# MONTE CARLO LIKELIHOOD INFERENCE FOR MISSING DATA MODELS

By Yun Ju Sung and Charles J. Geyer

*University of Washington and University of Minnesota*

Abbreviated title: Monte Carlo Likelihood Asymptotics

We describe a Monte Carlo method to approximate the maximum likelihood estimate (MLE), when there are missing data and the observed data likelihood is not available in closed form. This method uses simulated missing data that are independent and identically distributed and independent of the observed data. Our Monte Carlo approximation to the MLE is a consistent and asymptotically normal estimate of the minimizer $\theta^*$ of the Kullback-Leibler information, as both a Monte Carlo sample size and an observed data sample size go to infinity simultaneously. Plug-in estimates of the asymptotic variance are provided for constructing confidence regions for $\theta^*$. We give Logit-Normal generalized linear mixed model examples, calculated using an R package that we wrote.

*AMS* 2000 *subject classifications.* Primary 62F12; secondary 65C05.

*Key words and phrases.* Asymptotic theory, Monte Carlo, maximum likelihood, generalized linear mixed model, empirical process, model misspecification

**1. Introduction.** Missing data (Little and Rubin, 2002) either arise naturally—data that might have been observed are missing—or are intentionally chosen—a model includes random variables that are not observable (called *latent* variables or *random effects*). A mixture of normals or a generalized linear mixed model (GLMM) is an example of the latter. In either case, a model is specified for the *complete data* $(x, y)$, where $x$ is *missing* and $y$ is *observed*, by their joint density $f_\theta(x, y)$, also called the *complete data likelihood* (when considered as a function of $\theta$). The maximum likelihood estimator (MLE) maximizes the marginal density $f_\theta(y)$, also called the *observed data likelihood* (when considered as a function of $\theta$). This marginal density is only implicitly specified by the complete data model, $f_\theta(y) = \int f_\theta(x, y)\, dx$, and is often not available in closed form. This is what makes likelihood inference for missing data difficult.

Many Monte Carlo schemes for approximating the observed data likelihood in a missing data model have been proposed. We simulate missing data, independent of the observed data, using ordinary (independent sample) Monte Carlo. Others simulate missing data, dependent on the observed data, using either ordinary Monte Carlo (Ott, 1979; Kong, Liu, and Wong, 1994) or Markov chain Monte Carlo (MCMC) (Lange and Sobel, 1991; Thompson and Guo, 1991; Gelfand and Carlin, 1993; Geyer, 1994b; Thompson, 2003). There are also many Monte Carlo schemes for maximum likelihood without approximating the observed data likelihood: stochastic approximation (Younes, 1988; Moyeed and Baddeley, 1991), Monte Carlo EM (Wei and Tanner, 1990; Guo and Thompson, 1992), and Monte Carlo Newton-Raphson (Penttinen, 1984). There are also non–Monte Carlo schemes for maximum likelihood without approximating the observed data likelihood: EM (Dempster, Laird, and Rubin, 1977) and analytic approximation (Breslow and Clayton, 1993). There are so many methods because each has its strength and weakness. In theory, MCMC works for any problem, but in practice for complicated problems MCMC needs much trial and error and one can never be sure it has worked. Ordinary Monte Carlo with importance sampling is much simpler. It may not work for very complicated problems, but one always knows whether or not it worked and how well. Non–Monte Carlo

schemes are relatively easy to implement, but only approximate the desired answer and do not have error estimates so one cannot know how bad the approximation is. All of these are useful for some, but not all, problems.

Here, we provide rigorous asymptotic theory where both data and Monte Carlo sample sizes go to infinity. This is different from Geyer (1994b) where only the Monte Carlo sample size goes to infinity. When both sample sizes go to infinity, two sources of variability need to be considered: one from sampling of the observed data and the other from Monte Carlo sampling of the missing data. How to combine these two sources of variability is complicated even for ordinary Monte Carlo, as described below.

