THE SUBSET ARGUMENT AND CONSISTENCY OF MLE IN GLMM: ANSWER TO AN OPEN PROBLEM AND BEYOND

By Jiming Jiang[1]

University of California, Davis

We give answer to an open problem regarding consistency of the maximum likelihood estimators (MLEs) in generalized linear mixed models (GLMMs) involving crossed random effects. The solution to the open problem introduces an interesting, nonstandard approach to proving consistency of the MLEs in cases of dependent observations. Using the new technique, we extend the results to MLEs under a general GLMM. An example is used to further illustrate the technique.

1. Introduction. Generalized linear mixed models (GLMMs) have become a popular and very useful class of statistical models. See, for example, Jiang (2007), McCulloch, Searle and Neuhaus (2008) for some wide-ranging accounts of GLMMs with theory and applications. In the earlier years after GLMM was introduced, one of the biggest challenges in inference about these models was computation of the maximum likelihood estimators (MLEs). As is well known, the likelihood function under a GLMM typically involves integrals that cannot be computed analytically. The computational difficulty was highlighted by the infamous salamander mating data, first introduced by McCullagh and Nelder [(1989), Section 14.5]. A mixed logistic model, which is a special case of GLMM, was proposed for the salamander data that involved crossed random effects for the female and male animals. However, due to the fact that the random effects are crossed, the likelihood function involves a high-dimensional integral that not only does not have an analytic expression, but is also difficult to evaluate numerically [e.g., Jiang (2007), Section 4.4.3]. For years, the salamander data has been a driving force for the computational developments in GLMM. Virtually every numerical procedure that was proposed used this data as a "gold standard" to evaluate, or demonstrate, the procedure. See, for example, Karim and Zeger (1992), Breslow and Clayton (1993), Drum and McCullagh (1993), McCulloch (1994), Breslow and Lin (1995), Lin and Breslow (1996), Jiang (1998), Booth and Hobert (1999), Jiang and Zhang (2001), Sutradhar and Rao (2003), and Torabi (2012).

1.1. *A theoretical challenge and an open problem.* To illustrate the numerical difficulty as well as a theoretical challenge, which is the main objective of the current paper, let us begin with an example.

EXAMPLE 1. A mixed logistic model was proposed by Breslow and Clayton (1993) for the salamander data, and has since been used [e.g., Breslow and Lin (1995), Jiang (1998), Lin and Breslow (1996)]. Some alternative models, but only in terms of reparametrizations, have been considered [e.g., Booth and Hobert (1999)]. Jiang and Zhang (2001) noted that some of these models have ignored the fact that a group of salamanders were used in both the summer experiment and one of the fall experiments; in other words, there were replicates for some of the pairs of female and male animals. Nevertheless, all of these models are special cases of the following, more general setting. Suppose that, given the random effects $u_i, v_j, (i, j) \in S$, where S is a subset of $\mathcal{I} = \{(i, j) : 1 \leq i \leq m, 1 \leq j \leq n\}$, binary responses $y_{ijk}, (i, j) \in S, k = 1, \ldots, c_{ij}$ are conditionally independent such that, with $p_{ijk} = \mathrm{P}(y_{ijk} = 1 | u, v)$, we have $\mathrm{logit}(p_{ijk}) = x'_{ijk}\beta + u_i + v_j$, where $\mathrm{logit}(p) = \log\{p/(1 - p)\}, p \in (0, 1), x_{ijk}$ is an known vector of covariates, β is a unknown vector of parameters, and u, v denote all the random effects u_i and v_j that are involved. Here c_{ij} is the number of replicates for the (i, j) cell. Without loss of generality, assume that S is a irreducible subset of $\mathcal{I}$ in that m, n are the smallest positive integers such that $S \subset \mathcal{I}$. Furthermore, suppose that the random effects u_i's and v_j's are independent with $u_i \sim N(0, \sigma^2)$ and $v_j \sim N(0, \tau^2)$, where σ^2, τ^2 are unknown variances. One may think of the random effects u_i and v_j as corresponding to the female and male animals, as in the salamander problem. In fact, for the salamander data, $c_{ij} = 2$ for half of the pairs (i, j), and $c_{ij} = 1$ for the rest of the pairs. It can be shown [e.g., Jiang (2007), page 126; also see Section 4 in the sequel] that the log-likelihood function for estimating β, σ^2, τ^2 involves an integral of dimension $m + n$, which, in particular, increases with the sample size, and the integral cannot be further simplified.

The fact that the random effects are crossed, as in Example 1, presents not only a computational challenge but also a theoretical one, that is, to prove that the MLE is consistent in such a model. In contrast, the situation is very different if the GLMM has clustered, rather than crossed, random effects. For example, consider the following.

EXAMPLE 2. Suppose that, given the random effects $u_1, \ldots, u_m$, binary responses $y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i$ are conditionally independent such that, with $p_{ij} = \mathrm{P}(y_{ij} = 1 | u)$, we have $\mathrm{logit}(p_{ij}) = x'_{ij}\beta + u_i$, where x_{ij} is a vector of known covariates, β a vector of unknown coefficients, and $u = (u_i)_{1 \leq i \leq m}$. Furthermore, suppose that the u_i's are independent with $u_i \sim N(0, \sigma^2)$, where σ^2 is unknown. It is easy to show that the log-likelihood function for estimating β, σ^2

only involves one-dimensional integrals. Not only that, a major theoretical advantage of this case is that the log-likelihood can be expressed as a sum of independent random variables. In fact, this is a main characteristic of GLMMs with clustered random effects. Therefore, limit theorems for sums of independent random variables [e.g., Jiang (2010), Chapter 6] can be utilized to obtain asymptotic properties of the MLE.

Generally speaking, the classical approach to proving consistency of the MLE [e.g., Lehmann and Casella (1998), Chapter 6; Jiang (2010)] relies on asymptotic theory for sum of random variables, independent or not. However, one cannot express the log-likelihood in Example 1 as a sum of random variables with manageable properties. For this reason, it is very difficult to tackle asymptotic behavior of the MLE in the salamander problem, or any GLMM with crossed random effects, assuming that the numbers of random effects in all of the crossed factors increase. In fact, the problem is difficult to solve even for the simplest case, as stated in the open problem below.

> Open problem [*e.g., Jiang (2010), page 541*]: *Suppose that $x'_{ijk}\beta = \mu$, an unknown parameter, $c_{ij} = 1$ for all i, j, $S = \mathcal{I}$, and σ^2, τ^2 are known, say, $\sigma^2 = \tau^2 = 1$ in Example 1. Thus, μ is the only unknown parameter. Suppose that $m, n \to \infty$. Is the MLE of μ consistent?*

It was claimed [Jiang (2010), pages 541, 550] that even for this seemingly trivial case, the answer was not known but expected to be anything but trivial.

1.2. *Origination of the open problem.*

The problem regarding consistency of the MLE in GLMMs with crossed random effects began to draw attention in early 1997. It remained unsolved over the past 15 years, and was twice cited as an open problem in the literature, first in Jiang [(2007), page 173] and later in Jiang [(2010), page 541]. The latter also provided the following supporting evidence for a positive answer [Jiang (2010), page 550].

Let $k = m \wedge n$. Consider a subset of the data, $y_{ii}, i = 1, \ldots, k$. Note that the subset is a sequence of i.i.d. random variables. It follows, by the standard arguments, that the MLE of μ based on the subset, denoted by $\tilde{\mu}$, is consistent. Let $\hat{\mu}$ denote the MLE of μ based on the full data, $y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n$. The point is that even the MLE based on a subset of the data, $\tilde{\mu}$, is consistent; and if one has more data (information), one is expected to do better. Therefore, $\hat{\mu}$ has to be consistent as well.

