to 0$ for any $b(N) \to \infty$; if α_0 satisfies (2.4) and $\limsup \lambda_{\max}(D_0) < \infty$, then $P(\|\alpha_0\| \leq a(N)) \to 1$ provided $\log m/a(N)^2 \to 0$.

In the following theorem, p, the dimension of β, may be unbounded. As a result, the coefficients β_j are allowed to be dependent on p, and hence on N. But the number q, which is the number of random factors in a model with standard design, is assumed to be fixed.

Theorem 3.1. *Consider a GLMM* (2.1)—(2.3). *Suppose that* (3.1) *is satisfied,* $\limsup \lambda_{\max}(D_0) < \infty$, *the model is asymptotically identifiable for the fixed and random effects in the sense that it satisfies* (3.2) *for any* $\theta \in \Theta_0$, (3.3), *and* (3.4). *Furthermore, suppose that there are sequences* $\{a(N)\}$, $\{b(N)\}$ *satisfying* $P(\|\beta_0\| < b(N), \|\alpha_0\| < a(N)) \to 1$ *such that* (3.6) *holds for any* $\phi \in \Phi_0$. *Let* $(\tilde{\beta}, \tilde{\alpha})$ *be the estimate defined by* (3.7). *Then, for any* $\theta \in \Theta_0$, $\phi \in \Phi_0$, *we have*

$$\frac{1}{N}\Big\{\sum_{j=1}^{p}|X_j|^2(\tilde{\beta}_j-\beta_{0j})^2+\sum_{k=1}^{m}|Z_k|^2(\tilde{\alpha}_k-\alpha_{0k})^2\Big\} \xrightarrow{P} 0 \tag{3.8}$$

and $|\tilde{\beta}-\beta_0| \xrightarrow{P} 0$. *In particular, if the model has standard design for the random effects, then*

$$\Big(\sum_{v=1}^{m_u} n_{uv}\Big)^{-1}\sum_{v=1}^{m_u} n_{uv}(\tilde{\alpha}_{uv}-\alpha_{0uv})^2 \xrightarrow{P} 0\,,\quad 1\le u\le q\,, \tag{3.9}$$

where $\tilde{\alpha}=(\tilde{\alpha}_{(1)},\ldots,\tilde{\alpha}_{(q)})$, $\alpha_0=(\alpha_{0(1)},\ldots,\alpha_{0(q)})$ *with* $\tilde{\alpha}_{(u)}=(\tilde{\alpha}_{uv})_{1\le v\le m_u}$ *and* $\alpha_{0(u)}=(\alpha_{0uv})_{1\le v\le m_u}$.

The proof of Theorem 3.1 is given in the appendix.

Remark 1. Although the MPE is derived under the normality assumption (2.4), the above theorem does not require that the actual random effects be normally distributed.

Remark 2. It is well known that in some cases, e.g., the Neyman-Scott problem (Neyman and Scott (1948)), the MLEs are not consistent when the number of nuisance parameters goes to infinity. This problem does not surface under the conditions of Theorem 3.1.

Remark 3. Consistency is proved only for the restricted estimate (3.7). The proof of Theorem 3.1 does not imply that $P((\tilde{\beta},\tilde{\alpha})=(\tilde{\beta},\tilde{\alpha}))\to 1$. However, since $P(\|\beta_0\|<b(N),\|\alpha_0\|<a(N))\to 1$, it is natural to consider an estimate in the same range as the true effects. Thus the restricted estimate $(\tilde{\beta},\tilde{\alpha})$ is not unreasonable (see (3.7) and (2.10)).

4. More Examples

Example 4.1. Consider the logit random effects model defined by (1.1). Clearly, the model has standard design with $X=1_{m_1}\otimes 1_{m_2}$, $Z_{(1)}=I_{m_1}\otimes 1_{m_2}$, $Z_{(2)}=1_{m_1}\otimes I_{m_2}$, where $\otimes$ represents Kronecker product. It is easy to see that $s=2$. Also, $W=\sqrt{m_1m_2}\,\mathrm{diag}(1,(1/\sqrt{m_1})I_{m_1},(1/\sqrt{m_2})I_{m_2})$. For any $(\mu,a,b)\in S=\{a.=b.=0\}$, where $x.$ means summation over the components of x, let

$(h,u,v) = W(\mu,a,b)$. Then

$$(h,u,v)^t W^{-1}HW^{-1}(h,u,v) = (\mu,a,b)^t H(\mu,a,b) = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2}(\mu+u_i+v_j)^2$$
$$= m_1m_2\mu^2 + m_2\sum_{i=1}^{m_1}u_i^2 + m_1\sum_{j=1}^{m_2}v_j^2$$
$$= (h,u,v)^t(h,u,v)\ .$$

Therefore, $\lambda_{\min}(W^{-1}HW^{-1})|_{WS} = 1$. We have (3.4) since $X^tX = N = m_1m_2$, and (3.3) is satisfied if $m_1 \wedge m_2 \to \infty$.