Let the observed data $Y_1, \ldots, Y_n$ be independent and identically distributed (i. i. d.) from a density $g$, which is not assumed to be some $f_\theta$. That is, we do not assume the model is correctly specified, since an increase of generality makes the theory no more difficult. The MLE $\hat{\theta}_n$ is a maximizer of the log-likelihood

$$l_n(\theta) = \sum_{j=1}^{n} \log f_\theta(Y_j). \tag{1}$$

In our method, we generate an i. i. d. Monte Carlo sample $X_1, \ldots, X_m$, independent of $Y_1, \ldots, Y_n$, from an importance sampling density $h(x)$ and approximate $f_\theta(y)$ by

$$f_{\theta,m}(y) = \frac{1}{m} \sum_{i=1}^{m} \frac{f_\theta(X_i, y)}{h(X_i)}. \tag{2}$$

This makes heuristic sense because

$$f_{\theta,m}(y) \xrightarrow{as}_m E_h\left\{ \frac{f_\theta(X, y)}{h(X)} \right\} = f_\theta(y) \qquad \text{for each } y$$

by the strong law of large numbers. (The subscript $m$ on the arrow means as $m$ goes to infinity. Similarly, a subscript $m, n$ means as both $m$ and $n$ go to infinity.) Our estimate of $\hat{\theta}_n$ is the maximizer $\hat{\theta}_{m,n}$ of

$$l_{m,n}(\theta) = \sum_{j=1}^{n} \log f_{\theta,m}(Y_j), \tag{3}$$

an approximation to $l_n(\theta)$ with $f_{\theta,m}$ replacing $f_\theta$. We call $\hat{\theta}_{m,n}$ the Monte Carlo MLE (MCMLE). Note that the summands in (3) are dependent, since the same Monte Carlo sample $X_1, \ldots, X_m$ is used for each $\log f_{\theta,m}(Y_j)$.

Under the conditions of Theorem 2.3,

$$\hat{\theta}_{m,n} \approx \mathcal{N}\left(\theta^*, \frac{J^{-1}VJ^{-1}}{n} + \frac{J^{-1}WJ^{-1}}{m}\right), \tag{4}$$

for sufficiently large $m$ and $n$, where $\theta^*$ is the minimizer of the Kullback-Leibler information

$$K(\theta) = E_g \log \frac{g(Y)}{f_\theta(Y)}, \tag{5}$$

$J$ is minus the expectation of the second derivative of the log-likelihood, $V$ is the variance of the first derivative of the log-likelihood (score), and $W$ is the variance of the deviation of the score from its Monte Carlo approximation (given by (7) below). Under certain regularity conditions (Huber, 1967; White, 1982),

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta^*, \frac{J^{-1}VJ^{-1}}{n}\right). \tag{6}$$

We see that, in our method, $\hat{\theta}_{m,n}$ has nearly the same distribution when the Monte Carlo sample size $m$ is very large. If the model is correctly specified, that is, $g = f_{\theta_0}$, then $\theta^* = \theta_0$ and $J = V$, either of which is called Fisher information, and (6) becomes

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta^*, \frac{J^{-1}}{n}\right)$$

the familiar formula due to Fisher and Cramér. This replacement of $J^{-1}$ by the so-called "sandwich" $J^{-1}VJ^{-1}$ is the only complication arising from model misspecification.

The asymptotic variance of the MCMLE $\hat{\theta}_{m,n}$ in (4) consists of two terms: the first term (which is also the asymptotic variance of $\hat{\theta}_n$) reflects sampling variability of the observed data ($Y$'s) and the second term reflects the variability of the Monte Carlo sample ($X$'s). Increasing the Monte Carlo sample size $m$ can make the second term as small we please so that the MCMLE $\hat{\theta}_{m,n}$ is almost as good as the MLE

$\hat{\theta}_n$. In (4), $W$ is the only term related to the importance sampling density $h$ that generates the Monte Carlo sample. Choosing an $h$ that makes $W$ smaller makes $\hat{\theta}_{m,n}$ more accurate.