1.3. *The rest of the paper.*

In Section 2, we give a positive answer to the open problem as well as the proof. Surprisingly, the proof is fairly short, thanks to a new, nonstandard technique that we introduce, known as the *subset argument*. Using this argument, we are able to establish both Cramér (1946) and Wald (1949) types of consistency results for the MLE. It is fascinating that a 15-year-old prob-

lem can be solved in such a simple way. The new technique may be useful well beyond solving the open problem—for proving consistency of the MLE in cases of dependent observations. We consider some applications of the subset argument in Section 3 regarding consistency of the MLE in a general GLMM. An example is used in Section 4 to further illustrate the new technique. Remark and discussion on a number of theoretical and practical issues are offered in Section 5.

2. Answer to open problem. Throughout this section, we focus on the open problem stated in Section 1. Let μ denote the true parameter.

THEOREM 1 (Cramér consistency). *There is, with probability tending to one, a root to the likelihood equation, $\hat{\mu}$, such that $\hat{\mu} \xrightarrow{\text{P}} \mu$.*

PROOF. The idea was actually hinted in Jiang [(2010), page 550] as "evidence" that supports a positive answer (see the last paragraph of Section 1.2 of the current paper). Basically, the idea suggests that, perhaps, one could use the fact that the MLE based on the subset data is consistent to argue that the MLE based on the full data is also consistent. The question is how to execute the idea. Recall that, in the original proof of Wald [(1949); also see Wolfowitz (1949)], the focus was on the likelihood ratio $p_\theta(y)/p_{\theta_0}(y)$, and showing that the ratio converges to zero outside any (small) neighborhood of θ_0, the true parameter vector. Can we execute the subset idea in terms of the likelihood ratio? This leads to consideration of the relationship between the likelihood ratio under the full data and that under the subset data. It is in this context that the following *subset inequality* (2) is derived (see Section 5.1 for further discussion), which is the key to the proof.

Let $y_{[1]}$ denote the (row) vector of $y_{ii}, i = 1, \ldots, m \wedge n$, and $y_{[2]}$ the (row) vector of the rest of the $y_{ij}, i = 1, \ldots, m, j = 1, \ldots, n$. Let $p_\mu(y_{[1]}, y_{[2]})$ denote the probability mass function (p.m.f.) of $(y_{[1]}, y_{[2]})$, $p_\mu(y_{[1]})$ the p.m.f. of $y_{[1]}$,

$$(1) \qquad p_\mu(y_{[2]}|y_{[1]}) = \frac{p_\mu(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]})}$$

the conditional p.m.f. of $y_{[2]}$ given $y_{[1]}$, and P_μ the probability distribution, respectively, when μ is the true parameter. For any $\varepsilon > 0$, we have

$$\mathrm{P}_\mu\{p_\mu(y_{[1]}, y_{[2]}) \leq p_{\mu+\varepsilon}(y_{[1]}, y_{[2]})|y_{[1]}\} = \mathrm{P}_\mu\left\{\frac{p_{\mu+\varepsilon}(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} \geq 1 \Big| y_{[1]}\right\}$$

$$\leq \mathrm{E}\left\{\frac{p_{\mu+\varepsilon}(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} \Big| y_{[1]}\right\}$$

$$(2) \qquad = \sum_{y_{[2]}} \frac{p_{\mu+\varepsilon}(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} p_\mu(y_{[2]}|y_{[1]})$$

$$= \sum_{y_{[2]}} \frac{p_{\mu+\varepsilon}(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]})}$$

$$= \frac{p_{\mu+\varepsilon}(y_{[1]})}{p_\mu(y_{[1]})},$$

using (1). A more general form of (2) is given in Section 5.1.

On the other hand, by the standard asymptotic arguments [e.g., Jiang (2010), page 9], it can be shown that the likelihood ratio $p_{\mu+\varepsilon}(y_{[1]})/p_\mu(y_{[1]})$ converges to zero in probability, as $m \wedge n \to \infty$. Here we use the fact that the components of $y_{[1]}$, y_{ii}, $1 \le i \le m \wedge n$ are independent Bernoulli random variables. It follows that, for any $\eta > 0$, there is $N_\eta \ge 1$ such that, with probability $\ge 1 - \eta$, we have $\zeta_N = P_\mu\{p_\mu(y_{[1]}, y_{[2]}) \le p_{\mu+\varepsilon}(y_{[1]}, y_{[2]})|y_{[1]}\} \le \gamma^{m \wedge n}$ for some $0 < \gamma < 1$, if $m \wedge n \ge N_\eta$. The argument shows that $\zeta_N = O_P(\gamma^{m \wedge n})$, hence converges to 0 in probability. It follows, by the dominated convergence theorem, that $E_\mu(\zeta_N) = P_\mu\{p_\mu(y_{[1]}, y_{[2]}) \le p_{\mu+\varepsilon}(y_{[1]}, y_{[2]})\} \to 0$. Similarly, we have $P_\mu\{p_\mu(y_{[1]}, y_{[2]}) \le p_{\mu-\varepsilon}(y_{[1]}, y_{[2]})\} \to 0$. The rest of the proof follows by the standard arguments [e.g., Jiang (2010), pages 9–10]. □

The result of Theorem 1 is usually referred to as Cramér-type consistency [Cramér (1946)], which states that a root to the likelihood equation is consistent. However, it does not always imply that the MLE, which by definition is the (global) maximizer of the likelihood function, is consistent. A stronger result is called Wald-type consistency [Wald (1949); also see Wolfowitz (1949)], which states that the MLE is consistent. Note that the limiting process in Theorem 1 is $m, n \to \infty$, or, equivalently, $m \wedge n \to \infty$ (see Section 5.4 for discussion). With a slightly more restrictive limiting process, the Wald-consistency can actually be established, as follows.

THEOREM 2 (Wald consistency). *If* $(m \wedge n)^{-1} \log(m \vee n) \to 0$ *as* $m, n \to \infty$, *then the MLE of* μ *is consistent.*

PROOF. Define $p_0(\lambda) = E\{h(\lambda + \xi)\}$, where $h(x) = e^x/(1 + e^x)$ and $\xi \sim N(0, 2)$. Write $p_0 = p_0(\mu)$. For any integer k, divide the interval $[k, k+1)$ by $\lambda_{k,j} = k + \delta(mn)^{-1}(m \wedge n)j$, $j = 1, \ldots, J$, where $J = [mn/\delta(m \wedge n)]$ and $0 < \delta < 1 - p_0$. It is easy to show that $|(\partial/\partial\mu) \log p_\mu(y_{[1]}, y_{[2]})| \le mn$ uniformly for all μ. Thus, for any $\lambda \in [k, k+1)$, there is $1 \le j \le J$, such that $\log p_\lambda(y_{[1]}, y_{[2]}) - \log p_{\lambda_{k,j}}(y_{[1]}, y_{[2]}) = \{(\partial/\partial\mu) \log p_\mu(y_{[1]}, y_{[2]})|_{\mu=\tilde{\lambda}}\}(\lambda - \lambda_{k,j}) \le \delta(m \wedge n)$, where $\tilde{\lambda}$ lies between λ and $\lambda_{k,j}$. It follows that

$$\sup_{\lambda \in [k,k+1)} \frac{p_\lambda(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} \le e^{\delta(m \wedge n)} \max_{1 \le j \le J} \frac{p_{\lambda_{k,j}}(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})}.$$

Therefore, by the subset argument [see (2)], we have

$$P_\mu \left\{ \sup_{\lambda \in [k,k+1)} \frac{p_\lambda(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} > 1 \Big| y_{[1]} \right\}$$

(3)
$$\leq \sum_{j=1}^{J} P_\mu \left\{ \frac{p_{\lambda_{k,j}}(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} > e^{-\delta(m \wedge n)} \Big| y_{[1]} \right\}$$

$$\leq e^{\delta(m \wedge n)} \sum_{j=1}^{J} \frac{p_{\lambda_{k,j}}(y_{[1]})}{p_\mu(y_{[1]})}.$$