Suppose $m_1, m_2 \to \infty$ such that $\log m_1/(\log m_2)^2 \to 0$, $\log m_2/(\log m_1)^2 \to 0$. Let $\{d_N\}$, $\{b(N)\}$ be such that d_N, $b(N) \to \infty$, $b(N)/\log(m_1 \wedge m_2) \to 0$, and $d_N\sqrt{\log(m_1 \vee m_2)}/\log(m_1 \wedge m_2) \to 0$. Let $a(N) = d_N\sqrt{\log(m_1 \vee m_2)}$, $c(N) = b(N) + 2a(N)$. Then, $(\|a_0\| \vee \|b_0\|)/a(N) \to 0$. Also, $\rho_N = \min_{i,j}\{\inf_{|\lambda|\le M_{N,(i,j)}} b''_{i,j}(\lambda)\} \ge \inf_{|\lambda|\le c(N)}\{e^\lambda/(1+e^\lambda)^2\} \ge (1/4)e^{-2c(N)}$. Thus it is easy to show that (3.6) is satisfied.

It follows from Theorem 3.1 that $\tilde\mu \xrightarrow{P} \mu_0$, $(1/m_1)\sum_{i=1}^{m_1}(\tilde u_i - u_{0i})^2 \xrightarrow{P} 0$, and $(1/m_2)\sum_{j=1}^{m_2}(\tilde v_j - v_{0j})^2 \xrightarrow{P} 0$ no matter at which $\sigma^2 > 0$, $\tau^2 > 0$ the $\tilde\mu$, $\tilde u$, and $\tilde v$ are computed.

Note. Haberman (1977) has studied a model initially considered by Rasch (1960, 1961) for educational tests. The Rasch model has similar structure as Example 4.1 but assumes that the effects u_i and v_j are fixed. Nevertheless, the method of Haberman (1977) is applicable to Example 4.1. One finds improved results, namely that $\max_i|\tilde u_i - u_{0i}| \xrightarrow{P} 0$ and $\max_j|\tilde v_j - v_{0j}| \xrightarrow{P} 0$, under weaker conditions on (m_1,m_2) and stronger conditions on the random effects. More specifically, Haberman (1977) has assumed that $u_i + v_j$ is bounded. Although such an assumption may seem reasonable for fixed parameters, it may not be so realistic for random effects. It should also be noted that in many cases of GLMM, there may not be sufficient information for each individual random effect.

Example 4.2. Suppose $y_{ij}, 1 \le i \le m, 1 \le j \le n_i$ are binary responses with $\text{logit}(P(y_{ij} = 1|\alpha)) = \eta_{ij} = \beta_0 + \beta_1 x_{ij} + \alpha_i$, where x_{ij}'s are covariates and $\alpha_1,\dots,\alpha_m \sim N(0,\sigma^2)$. Then

$$X = \begin{pmatrix} 1_{n_1} & X_1 \\ \vdots & \vdots \\ 1_{n_m} & X_m \end{pmatrix},\quad Z = \begin{pmatrix} 1_{n_1} & & \\ & \ddots & \\ & & 1_{n_m} \end{pmatrix},\quad \text{where } X_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{in_i} \end{pmatrix},\ 1 \le i \le m\ .$$

Let $c = (1/N)\sum_{i=1}^m\sum_{j=1}^{n_i}(x_{ij} - w_i)^2$ and $d = (1/N)\sum_{i=1}^m\sum_{j=1}^{n_i}x_{ij}^2$, where $N = \sum_{i=1}^m n_i$ is the total sample size and $w_i = \bar x_{i\cdot} = \sum_{j=1}^{n_i}x_{ij}/n_i$. We assume

$c > 0$, i.e., there is variation within cells. Note that in this case $s = 1$. Let

$$T = \begin{pmatrix} 1 & \bar{w} & 0 \cdots 0 \\ 0 & 1 & 0 \cdots 0 \\ 0 & \bar{w}_1 - \bar{w} & 1 \cdots 0 \\ \vdots & \vdots & \vdots \ddots \vdots \\ 0 & \bar{w}_m - \bar{w} & 0 \cdots 1 \end{pmatrix},$$

where $\bar{w} = \sum_{i=1}^m w_i/m$. For any number $v > 0$, define $W_v = \sqrt{N}\text{diag}(1, \sqrt{v}, \sqrt{n_1/N}, \ldots, \sqrt{n_m/N})$. For any $(\beta_0, \beta_1, \alpha) \in S = \{\alpha_\cdot = 0\}$, let $(v_0, v_1, u) = W_d(\beta_0, \beta_1, \alpha)$, $(\beta_0^*, \beta_1^*, \alpha^*) = T(\beta_0, \beta_1, \alpha)$. Then

$$\begin{aligned} (v_0, v_1, u)^t W^{-1} H W^{-1} (v_0, v_1, u) &= \sum_{i=1}^m \sum_{j=1}^{n_i} (\beta_0 + \beta_1 x_{ij} + \alpha_i)^2 \\ &= N[(\beta_0^*)^2 + c(\beta_1^*)^2] + 2\beta_0^* \sum_{i=1}^m (n_i - \lambda)\alpha_i^* + \sum_{i=1}^m n_i (\alpha_i^*)^2 \\ &\geq N[(\beta_0^*)^2 + c(\beta_1^*)^2] - 2\kappa\sqrt{N}|\beta_0^*| \Big(\sum_{i=1}^m n_i (\alpha_i^*)^2 \Big)^{1/2} + \sum_{i=1}^m n_i (\alpha_i^*)^2 , \quad (4.1) \end{aligned}$$