The asymptotic distribution of $\hat{\theta}_{m,n}$ in (4) is a convolution of two independent normal distributions. The proof of this is not simple, however, for three reasons. First, the finite sample terms from which these arise (the two terms in the right hand side of (9)) are dependent. Second, one of these is itself a sum of dependent terms, because of the reuse of the $X$'s. Third, our two sample sizes $m$ and $n$ tend to infinity simultaneously, and we must show that the result does not depend on the way in which $m$ and $n$ go to infinity.

**2. Asymptotics of $\hat{\theta}_{m,n}$.** In this section, we state theorems about strong consistency and asymptotic normality of the MCMLE $\hat{\theta}_{m,n}$. Proofs are in the appendix. Epi-convergence is described in Section 2.1. Epi-convergence of $K_{m,n}$ to $K$ is in Section 2.2. This implies consistency of $\hat{\theta}_{m,n}$. Asymptotic normality of $\hat{\theta}_{m,n}$ is in Section 2.3. Plug-in estimates of the asymptotic variance for constructing confidence regions for $\theta^*$ are in Section 2.4.

We use empirical process notation throughout. We let $P$ denote the probability measure induced by the importance sampling density $h$ and $\mathbb{P}_m$ denote the empirical measure induced by $X_1, \ldots, X_m$ (that are i. i. d. from $P$). Similarly, we let $Q$ denote the probability measure induced by the true density $g$ and $\mathbb{Q}_n$ denote the empirical measure induced by $Y_1, \ldots, Y_n$ (that are i. i. d. from $Q$). Given a measurable function $f : \mathcal{X} \mapsto \mathbb{R}$, we write $\mathbb{P}_m f(X)$ for the expectation of $f$ under $\mathbb{P}_m$ and $Pf(X)$ for the expectation under $P$. Similarly we use $\mathbb{Q}_n f(Y)$ and $Qf(Y)$. Note that $\mathbb{P}_m f(X) = \frac{1}{n}\sum_{i=1}^n f(X_i)$ is just another notation for a particular sample mean.

The Kullback-Leibler information in (5) can be written as

$$K(\theta) = Q \log \frac{g(Y)}{f_\theta(Y)},$$

its empirical version as

$$K_n(\theta) = \mathbb{Q}_n \log \frac{g(Y)}{f_\theta(Y)},$$

and our approximation to $K_n(\theta)$ as

$$K_{m,n}(\theta) = \mathbb{Q}_n \log \frac{g(Y)}{f_{\theta,m}(Y)},$$

with

$$f_{\theta,m}(y) = \mathbb{P}_m \frac{f_\theta(X, y)}{h(X)}.$$

Then

$$K_n(\theta) = \mathbb{Q}_n \log g(Y) - \frac{1}{n} l_n(\theta),$$

and

$$K_{m,n}(\theta) = \mathbb{Q}_n \log g(Y) - \frac{1}{n} l_{m,n}(\theta).$$

Thus the MLE $\hat{\theta}_n$, the maximizer of $l_n$, is also the minimizer of $K_n$ and the MCMLE $\hat{\theta}_{m,n}$, the maximizer of $l_{m,n}$, is also the minimizer of $K_{m,n}$. By Jensen's inequality $K(\theta) \geq 0$. This allows $K(\theta) = \infty$ for some $\theta$, but we assume $K(\theta^*)$ is finite. (This excludes only the uninteresting case of the function $\theta \mapsto K(\theta)$ being identically $\infty$.)

2.1. *Epi-convergence.* To get the convergence of $\hat{\theta}_{m,n}$ to $\theta^*$ we use *epi-convergence* of the function $K_{m,n}$ to the function $K$. Epi-convergence is a "one-sided" uniform convergence that was first introduced by Wijsman (1964, 1966), developed in optimization theory (Attouch, 1984; Aubin and Frankowska, 1990; Rockafellar and Wets, 1998), and used in statistics (Geyer, 1994b,a). It is weaker than uniform convergence yet insures the convergence of minimizers as the following proposition due to Attouch (1984, Theorem 1.10) describes.