On the other hand, we have $0 \leq 1 - p_0(\lambda) = E\{1 + \exp(\lambda + \xi)\}^{-1} \leq e^{-\lambda}E(e^{-\xi}) = e^{1-\lambda}$; and, similarly, $0 \leq p_0(\lambda) \leq e^{1+\lambda}$. Let $\mathcal{A}_\delta = \{|\Delta| \leq \delta\}$ with $\Delta = (m \wedge n)^{-1} \sum_{i=1}^{m \wedge n} y_{ii} - p_0$. If $k \geq 1$, then, for any $1 \leq j \leq J$, write $p_1 = p_0(\lambda_{k,j})$. We have, on $\mathcal{A}_\delta$,

$$\frac{p_{\lambda_{k,j}}(y_{[1]})}{p_\mu(y_{[1]})} = \left\{ \left(\frac{p_1}{p_0} \right)^{p_0+\Delta} \left(\frac{1-p_1}{1-p_0} \right)^{1-p_0-\Delta} \right\}^{m \wedge n}$$

$$\leq \{a_\delta^{-1}(1-p_1)^{1-p_0-\delta}\}^{m \wedge n}$$

$$\leq [a_\delta^{-1} \exp\{(1-\lambda_{k,j})(1-p_0-\delta)\}]^{m \wedge n}$$

$$\leq \exp[\{1 - p_0 - \delta - \log a_\delta - (1-p_0-\delta)k\}(m \wedge n)],$$

where $a_\delta = \inf_{|x| \leq \delta} p_0^{p_0+x}(1-p_0)^{1-p_0-x} > 0$. It follows, by (3), that

$$P_\mu \left\{ \sup_{\lambda \in [k,k+1)} \frac{p_\lambda(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} > 1 \Big| y_{[1]} \right\}$$

$$\leq \frac{mn}{\delta(m \wedge n)} \exp[\{1 - p_0 - \log a_\delta - (1-p_0-\delta)k\}(m \wedge n)]$$

on $\mathcal{A}_\delta$, or, equivalently, that

$$P_\mu \left\{ \sup_{\lambda \in [k,k+1)} \frac{p_\lambda(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} > 1, |\Delta| \leq \delta \Big| y_{[1]} \right\}$$

(4)
$$\leq \frac{mn}{\delta(m \wedge n)} \exp[\{1 - p_0 - \log a_\delta - (1-p_0-\delta)k\}(m \wedge n)]1_{\mathcal{A}_\delta}.$$

Note that $\mathcal{A}_\delta \in \mathcal{F}(y_{[1]})$. By taking expectations on both sides of (4), it follows that the unconditional probability corresponding to the left side is bounded by the right side without $1_{\mathcal{A}_\delta}$, for $k = 1, 2, \ldots$. Therefore, we have

$$P_\mu \left\{ \sup_{\lambda \in [k,k+1)} \frac{p_\lambda(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} > 1 \text{ for some } k \geq K, |\Delta| \leq \delta \right\}$$

$$\leq \sum_{k=K}^{\infty} P_\mu \left\{ \sup_{\lambda \in [k,k+1)} \frac{p_\lambda(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} > 1, |\Delta| \leq \delta \right\}$$

$$\leq \frac{mn}{\delta(m \wedge n)} \exp\{(1 - p_0 - \log a_\delta)(m \wedge n)\} \sum_{k=K}^{\infty} e^{-(1-p_0-\delta)(m \wedge n)k}$$

(5)
$$= \frac{mn}{\delta(m \wedge n)} \exp\{(1 - p_0 - \log a_\delta)(m \wedge n)\} \frac{e^{-(1-p_0-\delta)(m \wedge n)K}}{1 - e^{-(1-p_0-\delta)(m \wedge n)}}$$

$$= \{1 - e^{-(1-p_0-\delta)(m \wedge n)}\}^{-1}$$

$$\times \exp[-(m \wedge n)\{(1 - p_0 - \delta)K - 1 + p_0 + \log a_\delta$$

$$- (m \wedge n)^{-1} \log(m \vee n) + (m \wedge n)^{-1} \log \delta\}].$$

Thus, if we choose K such that $(1 - p_0 - \delta)K - 1 + p_0 + \log a_\delta \geq 1$, then, for large $m \wedge n$, the probability on the left side of (5) is bounded by $2e^{-(m \wedge n)/2}$. On the other hand, we have $P_\mu(\mathcal{A}_\delta^c) \to 0$, as $m \wedge n \to \infty$. Thus, we have

(6)
$$P\left\{ \frac{p_\lambda(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} > 1 \text{ for some } \lambda \geq K \right\}$$

$$\leq 2e^{-(m \wedge n)/2} + P(\mathcal{A}_\delta^c) \longrightarrow 0$$

as $m \wedge n \to \infty$. Similarly, the left side of (6), with the words "$\lambda \geq K$" replaced by "$\lambda \leq -K$," goes to zero, as $m \wedge n \to \infty$, if K is chosen sufficiently large.

On the other hand, again by the subset argument, it can be shown (see the supplementary material [Jiang (2013)]) that for any $\varepsilon > 0$ and $K > |\mu| + \varepsilon$, we have

(7)
$$P_\mu \left\{ \sup_{\lambda \in [-K, \mu-\varepsilon) \cup (\mu+\varepsilon, K]} \frac{p_\lambda(y_{[1]}, y_{[2]})}{p_\mu(y_{[1]}, y_{[2]})} > 1 \right\} \longrightarrow 0$$

as $m, n \to \infty$. The consistency of the MLE then follows by combining (7) with the previously proved results. □

3. Beyond. We consider a few more applications of the subset argument, introduced in the previous section. All applications are regarding a general GLMM, whose definition is given below for the sake of completeness [see, e.g., Jiang (2007) for further details].

(i) Suppose that, given a vector u of random effects, responses $y_1, \ldots, y_N$ are conditionally independent with conditional density function, with respect to a σ-finite measure ν, given by the exponential family $f_i(y_i|u) = \exp[a_i^{-1}(\phi)\{y_i \xi_i - b(\xi_i)\} + c_i(y_i, \phi)]$, where ϕ is a dispersion parameter (which in some cases is known), and $b(\cdot), a_i(\cdot), c_i(\cdot, \cdot)$ are known, continuously differentiable functions with respect to ξ_i and ϕ. The natural parameter of the conditional exponential family, ξ_i, is therefore associated with the conditional mean, $\mu_i = E(y_i|u)$, according to the properties of the exponential family [e.g., McCullagh and Nelder (1989),

Section 2.2.2]. (ii) Furthermore, suppose that μ_i satisfies $g(\mu_i) = x_i'\beta + z_i'u$, where x_i, z_i are known vectors, β is a vector of unknown parameters, and $g(\cdot)$ is a link function. (iii) Finally, assume that $u \sim N(0, G)$, where the covariance matrix G may depend on a vector φ of dispersion parameters.

It is typically possible to find a subset of the data that are independent, in some way, under a general GLMM. For example, under the so-called ANOVA GLMM [e.g., Lin (1997)], a subset of independent data can always be found. Here an ANOVA GLMM satisfies $g(\mu) = X\beta + Z_1u_1 + \cdots + Z_su_s$, where $\mu = (\mu_i)_{1\leq i\leq N}$, $g(\mu) = [g(\mu_i)]_{1\leq i\leq N}$, $X = (x_i')_{1\leq i\leq N}$, $Z_r = (z_{ir}')_{1\leq i\leq N}$, $1 \leq r \leq s$, are known matrices, u_r, $1 \leq r \leq s$ are vectors of independent random effects, and $u_1, \ldots, u_s$ are independent. Examples 1 and 2 are special cases of the ANOVA GLMM. Note that in both examples the responses are indexed by (i, j), instead of i, but this difference is trivial. Nevertheless, the "trick" is to select a subset, or more than one subsets if necessary, with the following desirable properties: (I) the subset(s) can be divided into independent clusters with the number(s) of clusters increasing with the sample size; and (II) the combination of the subset(s) jointly identify all the unknown parameters. More specifically, let $y_i^{(a)}$, $i = 1, \ldots, N_a$ be the ath subset of the data, $1 \leq a \leq b$, where b is a fixed positive integer. Suppose that, for each a, there is a partition, $\{1, \ldots, N_a\} = \bigcup_{j=1}^{m_a} S_{a,j}$. Let $y_{a,j} = [y_i^{(a)}]_{i\in S_{a,j}}$, and $p_\theta(y_{a,j})$ be the probability density function (p.d.f.) of $y_{a,j}$, with respect to the measure ν (or the product measure induced by ν if $y_{a,j}$ is multivariate), when θ is the true parameter vector. Let Θ denote the parameter space, and θ_0 the true parameter vector. Then, (I) and (II) can be formally stated as follows:

(A1) $y_{a,j}$, $1 \leq j \leq m_a$ are independent with $m_a \to \infty$ as $N \to \infty$, $1 \leq a \leq b$;
(A2) for every $\theta \in \Theta \setminus \{\theta_0\}$, we have

$$\min_{1\leq a\leq b} \limsup_{N\to\infty} \frac{1}{m_a} \sum_{j=1}^{m_a} \mathrm{E}_{\theta_0}\left[\log\left\{\frac{p_\theta(y_{a,j})}{p_{\theta_0}(y_{a,j})}\right\}\right] < 0.$$

Note that (A2) controls the average Kullback–Leibler information [Kullback and Leibler (1951)]; thus, the inequality always holds if $<$ is replaced by $\leq$.

3.1. *Finite parameter space.*   Let us first consider a simpler case by assuming that Θ is finite. Although the assumption may seem restrictive, it is not totally unrealistic. For example, any computer system only allows a finite number of digits. This means that the parameter space that is practically stored in a computer system is finite. Using the subset argument, it is fairly straightforward to prove the following (see the supplementary material [Jiang (2013)]).

THEOREM 3.   *Under assumptions* (A1) *and* (A2), *if, in addition,*

(A3) *for every $\theta \in \Theta \setminus \{\theta_0\}$, we have*

$$\frac{1}{m_a^2} \sum_{j=1}^{m_a} \mathrm{var}_{\theta_0}\left[\log\left\{\frac{p_\theta(y_{a,j})}{p_{\theta_0}(y_{a,j})}\right\}\right] \longrightarrow 0, \qquad 1 \le a \le b,$$

then $\mathrm{P}_{\theta_0}(\hat{\theta} = \theta_0) \to 1$, *as* $N \to \infty$, *where* $\hat{\theta}$ *is the MLE of* θ.

3.2. *Euclidean parameter space.* We now consider the case that Θ is a convex subspace of R^d, the d-dimensional Euclidean space, in the sense that $\theta_1, \theta_2 \in \Theta$ implies $(1-t)\theta_1 + t\theta_2 \in \Theta$ for every $t \in (0, 1)$. In this case, we need to strengthen assumptions (A2), (A3) to the following:

(B2) $\theta_0 \in \Theta^o$, the interior of Θ, and there is $0 < M < \infty$ [same as in (B3) below] such that, for every $\varepsilon > 0$, we have

$$(8) \qquad \limsup_{N \to \infty} \sup_{\theta \in \Theta, \varepsilon \le |\theta - \theta_0| \le M} \min_{1 \le a \le b} \frac{1}{m_a} \sum_{j=1}^{m_a} \mathrm{E}_{\theta_0}\left[\log\left\{\frac{p_\theta(y_{a,j})}{p_{\theta_0}(y_{a,j})}\right\}\right] < 0.$$

(B3) There are positive constant sequences $s_N, s_{a,N}, 1 \le a \le b$ such that

$$(9) \qquad \sup_{\theta \in \Theta, |\theta - \theta_0| \le M} \max_{1 \le c \le d} \left|\frac{\partial}{\partial \theta_c} \log\{p_\theta(y)\}\right| = O_\mathrm{P}(s_N)$$

with $\log(s_N) / \min_{1 \le a \le b} m_a \to 0$, where $p_\theta(y)$ is the p.d.f. of $y = (y_i)_{1 \le i \le N}$ given that $\theta = (\theta_c)_{1 \le c \le d}$ is the true parameter vector,

$$(10) \qquad \sup_{\theta \in \Theta, |\theta - \theta_0| \le M} \frac{1}{m_a} \sum_{j=1}^{m_a} \max_{1 \le c \le d} \left|\frac{\partial}{\partial \theta_c} \log\{p_\theta(y_{a,j})\}\right| = o_\mathrm{P}(s_{a,N})$$

with $\log(s_{a,N}) / m_a \to 0$; and (for the same $s_{a,N}$)

$$(11) \qquad \sup_{\theta \in \Theta, |\theta - \theta_0| \le M} \frac{s_{a,N}^{d-1}}{m_a^2} \sum_{j=1}^{m_a} \mathrm{var}_{\theta_0}\left[\log\left\{\frac{p_\theta(y_{a,j})}{p_{\theta_0}(y_{a,j})}\right\}\right] \longrightarrow 0, \qquad 1 \le a \le b.$$

THEOREM 4. *Under assumptions* (A1), (B2) *and* (B3), *there is, with probability* $\to 1$, *a root to the likelihood equation,* $\hat{\theta}$, *such that* $\hat{\theta} \xrightarrow{\mathrm{P}} \theta_0$, *as* $N \to \infty$.

PROOF. Aside from the use of the subset argument, the lines of the proof are similar to, for example, the standard arguments of Lehmann and Casella [(1998), the beginning part of the proof of Theorem 5.1], although some details are more similar to Wolfowitz (1949). We outline the key steps below and refer the details to the supplementary material [Jiang (2013)]. Once again, the innovative part is the consideration of the conditional probability given the subset data and, most importantly, the subset inequality (15) in the sequel.

For any $\varepsilon > 0$, assume, without loss of generality, that $\{\theta : |\theta - \theta_0| \leq \varepsilon\} \subset \Theta$ and $C_\varepsilon = \{\theta \in R^d : |\theta_c - \theta_{0c}| \leq \varepsilon, 1 \leq c \leq d\} \subset \{\theta \in \Theta : |\theta - \theta_0| \leq M\}$. Essentially, all we need to show is that, as $N \to \infty$,

$$(12) \qquad P(\varepsilon) \equiv P_{\theta_0}\left\{ p_{\theta_0}(y) \leq \sup_{\theta \in \partial C_\varepsilon} p_\theta(y) \right\} \longrightarrow 0,$$

where ∂C_ε is the boundary of C_ε, which consists of $\theta \in C_\varepsilon$ such that $|\theta_c - \theta_{0c}| = \varepsilon$ for some $1 \leq c \leq d$. Define

$$S_{N,a}(\theta) = \frac{1}{m_a} \sum_{j=1}^{m_a} E_{\theta_0}\left[\log\left\{ \frac{p_\theta(y_{a,j})}{p_{\theta_0}(y_{a,j})} \right\} \right], \qquad 1 \leq a \leq b,$$

and $I_N(\theta) = \min\{1 \leq a \leq b : S_{N,a}(\theta) = \min_{1 \leq a' \leq b} S_{N,a'}(\theta)\}$. Then, $\partial C_\varepsilon = \bigcup_{a=1}^b \partial C_\varepsilon \cap \Theta_{N,a}$, where $\Theta_{N,a} = \{\theta \in \Theta : I_N(\theta) = a\}$. Then, we have

$$(13) \qquad P(\varepsilon) \leq \sum_{a=1}^b P_{\theta_0}\left\{ p_{\theta_0}(y) \leq \sup_{\theta \in \partial C_\varepsilon \cap \Theta_{N,a}} p_\theta(y) \right\}.$$