where λ is an arbitrary number and $\kappa = (N^{-1} \sum_{i=1}^m n_i^{-1} (n_i - \lambda)^2)^{1/2}$. If we pick λ to minimize κ, we find $\lambda = m(\sum_{i=1}^m n_i^{-1})^{-1}$. With such a λ, $\kappa = (1 - \tau_N^{-1})^{1/2}$, where $\tau_N = (\sum_{i=1}^m n_i/m)(\sum_{i=1}^m n_i^{-1}/m) \geq 1$. By Lemma 4.1 in the following and the fact that $N\bar{w}^2 + \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - w_i)^2 + \sum_{i=1}^m n_i (w_i - \bar{w})^2 = N\bar{w}^2 + \sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \bar{w})^2 \leq (2 + 3\tau_N) \sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij}^2$, it is easy to show that

$$\begin{aligned} &\lambda_{\min}(W^{-1} H W^{-1})|_{WS} \\ &\geq \Big(1 - (1 - \tau_N^{-1})^{1/2}\Big)\Big(\frac{c}{d}\Big) \Big/ \Big(1 + \frac{\bar{w}^2}{d} + \frac{c}{d} + \sum_{i=1}^m \Big(\frac{n_i}{N}\Big)\Big(\frac{w_i - \bar{w}}{\sqrt{d}}\Big)^2\Big) \\ &\geq \frac{1}{3}\Big(\frac{1 - (1 - \tau_N^{-1})^{1/2}}{1 + \tau_N}\Big)\Big(\frac{c}{d}\Big) . \end{aligned}$$

Thus (3.2) is satisfied provided

$$\liminf \frac{\sum_{i=1}^m \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\cdot})^2}{\sum_{i=1}^m \sum_{j=1}^{n_i} x_{ij}^2} > 0 , \quad (4.2)$$

$$\limsup \Big(\frac{1}{m} \sum_{i=1}^m n_i\Big)\Big(\frac{1}{m} \sum_{i=1}^m \frac{1}{n_i}\Big) < \infty . \quad (4.3)$$

The equality (4.2) may be regarded as a normalized asymptotic version of our earlier assumption that $c > 0$, i.e., asymptotically, there is variation within cells.

To see what (4.3) means, imagine that a population consists of a large number of subpopulations. Consider the following two-stage sampling scheme. In the first stage a random sample of m subpopulations $P_1, \ldots, P_m$ is picked. In the second stage a sample of size n_i is drawn from P_i, where n_i is proportional to the size of P_i, i.e., $n_i = pN_i$, where N_i is the size of P_i, and p is a fixed proportion. In this case, we have $(\sum_{i=1}^m n_i/m)(\sum_{i=1}^m n_i^{-1}/m) = (\sum_{i=1}^m N_i/m)(\sum_{i=1}^m N_i^{-1}/m)$, which, for large m, is close to $E(\xi)E(\xi^{-1})$, where ξ is the size of a randomly picked subpopulation. Therefore (4.3) may be understood as requiring $E(\xi)E(\xi^{-1})$ to be not large. To see one example in which this is true, suppose $\xi = N_0 + \zeta$, where N_0 is a fixed integer representing the smallest subpopulation size, and $\zeta \sim$ Poisson(λ). Then it is easy to show that $E(\xi)E(\xi^{-1}) \le 2$ regardless of the values of N_0 and λ. As another example, suppose ξ is uniformly distributed over the set of integers $\{N_0[k^q], 1 \le k \le K\}$, where N_0 is a positive integer, $0 < q < 1$, and $[x]$ represents the largest integer $\le x$. Then there is a constant c, depending only on q, such that $E(\xi)E(\xi^{-1}) \le c$ regardless of the values of N_0 and K.

Furthermore, suppose $\liminf \sum_{1\le i\le m}\sum_{1\le j\le n_i}(x_{ij}-\bar{x}_{..})^2/N > 0$, where $\bar{x}_{..} = \sum_{1\le i\le m}\sum_{1\le j\le n_i} x_{ij}/N$, the x_{ij}'s are bounded, and $\max_{1\le i\le m} n_i/N \to 0$, $\log m/(\log(N/m))^2 \to 0$. Then it is easy to show that (3.3), (3.4) and (3.6) are satisfied, where for (3.6) we take $b(N) \sim \sqrt{\log(N/m)}$ and $a(N) \sim (\log(N/m))^{1/2}(\log m)^{1/4}$.

It follows from Theorem 3.1 that $\tilde{\beta}_j \xrightarrow{P} \beta_{0j}$, $j = 0, 1$ and $\sum_{i=1}^m n_i(\tilde{\alpha}_i - \alpha_{0i})^2/N \xrightarrow{P} 0$ no matter at which $\sigma^2 > 0$ the $\tilde{\beta}$ and $\tilde{\alpha}$ are computed.

Lemma 4.1. *Let*

$$A = \begin{pmatrix} 1 & a & 0 \cdots 0 \\ a & a^2 + b + \sum_{i=1}^m c_i^2 & c_1 \cdots c_m \\ 0 & c_1 & 1 \cdots 0 \\ \vdots & \vdots & \vdots \ddots \vdots \\ 0 & c_m & 0 \cdots 1 \end{pmatrix},$$

then $\lambda_{\min}(A) \ge b/(1 + a^2 + b + \sum_{i=1}^m c_i^2)$.

This follows from the fact that $\det(\lambda I_{m+2} - A) = (\lambda - 1)^m[\lambda^2 - (1 + a^2 + b + \sum_{i=1}^m c_i^2)\lambda + b]$.

In some cases, the estimation of the fixed effects is the main concern. It is interesting to know if the MPE for β will be consistent under weaker assumptions. For example, in Example 2.2, if $m \to \infty$ but n remains bounded will the MPE for μ still be consistent regardless of the value of σ^2 at which the MPE is computed? Figure 2.1 seems to suggest something in this direction, but this is not correct.