PROPOSITION 2.1. *Let $X$ be a general topological space, $\{f_n\}$ a sequence of functions from $X$ to $\overline{\mathbb{R}}$ that epi-converges to $f$, and $\{x_n\}$ a sequence of points in $X$ satisfying $f_n(x_n) \leq \inf f_n + \epsilon_n$ with $\epsilon_n \downarrow 0$. Then for every converging subsequence $x_{n_k} \to x_0$*

$$f(x_0) = \inf f = \lim_k f_{n_k}(x_{n_k}).$$

The conditions that Wald (1949) imposed to get consistency of the MLE imply epi-convergence of $K_n$ to $K$ (when there are no missing data and no Monte Carlo).

If $f$ has a unique minimizer $x$, then $x$ is the only cluster point of the sequence $\{x_n\}$. Otherwise, there may be many cluster points, but all must minimize $f$. There may not be any convergent subsequence. If the sequence $\{x_n\}$ is in a compact set and $X$ is sequentially compact, however, there is always a convergent subsequence.

2.2. *Epi-convergence of $K_{m,n}$.* Now we state our theorem about epi-convergence of $K_{m,n}$.

THEOREM 2.2. *Let $\{ f_\theta(x,y) : \theta \in \Theta \}$, where $\Theta \subset \mathbb{R}^d$, be a family of densities with respect to a $\sigma$-finite measure $\mu \times \nu$ on $\mathcal{X} \times \mathcal{Y}$, let $X_1$, $X_2$, ... be i. i. d. from a probability distribution $P$ that has a density $h$ with respect to $\mu$, and let $Y_1$, $Y_2$, ... be i. i. d. from a probability distribution $Q$ that has a density $g$ with respect to $\nu$. Suppose*

(1) *$\Theta$ is a second countable topological space,*

(2) *for each $(x,y)$, the function $\theta \mapsto f_\theta(x,y)$ is upper semicontinuous on $\Theta$,*

(3) *for each $\theta$, there exists a neighborhood $B_\theta$ of $\theta$ such that*

$$Q \log \left\{ P \sup_{\phi \in B_\theta} \frac{f_\phi(X,Y)}{h(X)g(Y)} \right\} < \infty,$$

(4) *for each $\theta$, there exists a neighborhood $C_\theta$ of $\theta$ such that for any subset $B$ of $C_\theta$, the family of functions*

$$\left\{ \sup_{\phi \in B} \frac{f_\phi(\,\cdot\,,y)}{h(\,\cdot\,)g(y)} : y \in \mathcal{Y} \right\}$$

*is $P$-Glivenko-Cantelli,*

(5) *for each $\theta$, the family of functions $\{ f_\theta(\,\cdot\,|y)/h(\,\cdot\,) : y \in \mathcal{Y} \}$ is $P$-Glivenko-Cantelli.*

*Then $K_{m,n}$ epi-converges to $K$ with probability one.*

$P$-Glivenko-Cantelli, a set of measurable functions on which the uniform strong law of large numbers holds (van der Vaart and Wellner, 1996, page 81), is required in condition (4) and (5). Other conditions are similar to those of Theorem 1 in Geyer (1994b).

2.3. *Asymptotic Normality of $\hat{\theta}_{m,n}$.* We now state our theorem about asymptotic normality of $\hat{\theta}_{m,n}$. If the minimizer $\theta^*$ of $K$ is an interior point of $\Theta$, then

$$\nabla K(\theta^*) = 0$$

assuming $K$ is differentiable. Here $\nabla$ is used to mean differentiation with respect to $\theta$. Define the matrix norm

$$\|A\|_\infty = \max_{i,j} |a_{ij}|$$

for a matrix $A$ with components $a_{ij}$.