For a fixed $1 \leq a \leq b$, let δ be a small, positive number to be determined latter, and $K = [e^{\delta m_a}] + 1$. For any $l = (l_1, \ldots, l_d)$, where $0 \leq l_c \leq K - 1, 1 \leq c \leq d$, select a point θ_l from the subset $\{\theta : \theta_{0c} - \varepsilon + 2\varepsilon l_c/K \leq \theta_c \leq \theta_{0c} - \varepsilon + 2\varepsilon(l_c + 1)/K, 1 \leq c \leq d\} \cap \partial C_\varepsilon \cap \Theta_{N,a}$, if the latter is not empty; otherwise, do not select. Let D denote the collection of all such points. Also let B denote the left side of (9). It can be shown that

$$(14) \qquad \begin{aligned} &P_{\theta_0}\left\{ p_{\theta_0}(y) \leq \sup_{\theta \in \partial C_\varepsilon \cap \Theta_{N,a}} p_\theta(y) \right\} \\ &\leq P_{\theta_0}\left\{ \exp\left(\frac{2d\varepsilon B}{K} \right) > 2 \right\} + P_{\theta_0}\left\{ p_{\theta_0}(y) \leq 2\max_{\theta \in D} p_\theta(y) \right\}. \end{aligned}$$

We now apply the subset argument. Let $y_{[1]}$ denote the combined vector of $y_{a,j}, 1 \leq j \leq m_a$, and $y_{[2]}$ the vector of the rest of $y_1, \ldots, y_N$. Then, similar to the argument of (2), we have, for any $\theta \in D$,

$$(15) \qquad P_{\theta_0}\left\{ p_{\theta_0}(y) \leq 2p_\theta(y)|y_{[1]} \right\} \leq 2\frac{p_\theta(y_{[1]})}{p_{\theta_0}(y_{[1]})}.$$

Using this result, it can be shown that $P_{\theta_0}\{p_{\theta_0}(y) \leq 2\max_{\theta \in D} p_\theta(y)|y_{[1]}\} = o_P(1)$. From here, (12) can be established. $\quad \square$

Again, Theorem 4 is a Cramér-consistency result. On the other hand, Wald-consistency can be established under additional assumptions that control the behavior of the likelihood function in a neighborhood of infinity. For example, the following result may be viewed as an extension of Theorem 2. The proof is given in the supplementary material [Jiang (2013)]. Once again, the subset argument

plays a critical role in the proof. For simplicity, we focus on the case of discrete responses, which is typical for GLMMs. In addition, we assume the following. For any $0 \leq v < w$, define $S_d[v, w) = \{x \in R^d : v \leq |x| < w\}$ and write, in short, $S_d(k) = S_d[k, k+1)$ for $k = 1, 2, \dots$.

(C1) There are sequences of constants, $b_k, c_N \geq 1$, and random variables, ζ_N, where c_N, ζ_N do not depend on k, such that $\zeta_N = O_P(1)$ and

$$\sup_{\theta \in \Theta \cap S_d[k-1, k+2)} \max_{1 \leq c \leq d} \left| \frac{\partial}{\partial \theta_c} \log\{p_\theta(y)\} \right| \leq b_k c_N \zeta_N, \qquad k = 1, 2, \dots$$

(C2) There is a subset of independent data vectors, $y_{(j)}, 1 \leq j \leq m_N$ [not necessarily among those in (A1)] so that: (i) $E_{\theta_0} |\log\{p_{j,\theta_0}(y_{(j)})\}|$ is bounded, $p_{j,\theta}(\cdot)$ being the p.m.f. of $y_{(j)}$ under θ; (ii) there is a sequence of positive constants, γ_k, with $\lim_{k\to\infty} \gamma_k = \infty$, and a subset $\mathcal{T}_N$ of possible values of $y_{(j)}$, such that for every $k \geq 1$ and $\theta \in \Theta \cap S_d(k)$, there is $t \in \mathcal{T}_N$ satisfying $\max_{1 \leq j \leq m_N} \log\{p_{j,\theta}(t)\} \leq -\gamma_k$; (iii) $\inf_{t \in \mathcal{T}_N} m_N^{-1} \sum_{j=1}^{m_N} p_{j,\theta_0}(t) \geq \rho$ for some constant $\rho > 0$; and (iv) $|\mathcal{T}_N|/m_N = o(1)$, and $c_N^d \sum_{k=K}^{\infty} k^{d_1} b_k^d e^{-\delta m_N \gamma_k} = o(1)$ for some $K \geq 1$ and $\delta < \rho$, where $d_1 = d 1_{(d>1)}$.

It is easy to verify that the new assumptions (C1), (C2) are satisfied in the case of Theorem 2 for the open problem (see the supplementary material [Jiang (2013)]). Another example is considered in the next section.

THEOREM 5. *Suppose that* (A1) *holds*; (B2), (B3) *hold for any fixed $M > 0$ (instead of some $M > 0$), and with the $s_{a,N}^{d-1}$ in* (11) *replaced by $s_{a,N}^d$. In addition, suppose that* (C1), (C2) *hold. Then, the MLE of θ_0 is consistent.*

4. Example. Let us consider a special case of Example 1 with $x_{ijk}'\beta = \mu$, but σ^2 and τ^2 unknown. We change the notation slightly, namely, $y_{i,j,k}$ instead of y_{ijk}. Suppose that $S = S_1 \cup S_2$ such that $c_{ij} = r, (i, j) \in S_r, r = 1, 2$ (as in the case of the salamander data). We use two subsets to jointly identify all the unknown parameters. The first subset is similar to that used in the proofs of Theorems 1 and 2, namely, $y_{i,i} = (y_{i,i,k})_{k=1,2}, (i, i) \in S_2$. Let m_1 be the total number of such (i, i)'s, and assume that $m_1 \to \infty$, as $m, n \to \infty$. Then, the subset satisfies (A1). Let $\theta = (\mu, \sigma^2, \tau^2)'$. It can be shown that the sequence $y_{i,i}, (i, i) \in S_2$ is a sequence of i.i.d. random vectors with the probability distribution, under θ, given by

$$(16) \qquad p_\theta(y_{i,i}) = E\left[\frac{\exp\{y_{i,i,\cdot}(\mu + \xi)\}}{\{1 + \exp(\mu + \xi)\}^2}\right],$$

where $\xi \sim N(0, \psi^2)$, with $\psi^2 = \sigma^2 + \tau^2$, and $y_{i,i,\cdot} = y_{i,i,1} + y_{i,i,2}$. By the strict concavity of the logarithm, we have

$$(17) \qquad E_{\theta_0}\left[\log\left\{\frac{p_\theta(y_{i,i})}{p_{\theta_0}(y_{i,i})}\right\}\right] < 0$$

unless $p_\theta(y_{i,i})/p_{\theta_0}(y_{i,i})$ is a.s. P_{θ_0} a constant, which must be one because both p_θ and p_{θ_0} are probability distributions. It is easy to show that the probability distribution of (16) is completely determined by the function $M(\vartheta) = [M_r(\vartheta)]_{r=1,2}$, where $M_r(\vartheta) = E\{h_\vartheta^r(\zeta)\}$ with $\vartheta = (\mu, \psi)'$, $h_\vartheta(\zeta) = \exp(\mu + \psi\zeta)/\{1 + \exp(\mu + \psi\zeta)\}$, and $\zeta \sim N(0,1)$. In other words, $p_\theta(y_{i,i}) = p_{\theta_0}(y_{i,i})$ for all values of $y_{i,i}$ if and only if $M(\vartheta) = M(\vartheta_0)$. Jiang (1998) showed that the function $M(\cdot)$ is injective [also see Jiang (2007), page 221]. Thus, (17) holds unless $\mu = \mu_0$ and $\psi^2 = \psi_0^2$.