Example 4.3. Consider Example 2.2. It can be shown that the MPE for μ is inconsistent as $m \to \infty$ and n remains bounded, even if it is evaluated at the

true σ^2 (see Jiang (1999)). This example shows that the asymptotic behavior of the MPE in a GLMM may be quite different from that of the BLUE in a linear mixed model. For example, in Example 2.1, $\hat{\beta}$ is consistent even if n is fixed and $m \to \infty$.

A further topic of research in this regard is the asymptotic distribution of the MPE. Such results would be useful in obtaining interval estimates for the fixed and random effects. It is also of interest to know the convergence rates of the MPE to the true fixed and random effects, and whether the rates are affected by at which θ the MPE are evaluated.

5. Remarks on Estimation of Variance Components

Despite the results of Section 3, i.e., for certain large samples the consistency of the MPE is not affected by the variance components, these results do not mean to challenge the importance of the estimation of variance components. In fact, in many cases (e.g., in genetics) the variance components are of main interest. On the contrary, the MPE leads to an easy way of consistently estimating the variance components when the sample size is large. Consider, for example, Example 4.1. If the sample size is increasing in the specified way, it is easy to show that $\tilde{\sigma}^2 = (1/m)\sum_{i=1}^m \tilde{u}_i^2 \xrightarrow{P} \sigma_0^2$, $\tilde{\tau}^2 = (1/n)\sum_{j=1}^n \tilde{v}_j^2 \xrightarrow{P} \tau_0^2$. However, things may be different if the sample size is either not large or is large but not in a favorable way (e.g., in Example 2.2, m large but n small).

In cases where the sample size is small, one may consider the following modified pseudo-profile likelihood approach. Let $l_P(\theta) = \log f(y, \hat{\gamma}|\theta)$, where $\hat{\gamma}$ is the MPE for γ. This may be regarded as a log-pseudo profile likelihood. If one intends to estimate θ based on l_P, one might pick $\theta = \hat{\theta}_P$ to maximize l_P, or solve

$$\frac{\partial l_P}{\partial \theta_i} = 0 \,, \quad i = 1, \ldots, q \,, \tag{5.1}$$

where q is the dimension of θ. However these equations are biased in that $E_\theta(\partial l_P/\partial\theta_i) \neq 0$. Methods of adjusting the profile likelihoods have been studied (e.g., McCullagh and Tibshirani (1990)). Here we consider modification of the equations (5.1) instead of l_P itself. Note that every modification of l_P leads to a change of equations, but the converse is not true (because the modified equation may not correspond to a likelihood equation, if θ is multi-dimensional). The modified equations are

$$\frac{\partial l_P}{\partial \theta_i} = E_\theta\Big(\frac{\partial l_P}{\partial \theta_i}\Big) \,, \quad i = 1, \ldots, q \,. \tag{5.2}$$

It should be pointed out that the computation of the expectation in (5.2) is more complicated when the restricted estimate in Theorem 3.1 is used. Note

that if l_P were a log-likelihood, the RHS of (5.2) would be 0 and (5.2) would be the maximum likelihood (ML) equations. In fact it can be shown that, under linear mixed models, (5.2) reduces to the ML equations if γ consists of all the random effects and θ the fixed effects and variance components, or to the REML equations (e.g., Searle, Casella and McCulloch (1992, §6)) if γ consists of all the fixed and random effects and θ the variance components. Note that the RHS of (5.2) may be evaluated via Monte-Carlo methods.

6. An Application

One area of application of Theorem 3.1 is small area estimation. In sample surveys, direct-survey estimates for small geographic areas or subpopulations are likely to yield inaccurate results, because the sample sizes from such areas are usually small. Therefore, it is necessary to "borrow strength" from related areas to find more accurate estimates for a given area or, simultaneously, for several areas. For continuous responses, such an idea has led to a linear mixed model approach, treating the area effects as random (see Ghosh and Rao (1994) for a review). For binary responses, Malec *et al* (1997) used a mixed logistic model for inference about small areas.

Figure 6.1.

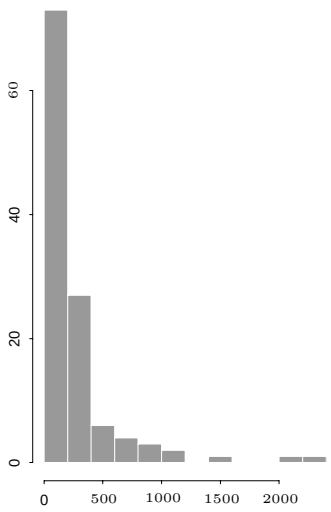
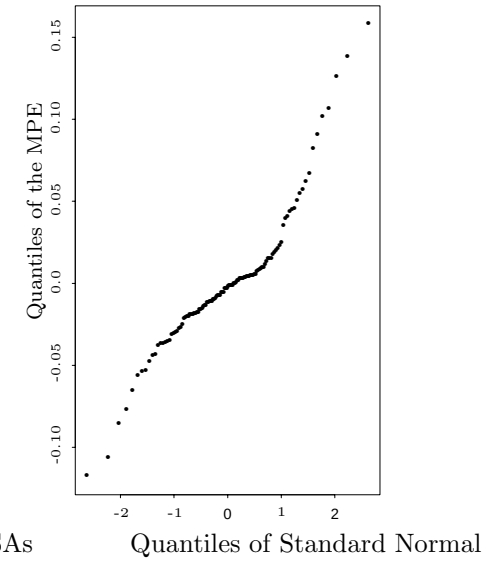


Figure 6.2.