THEOREM 2.3. *Let $\{ f_\theta(x, y) : \theta \in \Theta \}$, where $\Theta \subset \mathbb{R}^d$, be a family of densities with respect to a $\sigma$-finite measure $\mu \times \nu$ on $\mathcal{X} \times \mathcal{Y}$, let $X_1$, $X_2$, ... be i. i. d. from a probability distribution $P$ that has a density $h$ with respect to $\mu$, and let $Y_1$, $Y_2$, ... be i. i. d. from a probability distribution $Q$ that has a density $g$ with respect to $\nu$. Suppose*

(1) *second partial derivatives of $f_\theta(y)$ with respect to $\theta$ exist and are continuous on the interior of $\Theta$ for all $y$,*

(2) *there is an interior point $\theta^*$ of $\Theta$ such that $Q\nabla \log f_{\theta^*}(Y) = 0$, $V = \operatorname{var}_Q \nabla \log f_{\theta^*}(Y)$ is finite and $J = -Q\nabla^2 \log f_{\theta^*}(Y)$ is finite and nonsingular,*

(3) *there exists a $\rho > 0$ such that $S_\rho = \{ \theta : |\theta - \theta^*| \le \rho \}$ is contained in $\Theta$ and $\mathcal{F}_1 = \{ \nabla^2 f_\theta(\,\cdot\,) : \theta \in S_\rho \}$ is $Q$-Glivenko-Cantelli,*

(4) *second partial derivatives of $f_\theta(x|y)$ with respect to $\theta$ exist for all $x$ and $y$,*

(5) $\mathcal{Y}$ *is a separable metric space,*

(6) $\mathcal{F}_2 = \left\{ f_{\theta^*}(\,\cdot\,|y)/h(\,\cdot\,) : y \in \mathcal{Y} \right\}$ *is P-Glivenko-Cantelli,*

(7) $\mathcal{F}_3 = \left\{ \nabla f_{\theta^*}(\,\cdot\,|y)/h(\,\cdot\,) : y \in \mathcal{Y} \right\}$ *is P-Donsker,*

(8) *the envelope function* $F$ *of* $\mathcal{F}_3$ *has a finite second moment,*

(9) $P\nabla f_{\theta^*}(X|y)/h(X) = 0$ *for each* $y$*,*

(10) $y \mapsto \nabla f_{\theta^*}(x|y)$ *is continuous on* $\mathcal{Y}$ *for each* $x$*,*

(11) $\mathcal{F}_4 = \left\{ \nabla^2 f_\theta(\,\cdot\,|y)/h(\,\cdot\,) : y \in \mathcal{Y},\, \theta \in S_\rho \right\}$ *is P-Glivenko-Cantelli*

(12) $P\nabla^2 f_\theta(X|y)/h(X) = 0$ *for each* $y$ *and* $\theta \in S_\rho$*,*

(13) *there is a sequence* $\hat{\theta}_{m,n}$ *which converges to* $\theta^*$ *in probability such that*

$$\sqrt{\min(m,n)}\, \nabla K_{m,n}(\hat{\theta}_{m,n}) \xrightarrow[m,n]{P} 0.$$

*Then*

$$W = \operatorname{var}_P Q \frac{\nabla f_{\theta^*}(X|Y)}{h(X)} \tag{7}$$

*is finite and*

$$\left( \frac{V}{n} + \frac{W}{m} \right)^{-1/2} J \left( \hat{\theta}_{m,n} - \theta^* \right) \xrightarrow[m,n]{\mathcal{L}} \mathcal{N}(0, I). \tag{8}$$

Donsker, a set of measurable functions on which the "uniform" central limit theorem holds (van der Vaart and Wellner, 1996, page 81), is required for $\mathcal{F}_3$ and Glivenko-Cantelli for $\mathcal{F}_1$, $\mathcal{F}_2$, and $\mathcal{F}_4$. Conditions (1) through (3) and (13) are similar to the usual regularity conditions for asymptotic normality of the MLE, which can be found, for example, in Ferguson (1996, Chapter 18).

Note $\mathcal{F}_3$ is a family of vector-valued functions and $\mathcal{F}_1$ and $\mathcal{F}_4$ are families of matrix-valued functions. A class of vector-valued functions $f : \mathcal{X} \mapsto \mathbb{R}^d$ is defined to be Glivenko-Cantelli or Donsker if each of the classes of components $f_i : \mathcal{X} \mapsto \mathbb{R}$

with $f = (f_1, \ldots, f_d)$ ranging over $\mathcal{F}$ is Glivenko-Cantelli or Donsker (van der Vaart, 1998, page 270).