It remains to deal with a θ that satisfies $\mu = \mu_0$, $\psi^2 = \psi_0^2$, but $\theta \neq \theta_0$. For such a θ, we use the second subset, defined as $y_i = (y_{i,2i-1,1}, y_{i,2i,1})'$ such that $(i, 2i-1) \in S$ and $(i, 2i) \in S$. Let m_2 be the total number of all such i's, and assume that $m_2 \to \infty$ as $m, n \to \infty$. It is easy to see that (A1) is, again, satisfied for the new subset. Note that any θ satisfying $\mu = \mu_0$ and $\psi^2 = \psi_0^2$ is completely determined by the parameter $\gamma = \sigma^2/\psi^2$. Furthermore, the new subset is a sequence of i.i.d. random vectors with the probability distribution, under such a θ, given by

$$(18) \qquad p_\gamma(y_i) = E\left[\frac{\exp\{y_{i,2i-1,1}(\mu_0 + X)\}}{1 + \exp(\mu_0 + X)} \cdot \frac{\exp\{y_{i,2i,1}(\mu_0 + Y)\}}{1 + \exp(\mu_0 + Y)}\right],$$

where (X, Y) has the bivariate normal distribution with $\mathrm{var}(X) = \mathrm{var}(Y) = \psi_0^2$ and $\mathrm{cor}(X, Y) = \gamma$. Similar to (17), we have

$$(19) \qquad\qquad E_{\gamma_0}\left[\log\left\{\frac{p_\gamma(y_i)}{p_{\gamma_0}(y_i)}\right\}\right] < 0$$

unless $p_\gamma(y_i) = p_{\gamma_0}(y_i)$ for all values of y_i. Consider (18) with $y_i = (1, 1)$ and let P_γ denote the probability distribution of (X, Y) with the correlation coefficient γ. By Fubini's theorem, it can be shown that

$$(20) \quad p_\gamma(1,1) = \int_0^\infty \int_0^\infty P_\gamma\{X \geq \mathrm{logit}(s) - \mu_0, Y \geq \mathrm{logit}(t) - \mu_0\}\,ds\,dt.$$

Hereafter, we refer the detailed derivations to the supplementary material [Jiang (2013)]. By Slepian's inequality [e.g., Jiang (2010), pages 157–158], the integrand on the right side of (20) is strictly increasing with γ, hence so is the integral. Thus, if $\gamma \neq \gamma_0$, at least we have $p_\gamma(1,1) \neq p_{\gamma_0}(1,1)$, hence (19) holds.

In summary, for any $\theta \in \Theta, \theta \neq \theta_0$, we must have either (17) or (19) hold. Therefore, by continuity, assumption (B2) holds, provided that true variances, σ_0^2, τ_0^2 are positive. Note that, in the current case, the expectations involved in (B2) do not depend on either j or N, the total sample size.

To verify (B3), it can be shown that $|(\partial/\partial\mu)\log\{p_\theta(y)\}| \leq N$. Furthermore, we have $|(\partial/\partial\sigma^2)\log\{p_\theta(y)\}| \vee |(\partial/\partial\tau^2)\log\{p_\theta(y)\}| \leq (A + C + 1)N$ in a neighborhood of θ_0, $\mathcal{N}(\theta_0)$. Therefore, (9) holds with $s_N = N$.

As for (10), it is easy to show that the partial derivatives involved are uniformly bounded for $\theta \in \mathcal{N}(\theta_0)$. Thus, (10) holds for any $s_{a,N}$ such that $s_{a,N} \to \infty$, $a = 1, 2$. Furthermore, the left side of (11) is bounded by $c_a s_{a,N}^2/m_a$ for some constant $c_a > 0$, $a = 1, 2$ (note that $d = 3$ in this case). Thus, for example, we may choose

$s_{a,N} = \sqrt{m_a/\{1 + \log(m_a)\}}$, $a = 1, 2$, to ensure that $\log(s_{a,N})/m_a \to 0$, $a = 1, 2$, and (11) holds.

In conclusion, all the assumptions of Theorem 4 hold provided that $\sigma_0^2 > 0$, $\tau_0^2 > 0$, and $(m_1 \wedge m_2)^{-1} \log(N) \to 0$.

Similarly, the conditions of Theorem 5 can be verified. Essentially, what is new is to check assumptions (C1) and (C2). See the supplementary material [Jiang (2013)].

5. Discussion.

5.1. *Remark on subset argument.* In proving a number of results, we have demonstrated the usefulness of the subset argument. In principle, the method allows one to argue consistency of the MLE in any situation of dependent data, not necessarily under a GLMM, provided that one can identify some suitable subset(s) of the data whose asymptotic properties are easier to handle, such as collections of independent random vectors. The connection between the full data and subset data is made by the subset inequality, which, in a more general form, is a consequence of the martingale property of the likelihood-ratio [e.g., Jiang (2010), pages 244–246]: suppose that Y_1 is a subvector of a random vector Y. Let $p_\theta(\cdot)$ and $p_{1,\theta}(\cdot)$ denote the p.d.f.'s of Y and Y_1, respectively, with respect to a σ-finite measure ν, under the parameter vector θ. For simplicity, suppose that $p_{\theta_0}, p_{1,\theta_0}$ are positive a.e. ν, and $\lambda(\cdot)$ is a positive, measurable function. Then, for any θ, we have

$$P_{\theta_0}\{p_{\theta_0}(Y) \le \lambda(Y_1)p_\theta(Y)|Y_1\} \le \lambda(Y_1)\frac{p_{1,\theta}(Y_1)}{p_{1,\theta_0}(Y_1)} \qquad \text{a.e. } \nu,$$

where P_{θ_0} denotes the probability distribution corresponding to p_{θ_0}.

5.2. *Quantifying the information loss.* On the other hand, the subset argument merely provides a method of proof for the consistency of the full-data MLE—it by no means suggests the subset-data MLE as a replacement for the full-data MLE. In fact, there is an information loss if such a replacement takes place. To quantify the information loss, assume the regularity conditions for exchanging the order of differentiation and integration. Then, the Fisher information matrix based on the full data can be expressed as

$$I_f(\theta) = -E_\theta\left\{\frac{\partial^2}{\partial\theta\,\partial\theta'}\log p_\theta(y)\right\}$$

$$= E_\theta\left[\left\{\frac{\partial}{\partial\theta}\log p_\theta(y)\right\}\left\{\frac{\partial}{\partial\theta}\log p_\theta(y)\right\}'\right] - E_\theta\left\{\frac{1}{p_\theta(y)}\frac{\partial^2}{\partial\theta\,\partial\theta'}p_\theta(y)\right\}$$

$$= I_{f,1}(\theta) - I_{f,2}(\theta).$$

Similarly, the information matrix based on the subset data can be expressed as $I_s(\theta) = I_{s,1}(\theta) - I_{s,2}(\theta)$, where $I_{s,j}(\theta)$ is $I_{f,j}(\theta)$ with y replaced by $y_{[1]}$, $j = 1, 2$

$[p_\theta(y_{[1]})$ denotes the p.d.f. (or p.m.f.) of $y_{[1]}]$. By conditioning on $y_{[1]}$, it can be shown that $I_{f,2}(\theta) = I_{s,2}(\theta)$, while $I_{f,1}(\theta) \geq I_{s,1}(\theta)$. It follows that

$$(21) \qquad\qquad\qquad I_f(\theta) \geq I_s(\theta)$$

for all θ. Here the inequality means that the difference between the left side and right side is a nonnegative definite matrix. (21) suggests that the information contained in the full data is no less than that contained in the subset data, which, of course, is what one would expect. Furthermore, the information loss is given by

$$(22) \qquad\qquad I_f(\theta) - I_s(\theta) = E_\theta\left[\mathrm{Var}_\theta\left\{\frac{\partial}{\partial\theta}\log p_\theta(y)\Big|y_{[1]}\right\}\right],$$

where $\mathrm{Var}_\theta(\cdot|y_{[1]})$ denotes the conditional covariance matrix given $y_{[1]}$ under θ. The derivations of (21) and (22) are deferred to the supplementary material [Jiang (2013)]. It is seen from (22) that the information loss is determined by how much (additional) variation there is in the score function, $(\partial/\partial\theta)\log p_\theta(y)$, given the subset data $y_{[1]}$. In particular, if $y_{[1]} = y$, then the score function is a constant vector given $y_{[1]}$ (and θ); hence $\mathrm{Var}_\theta\{(\partial/\partial\theta)\log p_\theta(y)|y_{[1]}\} = 0$, thus, there is no information loss. In general, of course, the subset data $y_{[1]}$ is not chosen as y; therefore, there will be some loss of information.