Note that to apply Theorem 3.1, one does not have to assume that there are a large number of repetitions for all the random effects (see Example 4.2). Instead, it is important that the total number of random effects m be a small proportion

of the total sample size N. Situations like this (i.e., with large N and m and small m/N) are typical in small area estimation. In the following, we consider an application of Theorem 3.1 to a data set from the BRFSS (see the end of Section 1). The BRFSS is a Centers for Disease Control and Prevention coordinated, state-based random-digit-dialing telephone survey. The data we are particularly interested in is for the use of mammography among women aged 40 or older, from 1993 to 1995, and for areas from three Federal Regional Offices. The regional offices are Boston (Maine, Vermont, Mass., Conn., R.I., and N.H.), New York (N.Y. and N.J.), and Philadelphia (Penn., Del., D.C., Maryland, Va., and W.Va.). Our data suggests that mammography rates gradually increase from age groups 40-44 to 50-54, and then decrease. To catch this curvature phenomena, the use of a quadratic model to describe the age effect seems appropriate. The following model is proposed for the proportion p of women having had mammography:

$$\text{logit}(p)=\beta_0+\beta_1*\text{age}+\beta_2*\text{age}^2+\beta_3*\text{race}+\beta_4*\text{edu }\%+\text{HSA effect}, \quad (6.1)$$

where age is grouped as 40-44, 45-49, ..., 75-79, 80 and over; race as white and others; edu % means the percent of people in the HSA (Health Service Area) aged 25 or older with at least high school education; and the additional HSA effect is considered as random.

There are 118 HSA's in the region. The total sample size is $29,505$, and the sample sizes for the HSA's range from 4 to 2301 (see Figure 6.1 for the histogram of the sample sizes). The ratio of the number of HSA's to sample size is 0.004, so one would expect the MPE for the coefficients β's to be accurate, and the mean squared error of the MPE for the HSA effects to be small.

We compute the MPEs at $\sigma = 0.1$ (σ^2 is the variance of the random effects). By Theorem 3.1 this should not affect the accuracy of the MPE by much. The MPE for the β's are $\hat{\beta}_0 = -0.421$, $\hat{\beta}_1 = 0.390$, $\hat{\beta}_2 = -0.047$, $\hat{\beta}_3 = -0.175$, and $\hat{\beta}_4 = 2.155$. A Q-Q plot of the MPE for the HSA effects is shown in Figure 6.2. Although the random effects do not seem to be normally distributed, such an assumption is not required by Theorem 3.1. The MPE for both the fixed and random effects are obtained by the Gauss-Seidel algorithm (Jiang (2000)), which converges quickly in this case. Based on the MPE we obtain an estimate of σ (see the first paragraph of Section 5) as $\hat{\sigma} = 0.042$. Finally, based on the MPE for both the fixed and random effects, and the proportions for the age and race groups in the HSA's from the 1990 U.S. Census, we obtain estimates of the proportion of women having had mammography in the HSA's. A map is made based on these estimates, see Figure 6.3.

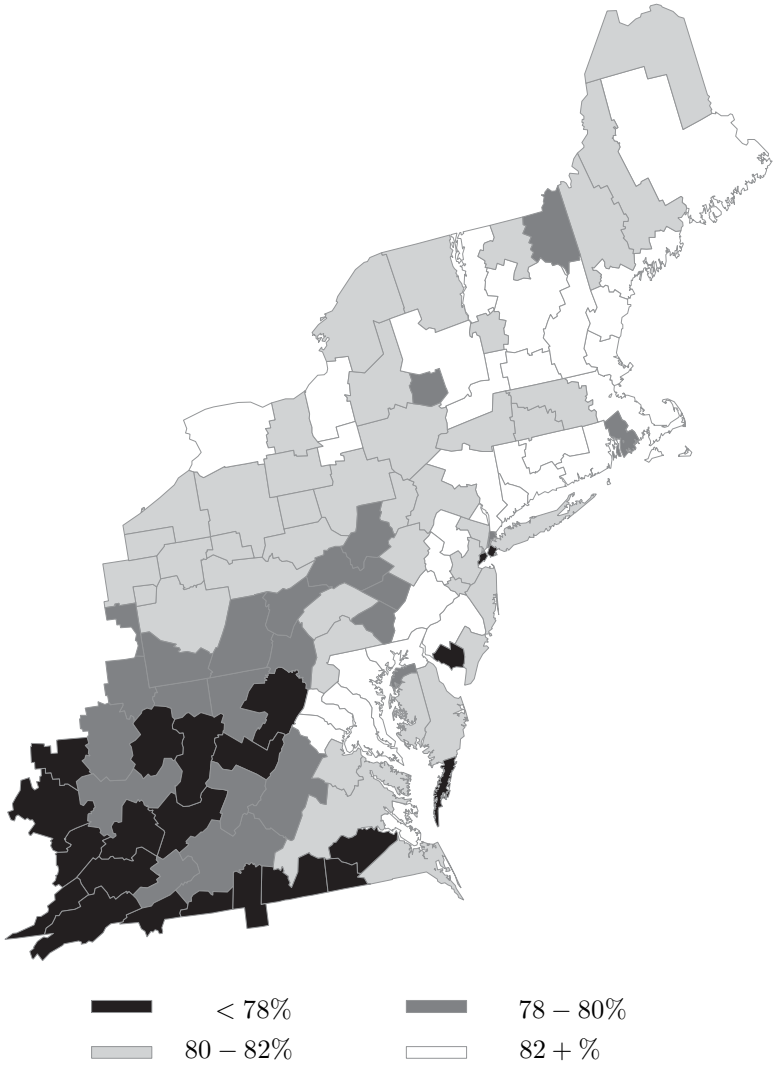


Figure 6.3

Acknowledgement

The authors wish to thank Professor J. Sedransk for helpful discussions. The authors are also grateful to an associate editor and a referee for their constructive and insightful comments which help improve the manuscript. The research in this paper is supported by NSA Grant MDA904-98-1-0038 and by funding from a CDC contract.