Under smoothness conditions imposed in this theorem, the asymptotics of $\hat{\theta}_{m,n}$ arises from the asymptotics of

$$\nabla K_{m,n}(\theta^*) = -\mathbb{Q}_n \nabla \log f_{\theta^*}(Y) - \mathbb{Q}_n \nabla \log \mathbb{P}_m f_{\theta^*}(X|Y)/h(X). \tag{9}$$

Note the two terms on the right hand side are dependent and the summands in the second term are dependent, which indicates the complexity of this problem and why the usual asymptotic arguments (requiring only the usual regularity conditions) do not work here.

The asymptotics for the first term on the right in (9) follows from the central limit theorem. The following diagram outlines the proof of the asymptotics for the second term (Lemma B.3).

$$\begin{array}{ccc} \sqrt{m}\,\mathbb{Q}_n \nabla \log \mathbb{P}_m f_{\theta^*}(X|Y)/h(X) & \xrightarrow{\ m\ } & \mathbb{Q}_n \mathbb{G}_P \nabla f_{\theta^*}(X|Y)/h(X) \\ & \searrow^{m,n} & \Big\downarrow n \\ & & Q\,\mathbb{G}_P \nabla f_{\theta^*}(X|Y)/h(X) \end{array} \tag{10}$$

The empirical process $\sqrt{m}(\mathbb{P}_m - P)$ converges in distribution to a tight Gaussian process $\mathbb{G}_P$. Applying the almost sure representation theorem to this convergence in distribution, for each $\omega$ in the almost sure representation, the upper left term in (10) goes to a constant, as first $m \to \infty$ then $n \to \infty$, and this makes the term asymptotically independent of the first term on the right in (9). The same representation also shows the asymptotic distribution for the term is the lower right term in (10).

2.4. *Plug-in Estimates for $J$, $V$, and $W$.* We can construct a confidence region for $\theta^*$ using (8) (or equivalently (4)). If we can evaluate the integrals defining $J$,

$V$, and $W$, then we may use those integrals with $\hat{\theta}_{m,n}$ plugged in for $\theta^*$ to estimate them, assuming enough continuity. Often we cannot evaluate the integrals or do not know $g$. Then we use their sample versions: sample variance instead of variance and sample mean instead of expectation

$$\begin{aligned}
\widehat{J}_{m,n} &= -\frac{1}{n}\sum_{j=1}^{n} \nabla^2 \log f_{\hat{\theta}_{m,n}}(Y_j) \\
\widehat{V}_{m,n} &= \frac{1}{n}\sum_{j=1}^{n} \Big\{\nabla \log f_{\hat{\theta}_{m,n}}(Y_j)\Big\} \Big\{\nabla \log f_{\hat{\theta}_{m,n}}(Y_j)\Big\}^T \\
\widehat{W}_{m,n} &= \frac{1}{m}\sum_{i=1}^{m} \widehat{S}_i \widehat{S}_i^T
\end{aligned} \tag{11}$$

where

$$\widehat{S}_i = \frac{1}{n}\sum_{j=1}^{n} \frac{\nabla f_{\hat{\theta}_{m,n}}(X_i|Y_j)}{h(X_i)}. \tag{12}$$

The resulting variance estimate $\widehat{J}_{m,n}^{-1}(\widehat{V}_{m,n}/n + \widehat{W}_{m,n}/m)\widehat{J}_{m,n}^{-1}$ is in a form often referred to as the "sandwich estimator" (Liang and Zeger, 1986).