Nevertheless, the information contained in the subset data is usually sufficient for identifying at least some of the parameters. Note that consistency is a relatively weak asymptotic property in the sense that various estimators, including those based on the subset data and, for example, the method of moments estimator of Jiang (1998), are consistent, even though they may not be asymptotically efficient. Essentially, for the consistency property to hold, one needs that, in spite of the potential information loss, the remaining information that the estimator is able to utilize grows with the sample size. For example, in the open problem (Sections 1 and 2), the information contained in y_{ii} grows at the rate of $m \wedge n$, which is sufficient for identifying μ; in the example of Section 4, the information contained in $y_{i,i}$ grows in the order of m_1, which is sufficient for identifying μ and $\psi^2 = \sigma^2 + \tau^2$, while the information contained in y_i grows at the rate of m_2, which is sufficient for identifying $\gamma = \sigma^2/\psi^2$. The identification of the "right" subset in a given problem is usually suggested by the nature of the parametrization. As mentioned (see the third paragraph of Section 3), a subset $y_{[1]}$ of independent data can always be found under the ANOVA GLMM (e.g., starting with the first observation, y_1, one finds the next observation such that it involves different random effects from those related to y_1, and so on). If the $y_{[1]}$ is such that $\liminf_{N\to\infty} \lambda_{\min}\{I_s(\theta)\} = \infty$, where $I_s(\theta)$ is as in (21) and $\lambda_{\min}$ denotes the smallest eigenvalue, the subset $y_{[1]}$ is sufficient for identifying all the components of θ; otherwise, more than one subsets are needed in order to identify all the parameters, as is shown in Section 4.

5.3. *Note on computation of MLE.* The subset argument offers a powerful tool for establishing consistency of the MLE in GLMM with crossed random effects. Note that the idea has not followed the traditional path of attempting to develop a (computational) procedure to approximate the MLE. In fact, this might explain why the computational advances over the past two decades [see, e.g., Jiang (2007), Section 4.1 for an overview] had not led to a major theoretical breakthrough for the MLE in GLMM in terms of asymptotic properties. Note that the MLE based on the subset data is a consistent estimator of the true parameter, and in that sense it is an approximation to the MLE based on the full data (two consistent estimators of the same parameter approximate each other). However, there is an information loss, as discussed in the previous subsection [see (22)], so one definitely wants to do better computationally.

One computational method that has been developed for computing the MLE in GLMMs, including those with crossed random effects, is Monte Carlo EM algorithm [e.g., McCullogh (1994, 1997), Booth and Hobert (1999)]. Here, however, we would like to discuss another, more recent, computational advance, known as *data cloning* [DC; Lele, Dennis and Lutscher (2007), Lele, Nadeem and Schmuland (2010)]. The DC uses the Bayesian computational approach for frequentist purposes. Let π denote the prior density function of θ. Then, one has the posterior,

$$(23) \qquad \pi(\theta|y) = \frac{p_\theta(y)\pi(\theta)}{p(y)},$$

where $p(y)$ is the integral of the numerator with respect to θ, which does not depend on θ. There are computational tools using the Markov chain Monte Carlo for posterior simulation that generate random variables from the posterior without having to compute the numerator or denominator of (23) [e.g., Gilks, Richardson and Spiegelhalter (1996); Spiegelhalter et al. (2004)]. Thus, we can assume that one can generate random variables from the posterior. If the observations y were repeated independently from K different individuals such that all of these individuals result in exactly the same data, y, denoted by $y^{(K)} = (y, \ldots, y)$, then the posterior based on $y^{(K)}$ is given by

$$(24) \qquad \pi_K\{\theta|y^{(K)}\} = \frac{\{p_\theta(y)\}^K \pi(\theta)}{\int \{p_\theta(y)\}^K \pi(\theta)\, d\theta}.$$

Lele, Dennis and Lutscher (2007), Lele, Nadeem and Schmuland (2010) showed that, as K increases, the right side of (24) converges to a multivariate normal distribution whose mean vector is equal to the MLE, $\hat{\theta}$, and whose covariance matrix is approximately equal to $K^{-1}I_f^{-1}(\hat{\theta})$. Therefore, for large K, one can approximate the MLE by the sample mean vector of, say, $\theta^{(1)}, \ldots, \theta^{(B)}$ generated from the posterior distribution (24). Denoted the sample mean by $\bar{\theta}^{(\cdot)}$, and call it the DC MLE. Furthermore, $I_f^{-1}(\hat{\theta})$ [see (21), (22)] can be approximated by K times the sample covariance matrix of $\theta^{(1)}, \ldots, \theta^{(B)}$. Torabi (2012) successfully applied the DC method to obtain the MLE for the salamander-mating data.

Note that the DC MLE is an approximate, rather than exact, MLE, in the sense that, as $K \to \infty$, the difference between $\bar{\theta}^{(\cdot)}$ and the exact MLE vanishes. Because we have established consistency of the exact MLE, it follows that the DC MLE is a consistent estimator as long as the number K increase with the sample size. More precisely, it is shown in the supplementary material [Jiang (2013)] that, for every $\varepsilon, \delta > 0$, there is $N_{\varepsilon,\delta}$ such that for every $n \geq N_{\varepsilon,\delta}$ and $B \geq 1$, there is $K(n, B)$ such that $\mathrm{P}\{|\bar{\theta}^{(\cdot)} - \theta_0| \geq \varepsilon\} < \delta$, if $K \geq K(n, B)$, where θ_0 is the true parameter vector. Note that, as far as consistency is concerned, one does not need that B goes to infinity. This makes sense because, as $K \to \infty$, the posterior (24) is becoming degenerate [the asymptotic covariance matrix is $K^{-1} I_{\mathrm{f}}^{-1}(\hat{\theta})$]; thus, one does not need a large B to "average out" the variation in $\bar{\theta}^{(\cdot)}$. Thus, from an asymptotic point of view, the result of the current paper provides a justification for the DC method.

More importantly, because B, K are up to one's choice, one can make sure that they are large enough so that there is virtually no information loss, as was concerned earlier. In this regard, a reasonably large B would reduce the sampling variation and therefore improve the DC approximation, and make the computation more efficient. See Lele, Nadeem and Schmuland (2010) for discussion on how to choose B and K from practical points of view.

As for the prior π, Lele, Nadeem and Schmuland (2010) only suggests that it be chosen according to computational convenience and be proper (to avoid improper posterior). Following the subset idea, an obvious choice for the prior would be the multivariate normal distribution with mean vector $\hat{\theta}_{\mathrm{s}}$, the subset-data MLE, and covariance matrix $I_{\mathrm{s}}^{-1}(\hat{\theta}_{\mathrm{s}})$ [defined above (21)]. Note that $I_{\mathrm{s}}(\theta)$ is much easier to evaluate than $I_{\mathrm{f}}(\theta)$. This would make the procedure more similar to the empirical Bayes than the hierarchical one. Nevertheless, the DC only uses the Bayesian computational tool, as mentioned.