Appendix. Proof of Theorem 3.1

To make the proof easy-to-read, we split it into five steps.

Step 1. A Taylor series expansion. By the argument above Lemma 3.1 and the fact that $(\beta^*, \alpha^*) \in S$ for any (β, α), it is easy to show that $(\tilde{\beta}, \tilde{\alpha}) \in S$. Again, recalling notation, let $\theta \in \Theta_0, \phi \in \Phi_0$. By (3.5) with (β_*, α_*) replaced by (β_0^*, α_0^*), a Taylor expansion of $\lambda(\cdot)$ at $\eta_0^* = X\beta_0^* + Z\alpha_0^*$, and the fact that $\eta_0^* = X\beta_0^* + Z\alpha_0^* = X\beta_0 + Z\alpha_0 = \eta_0$, we have

$$\begin{aligned} l(\beta,\alpha) - l(\beta_0^*,\alpha_0^*) &= \lambda(\eta) - \lambda(\eta_0^*) + (1/2)(\alpha_0^{*t}D^{-1}\alpha_0^* - \alpha^t D^{-1}\alpha) \\ &= \sum_{i=1}^N \frac{\partial\lambda}{\partial\eta_i}|_{\eta_0^*}(\eta_i - \eta_{0i}^*) + \frac{1}{2}\sum_{i=1}^N \frac{\partial^2\lambda}{\partial\eta_i^2}|_{\eta_*}(\eta_i - \eta_{0i}^*)^2 \\ &\quad + \frac{1}{2}(\alpha_0^{*t}D^{-1}\alpha_0^* - \alpha^t D^{-1}\alpha) \\ &= \sum_{i=1}^N \frac{y_i - b_i'(\eta_{0i})}{a_i(\phi)}(\eta_i - \eta_{0i}^*) + \frac{1}{2}(\alpha_0^{*t}D^{-1}\alpha_0^* - \alpha^t D^{-1}\alpha) \\ &\quad - \frac{1}{2}\sum_{i=1}^N \frac{b_i''(\eta_{*i})}{a_i(\phi)}(\eta_i - \eta_{0i}^*)^2 \\ &= I_1 + (1/2)I_2 - (1/2)I_3, \end{aligned} \tag{A.1}$$

where $\eta_* = (1-t)\eta_0^* + t\eta = (1-t)\eta_0 + t\eta$ for some $0 \le t \le 1$. Note that $\partial^2\lambda/\partial\eta_i\partial\eta_{i'} = 0$, $i \ne i'$.

Step 2. Bounds for I_1, I_2, and I_3. By (2.9) we have

$$\begin{aligned} I_1 &= \sum_{j=1}^p \Big(\sum_{i=1}^N \frac{x_{ij}}{a_i(\phi)}(y_i - E(y_i|\alpha_0))\Big)(\beta_j - \beta_{0j}^*) \\ &\quad + \sum_{k=1}^m \Big(\sum_{i=1}^N \frac{z_{ik}}{a_i(\phi)}(y_i - E(y_i|\alpha_0))\Big)(\alpha_k - \alpha_{0k}^*), \end{aligned}$$

$$\begin{aligned} &E\Big\{\sum_{j=1}^p \frac{1}{|X_j|^2}\Big(\sum_{i=1}^N \frac{x_{ij}}{a_i(\phi)}(y_i - E(y_i|\alpha_0))\Big)^2\Big\} \\ &= \sum_{j=1}^p \frac{1}{|X_j|^2} E\Big(E\{(\cdots)^2|\alpha_0\}\Big) \\ &= \sum_{j=1}^p \frac{1}{|X_j|^2} E\Big(\sum_{i=1}^N \frac{x_{ij}^2}{a_i^2(\phi)}\mathrm{var}(y_i|\alpha_0)\Big) \\ &\le K_N^2 p\ , \end{aligned}$$

where $K_N^2 = \max_{1\le i\le N} a_i^{-1}(\phi) E b_i''(\eta_0)$, and similarly

$$E\Big\{ \sum_{k=1}^{m} \frac{1}{|Z_k|^2} \Big(\sum_{i=1}^{N} \frac{z_{ik}}{a_i(\phi)} (y_i - E(y_i|\alpha_0)) \Big)^2 \Big\} \le K_N^2 m \, .$$