Recall, however, we started this paper with the case where $f_\theta(y)$ is not available in closed form. Then the marginal in (11) and conditional in (12) would not be closed form. So we use $f_{\theta,m}(y)$ given by (2) instead of the unknown marginal $f_\theta(y)$, and using

$$f_{\theta,m}(x|y) = \frac{f_\theta(x,y)}{f_{\theta,m}(y)},$$

we use

$$\nabla f_{\theta,m}(x|y) = \frac{\nabla f_\theta(x,y)}{f_{\theta,m}(y)} - \frac{f_\theta(x,y)\nabla f_{\theta,m}(y)}{f_{\theta,m}(y)^2}.$$

**3. Logit-Normal GLMM Examples.** In a Logit-Normal GLMM, the observed data is a vector $y$ whose components are conditionally independent given the missing data (also called random effects) vector $b$ with

$$y_i|b \sim \text{Bernoulli}(\text{logit}^{-1}(\eta_i)), \tag{13}$$

where

$$\eta = X\beta + Zb, \tag{14}$$

and $b$ is unconditionally jointly mean-zero multivariate normal. In (14) $X$ and $Z$ are known matrices (the design matrices for fixed and random effects, respectively) and $\beta$ is an unknown fixed effects parameter vector.

The unknown parameters to be estimated are $\beta$ and parameters determining the variance matrix of $b$. Usually this variance matrix has simple structure and involves only a few parameters. We have written an R package `bernor` that implements the methods of this paper for a class of Logit-Normal GLMM. The package and more detailed descriptions of its application to the examples in this paper are on the webpage `www.stat.umn.edu/geyer/bernor`. Our package restricts to a case where the variance matrix of $b$ is diagonal, so the random effects are (unconditionally) independent.

3.1. *Data from McCulloch's Model.* We use a data set given by Booth and Hobert (1999, Table 2) that was simulated using a model from McCulloch (1997). This model corresponds to a Logit-Normal GLMM with one-dimensional $\beta$ and $b$ in (14), and its log likelihood can be calculated exactly by numerical integration. The observed data consist of 10 i. i. d. pieces, each with length 15. The parameters that generated the data are $\beta = 5$ and $\sigma = \sqrt{1/2}$.

We generated a Monte Carlo sample of size $10^4$, using a standard normal distribution as $h$, obtained the MCMLE and plug-in estimates given by (11), and constructed a nominal 95% confidence ellipse (shown in Figure 1). This ellipse contains the simulation truth. Figure 1 also shows the dotted ellipse where the asymptotic variance is calculated exactly using the theoretical expected Fisher information and $W$, instead of plug-in estimates. Our estimate ($\hat{\beta}_{m,n} = 6.15, \hat{\sigma}_{m,n} = 1.31$) is close to the MLE ($\hat{\beta}_n = 6.13, \hat{\sigma}_n = 1.33$). But neither is close to the truth, and the solid and dotted ellipses are different, indicating that plug-in estimates are not close to their expected values. This merely indicates that sample size $n = 10$ is too small to

apply asymptotics.

To demonstrate our asymptotic theory, we did a simulation study using the same model with sample sizes $n = 500$ and $m = 100$. (We have chosen these sample sizes so that sampling and Monte Carlo variability, the two terms that make up the variance in (4), are roughly the same size.) Figure 2 gives the scatter plot of 100 MCMLE's. The solid ellipse is an asymptotic 95% coverage ellipse using the theoretical expected Fisher information and $W$. The dashed ellipse is what we would have if we had very large Monte Carlo sample size $m$, leaving $n$ the same. The solid ellipse contains 92 out of 100 points, thus the asymptotics appear to work well at these sample sizes. However, as the dashed curve shows, even if we were to use a Monte Carlo sample size $m$ so large that the Monte Carlo error is negligible, the (non–Monte Carlo) sampling variability of the estimator would still be large, even at $n = 500$. The estimator of the fixed effect $\mu$ is fairly precise (about one and a half significant figure accuracy), but the estimator of the random effect scale parameter $\sigma$ has zero significant figure accuracy. It appears that very large (data) sample sizes would be necessary for scientifically useful inference about this model.

3.2. *The Salamander Data.* We use the data in McCullagh and Nelder (1989, Section 14.5) that were obtained from a salamander mating experiment and have been analyzed many times (see Booth and Hobert, 1999, for one analysis and citations of others). We use "Model A" of Karim and Zeger (1992). According to this model, the data have only 3 i. i. d. pieces. Thus we assume $J = V$ (no model misspecification), since $n = 3$ is too small to obtain non-singular estimates of $V$.