5.4. *Regarding the limiting process.* In some applications of GLMM, the estimation of the random effects are of interest. There have also been developments in semiparametric GLM and nonparametric ANOVA. In those cases, the random effects are treated the same way as the fixed effects. As a result, the proof of the consistency results in those cases usually impose constraints on the ratio of the number of effects and number of observations falling in each cluster [e.g., Chen (1995), Jiang (1999), Wu and Liang (2004), and Wang, Tsai and Qu (2012)]. A major difference exists, however, between the case of clustered data (e.g., Example 2) and that with crossed random effects (e.g., Example 1) in that, in the latter case, the data cannot be divided into independent groups (with the number of groups increasing with the sample size). Furthermore, the necessary constraints are very different depending on the interest of estimation. Consider, for example, a very simple case of linear mixed model, $y_{ij} = \mu + u_i + v_j + e_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$, where the u_i's and v_j's are random effects, and e_{ij}'s are errors. Assume, for simplicity, that all the random effects and errors are i.i.d. $N(0, 1)$, so that μ is the

only unknown parameter. Suppose that $n \to \infty$, while m is fixed, say, $m = 1$. In this case, $\bar{y}_{1.} = n^{-1} \sum_{j=1}^{n} y_{1j} = \mu + u_1 + \bar{v}_. + \bar{e}_{1.}$ is a consistent estimator of the cluster mean, $\mu_1 = \mu + u_1$. On the other hand, the MLE of μ, which is also $\bar{y}_{1.}$, is inconsistent (because it converges in probability to $\mu + u_1$, which is not equal to μ with probability one). Note that here the ratio of the number of effects and number of observations in the cluster is $2/n$. Apparently, this is sufficient for consistently estimating the mixed effect $\mu + u_1$, but not the fixed effect μ. One might suspect that the case $m = 1$ is somewhat extreme, as μ and u_1 are "inseparable"; but it does not matter. In fact, for any $m \geq 1$, as long as it is fixed, the MLE of μ is $\bar{y}_{..} = (mn)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n} y_{ij} = \mu + \bar{u}_. + \bar{v}_. + \bar{e}_{..}$, which converges in probability to $\mu + \bar{u}_.$ as $n \to \infty$, and $\mu + \bar{u}_. \neq \mu$ with probability one. Thus, the only way that the MLE of μ can be consistent is to have both m and n go to ∞.

The example also helps to explain why it is necessary to consider the limiting process $m \wedge n \to \infty$, instead of something else, in the open problem. The result of Theorem 1 shows that $m \wedge n \to \infty$ is also sufficient for the consistency of the MLE. In fact, from the proof of Theorem 1 it follows that, for large m, n, we have with probability tending to one that the conditional probability that $p_\mu(y) \leq p_{\mu+\varepsilon}(y)$ given $y_{[1]}$ is bounded by $\gamma^{m \wedge n}$ for some constant $0 < \gamma < 1$. The corresponding upper bound under Theorem 3 is $e^{-\lambda m_a}$ for some constant $\lambda > 0$, where m_a is the number of independent vectors in the subset $y_{[1]}$, and a similar result holds under Theorem 4 with the upper bound being $\exp[-\lambda m_a \{1 + o(1)\}]$. The assumption of Theorem 3, namely, (A1), makes sure that $m_* = \min_{1 \leq a \leq b} m_a \to \infty$ as the sample size increases; the assumptions of Theorem 4, namely, (A1) and (B3), make sure that, in addition, the $o(1)$ in the above vanishes as $m_* \to \infty$.

Although estimation of the random effects is not an objective of this paper, in some cases this is of interest. For example, one may consider estimating the conditional mean of y_{ij} given u_i in the open problem (which may correspond to the conditional probability of successful mating with the ith female in the salamander problem). As mentioned, the data are not clustered in this case; in other words, all the data are in the same cluster, so the ratio of the number of effects over the number of observations is $(1 + m + n)/mn = m^{-1} + n^{-1} + (mn)^{-1}$, which goes to zero as $m \wedge n \to \infty$. It is easy to show that $\bar{y}_{i.} = n^{-1} \sum_{j=1}^{n} y_{ij}$ is a consistent estimator of $E_\mu(y_{ij}|u_i) = E\{h(\mu + u_i + \eta)\}$, where $h(x) = e^x/(1 + e^x)$ and the (unconditional) expectation is with respect to $\eta \sim N(0, 1)$, $1 \leq i \leq m$. Similarly, $\bar{y}_{.j} = m^{-1} \sum_{i=1}^{m} y_{ij}$ is a consistent estimator of $E_\mu(y_{ij}|v_j) = E\{h(\mu + \xi + v_j)\}$, where the (unconditional) expectation is with respect to $\xi \sim N(0, 1)$, $1 \leq j \leq n$.

SUPPLEMENTARY MATERIAL

Supplementary material (DOI: [10.1214/13-AOS1084SUPP](https://doi.org/10.1214/13-AOS1084SUPP); .pdf). The supplementary material is available online at http://anson.ucdavis.edu/~jiang/glmmmle.suppl.pdf.

REFERENCES

BOOTH, J. G. and HOBERT, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. Roy. Statist. Soc. B* **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.

BRESLOW, N. E. and LIN, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika* **82** 81–91. MR1332840

CHEN, H. (1995). Asymptotically efficient estimation in semiparametric generalized linear models. *Ann. Statist.* **23** 1102–1129. MR1353497

CRAMÉR, H. (1946). *Mathematical Methods of Statistics*. *Princeton Mathematical Series* **9**. Princeton Univ. Press, Princeton, NJ. MR0016588

DRUM, M. L. and MCCULLAGH, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics* **49** 677–689. MR1243484

GILKS, W. R., RICHARDSON, S. and SPIEGELHALTER, D. J., eds. (1996). *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London. MR1397966

JIANG, J. (1998). Consistent estimators in generalized linear mixed models. *J. Amer. Statist. Assoc.* **93** 720–729. MR1631373

JIANG, J. (1999). Conditional inference about generalized linear mixed models. *Ann. Statist.* **27** 1974–2007. MR1765625

JIANG, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York. MR2308058

JIANG, J. (2010). *Large Sample Techniques for Statistics*. Springer, New York. MR2675055

JIANG, J. (2013). Supplement to "The subset argument and consistency of MLE in GLMM: Answer to an open problem and beyond." DOI:10.1214/13-AOS1084SUPP.

JIANG, J. and ZHANG, W. (2001). Robust estimation in generalised linear mixed models. *Biometrika* **88** 753–765. MR1859407

KARIM, M. R. and ZEGER, S. L. (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics* **48** 631–644.

KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statistics* **22** 79–86. MR0039968

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. Springer, New York. MR1639875

LELE, S. R., DENNIS, B. and LUTSCHER, F. (2007). Data cloning: Easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol. Lett.* **10** 551–563.

LELE, S. R., NADEEM, K. and SCHMULAND, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *J. Amer. Statist. Assoc.* **105** 1617–1625. MR2796576

LIN, X. (1997). Variance component testing in generalised linear models with random effects. *Biometrika* **84** 309–326. MR1467049

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. MR1424603

MCCULLAGH, P. and NELDER, J. A. (1989). *Generalized Linear Models*, 2nd ed. Chapman & Hall, London.

MCCULLOCH, C. E. (1994). Maximum likelihood variance components estimation for binary data. *J. Amer. Statist. Assoc.* **89** 330–335.

MCCULLOCH, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92** 162–170. MR1436105

MCCULLOCH, C. E., SEARLE, S. R. and NEUHAUS, J. M. (2008). *Generalized, Linear, and Mixed Models*, 2nd ed. Wiley, Hoboken, NJ. MR2431553

SPIEGELHALTER, D. J., THOMAS, A., BEST, N. and LUNN, D. (2004). WinBUGS version 1.4 user manual. MRC Biostatistics Unit, Institute of Public Health, London.

SUTRADHAR, B. C. and RAO, R. P. (2003). On quasi-likelihood inference in generalized linear mixed models with two components of dispersion. *Canad. J. Statist.* **31** 415–435. MR2043151

TORABI, M. (2012). Likelihood inference in generalized linear mixed models with two components of dispersion using data cloning. *Comput. Statist. Data Anal.* **56** 4259–4265. MR2957869

WALD, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statistics* **20** 595–601. MR0032169

WANG, P., TSAI, G. F. and QU, A. (2012). Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *J. Amer. Statist. Assoc.* **107** 725–736.

WOLFOWITZ, J. (1949). On Wald's proof of the consistency of the maximum likelihood estimate. *Ann. Math. Statistics* **20** 601–602. MR0032170

WU, H. and LIANG, H. (2004). Backfitting random varying-coefficient models with time-dependent smoothing covariates. *Scand. J. Stat.* **31** 3–19. MR2042595

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, DAVIS
DAVIS, CALIFORNIA 95656
USA
E-MAIL: jiang@wald.ucdavis.edu