By Hölder's inequality we have

$$\begin{aligned} |I_1| \le K_N \Big\{ \sqrt{p} O_p^{(1)}(1) \Big(\sum_{j=1}^{p} |X_j|^2 (\beta_j - \beta_{0j}^*)^2 \Big)^{1/2} \\ + \sqrt{m} O_p^{(2)}(1) \Big(\sum_{k=1}^{m} |Z_k|^2 (\alpha_k - \alpha_{0k}^*)^2 \Big)^{1/2} \Big\}, \end{aligned} \tag{A.2}$$

where the $O_p(1)$'s do not depend on (β, α).
Now

$$\begin{aligned} E(\alpha_0^{*t} D^{-1} \alpha_0^*) &\le \operatorname{rank}(P_1^*) \sup_{|v|=1} E|v^t D^{-1/2} \alpha_0|^2 \\ &\le (m-s) \lambda_{\min}^{-1}(D) \lambda_{\max}(D_0) \, , \end{aligned} \tag{A.3}$$

and $(\alpha - \alpha_0^*)^t D^{-1} (\alpha - \alpha_0^*) \le \lambda_{\min}^{-1}(D) |\alpha - \alpha_0^*|^2 \le c_0^{-2} \lambda_{\min}^{-1}(D) \sum_{k=1}^m |Z_k|^2 (\alpha_k - \alpha_{0k}^*)^2$, by (3.1). Thus

$$\begin{aligned} I_2 &= -(\alpha - \alpha_0^*)^t D^{-1} (\alpha - \alpha_0^*) - 2\alpha_0^{*t} D^{-1} (\alpha - \alpha_0^*) \\ &\le 2(\alpha_0^{*t} D^{-1} \alpha_0^*)^{1/2} ((\alpha - \alpha_0^*)^t D^{-1} (\alpha - \alpha_0^*))^{1/2} \\ &\le \lambda_{\min}^{-1}(D) \lambda_{\max}^{1/2}(D_0) \sqrt{m-s} O_p^{(3)}(1) \Big(\sum_{k=1}^{m} |Z_k|^2 (\alpha_k - \alpha_{0k}^*)^2 \Big)^{\frac{1}{2}}, \end{aligned} \tag{A.4}$$

where the $O_p(1)$ does not depend on (β, α).
Since $|\eta_{*i}| \le (1-t)|\eta_{0i}| + t|\eta_i| \le M_{N,i}$ if both (β_0, α_0) and (β, α) are in R_N (see (2.10)) and $(\beta_0^*, \alpha_0^*) \in S$, we have, when $(\beta, \alpha) \in S \cap R_N$ and $(\beta_0, \alpha_0) \in R_N$, that

$$I_3 \ge \rho_N |\eta - \eta_0^*|^2 = \rho_N \begin{pmatrix} \beta - \beta_0^* \\ \alpha - \alpha_0^* \end{pmatrix}^t H \begin{pmatrix} \beta - \beta_0^* \\ \alpha - \alpha_0^* \end{pmatrix} \ge \rho_N \lambda_N d^2(\beta, \alpha) \tag{A.5}$$

(see (2.11)), where $\lambda_N = \lambda_{\min}(W^{-1} H W^{-1})|_{WS}$, $d^2(\beta, \alpha) = \sum_{j=1}^p |X_j|^2 (\beta_j - \beta_{0j}^*)^2 + \sum_{k=1}^m |Z_k|^2 (\alpha_k - \alpha_{0k}^*)^2$. Note that, by (3.2), λ_N is bounded away from 0.

Step 3. That

$$N^{-1} d^2(\tilde{\beta}, \tilde{\alpha}) \xrightarrow{P} 0 \, . \tag{A.6}$$

Combining (A.1), (A.2), (A.4), and (A.5) we have, whenever $(\beta,\alpha) \in S \cap R_N$ and $(\beta_0,\alpha_0) \in R_N$ (see (2.10)), that

$$\begin{aligned} l(\beta,\alpha) - l(\beta_0^*,\alpha_0^*) &\le [K_N(\sqrt{p}O_p^{(1)}(1) + \sqrt{m}O_p^{(2)}(1)) \\ &\quad + \lambda_{\min}^{-1}(D)\lambda_{\max}^{1/2}(D_0)\sqrt{m-s}O_p^{(3)}(1)]d(\beta,\alpha) \\ &\quad - (1/2)\rho_N\lambda_N d^2(\beta,\alpha)\ . \end{aligned} \tag{A.7}$$

Let $\epsilon > 0$, and $E_N = \{d^2(\beta,\alpha) \le \epsilon^2 N\}$. By (A.7) we see that, if $(\beta_0,\alpha_0) \in R_N$, then

$$\begin{aligned} &\sup_{(\beta,\alpha)\in S\cap R_N\cap E_N^c}\{d^{-2}(\beta,\alpha)(l(\beta,\alpha) - l(\beta_0^*,\alpha_0^*))\} \\ &\le \epsilon^{-1}\Big\{K_N\Big(\sqrt{\frac{p}{N}}O_p^{(1)} + \sqrt{\frac{m}{N}}O_p^{(2)}(1)\Big) \\ &\qquad + \lambda_{\min}^{-1}(D)\lambda_{\max}^{1/2}(D_0)\sqrt{\frac{m-s}{N}}O_p^{(3)}(1)\Big\} - \frac{1}{2}\lambda_N\rho_N \\ &= [(1/\epsilon)o_p(1) - (1/2)\lambda_N]\rho_N\ , \end{aligned} \tag{A.8}$$

using (3.6) for the last step. Since $P((\beta_0,\alpha_0) \in R_N) \to 1$ by (2.10) and the conditions of the theorem, (A.8), (3.2) and the fact that $(\tilde\beta,\tilde\alpha) \in S$ (see early result in Step I) imply that $P((\tilde\beta,\tilde\alpha) \in E_N) \to 1$. (A.6) thus follows by the arbitrariness of ϵ.

Step 4. That (3.8) holds. We have

$$\begin{aligned} \|(X^tX)^{-1}X^tZ\| &\le \|(X^tX)^{-1/2}\|\|(X^tX)^{-1/2}X^t\|\|Z\| \\ &= \lambda_{\min}^{-1/2}(X^tX)\|Z\|\ . \end{aligned} \tag{A.9}$$

It follows from (A.9), (3.3), and (3.4) that

$$s\|(X^tX)^{-1}X^tZ\|^2 \longrightarrow 0\ . \tag{A.10}$$

Finally, by the same argument as (A.3), we have $E|\alpha_0-\alpha_0^*|^2 \le \lambda_{\max}(D)$ $E(\alpha_0-\alpha_0^*)^tD^{-1}(\alpha_0-\alpha_0^*) \le s\lambda_{\max}(D_0)\lambda_{\max}(D)\lambda_{\min}^{-1}(D)$. Thus,

$$\begin{aligned} &\frac{1}{N}\Big\{\sum_{j=1}^p |X_j|^2(\tilde\beta_j-\beta_{0j})^2 + \sum_{k=1}^m |Z_k|^2(\tilde\alpha_k-\alpha_{0k})^2\Big\} \\ &\le \frac{2}{N}\Big\{d^2(\tilde\beta,\tilde\alpha) + \Big(\max_{1\le j\le p}|X_j|^2\Big)|(X^tX)^{-1}X^tZ(\alpha_0-\alpha_0^*)|^2 \\ &\qquad + \Big(\max_{1\le k\le m}|Z_k|^2\Big)|\alpha_0-\alpha_0^*|^2\Big\} \\ &\le 2\Big\{(d^2(\tilde\beta,\tilde\alpha)/N) + [\|(X^tX)^{-1}X^tZ\|^2(\|X\|^2/N) \\ &\qquad + (\|Z\|^2/N)]|\alpha_0-\alpha_0^*|^2\Big\} \\ &= 2\{N^{-1}d^2(\tilde\beta,\tilde\alpha) + o_p(1)\}\ , \end{aligned} \tag{A.11}$$

using (A.10), (3.4), and (3.3). Then (3.8) follows from (A.11) and (A.6).

Step 5. The rest of the conclusions. The only thing one needs to note is that, by (3.4),

$$\frac{1}{N}\sum_{j=1}^{p}|X_j|^2(\tilde{\beta}_j-\beta_{0j})^2 \geq \left[\frac{\lambda_{\min}(X^tX)}{N}\right]|\tilde{\beta}-\beta_0|^2 \quad \geq \quad \delta|\tilde{\beta}-\beta_0|^2$$

for some $\delta > 0$, if N is large.

References

Breslow, N. E. (1996). Comment on Lee & Nelder: hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58**, 667.

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

Clayton, D. G. (1996). Comment on Lee & Nelder: hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58**, 657-659.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statist. Sci.* **6**, 15-51.

Haberman, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5**, 815-841.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Ann. Math. Statist.* **21**, 309-310.

Jiang (1996). REML estimation: asymptotic behavior and related topics. *Ann. Statist.* **24**, 255-286.

Jiang, J. (1998a). Asymptotic properties of the empirical BLUP and BLUE in mixed linear models. *Statist. Sinica* **8**, 861-885.

Jiang, J. (1998b). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93**, 720-729.

Jiang, J. (1999). On maximum hierarchical likelihood estimators, *Commun. Statist. - Theory Meth.* **28**, 1769-1775.

Jiang, J. (2000). A nonlinear Gauss-Seidel algorithm for inference about GLMM. *Comp. Statist.* **15**, 229-241.

Kuk, A. Y. C. (1995). Asymptotically unbiased estimation in generalized linear models with random effects. *J. Roy. Statist. Soc. Ser. B* **57**, 395-407.

Lai, T. L. and Wei, C. Z. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *Ann. Statist.* **10**, 154-166.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *J. Roy. Statist. Soc. Ser. B* **58**, 619-678.

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *J. Amer. Statist. Assoc.* **91**, 1007-1016.

Malec, D., Sedransk, J., Moriarity, C. L. and LeClere, F. B. (1997). Small area inference for binary variables in the national health interview survey. *J. Amer. Statist. Assoc.* **92**, 815-826.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition. Chapman and Hall, New York.

McCullagh, P. and Tibshirani, R. (1990). A simple method for adjustment of profile likelihoods. *J. Roy. Statist. Soc. Ser. B* **52**, 325-344.

McGilchrist, C. A. (1994). Estimation in generalized mixed models. *J. Roy. Statist. Soc. Ser. B* **56**, 61-69.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrika* **16**, 1-32.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16**, 356-366.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* Danish Inst. of Edu. Research, Copenhagen.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. *Proc. Fourth Berkeley Symp. Math. Statist. Probab.* **4**, 321-334, Univ. of California Press.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects. *Statist. Sci.* **6**, 15-51.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika* **78**, 719-727.

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

Speed, T. (1991). Comment on Robinson: Estimation of random effects. *Statist. Sci.* **6**, 42-44.

Department of Statistics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, Ohio 44106-7054, U.S.A.

CHRG, Suite 309, Conference Center Building, Knoxville, TN 37996-4133, U.S.A.

Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455-0378, U.S.A.

(Received January 1998; accepted May 2000)