We obtained the MCMLE and standard errors, using $m = 10^4$ and also using $m = 10^7$ (shown in Table 1). The MLE given by Booth and Hobert (1999) is also shown in Table 1 (we have independently verified using MCMC that their MLE appears to be correct up to three significant figures). Our MCMLE's agree qualitatively but not quantatitively with the MLE. Increasing $m$ would improve the agreement.

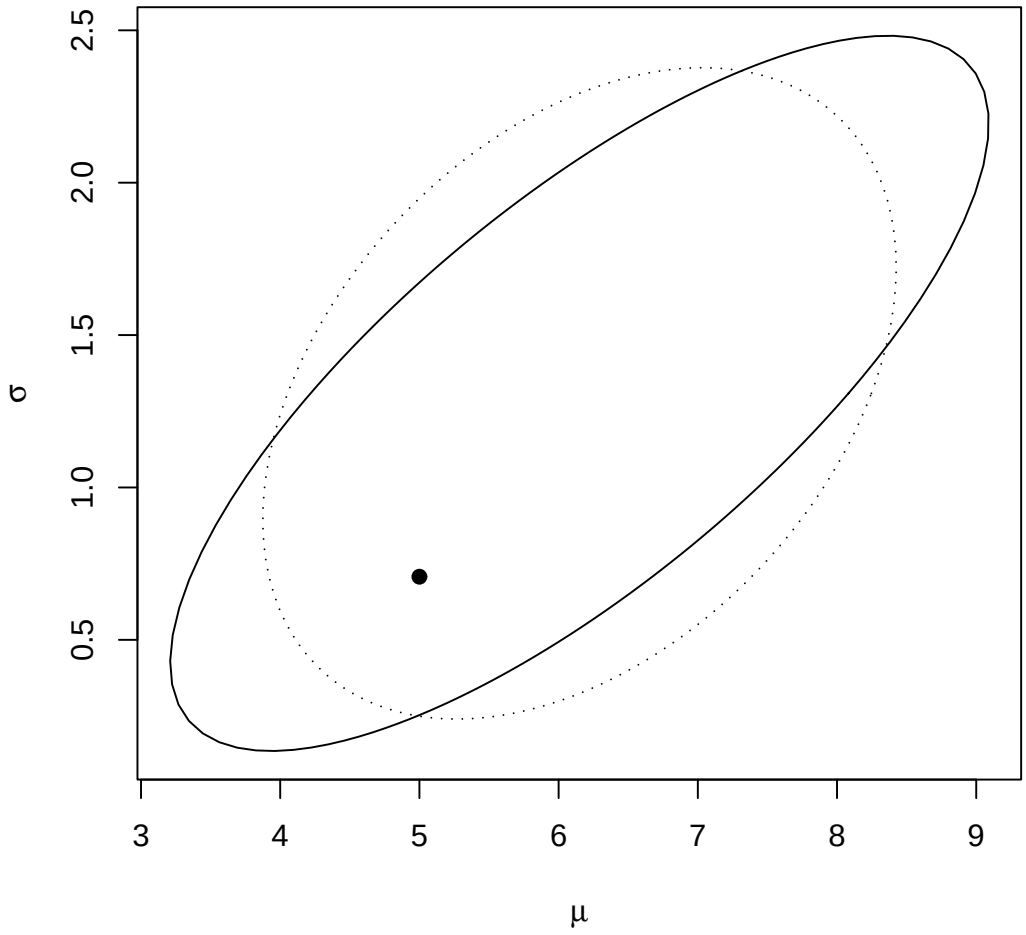



Figure 1: Nominal 95% confidence ellipse for our analysis of the Booth and Hobert data using $m = 10^4$. The solid dot is the "simulation truth" parameter value. The solid ellipse uses plug-in estimates of $J$, $V$, and $W$, whereas the dotted ellipse uses the Fisher information and $W$.

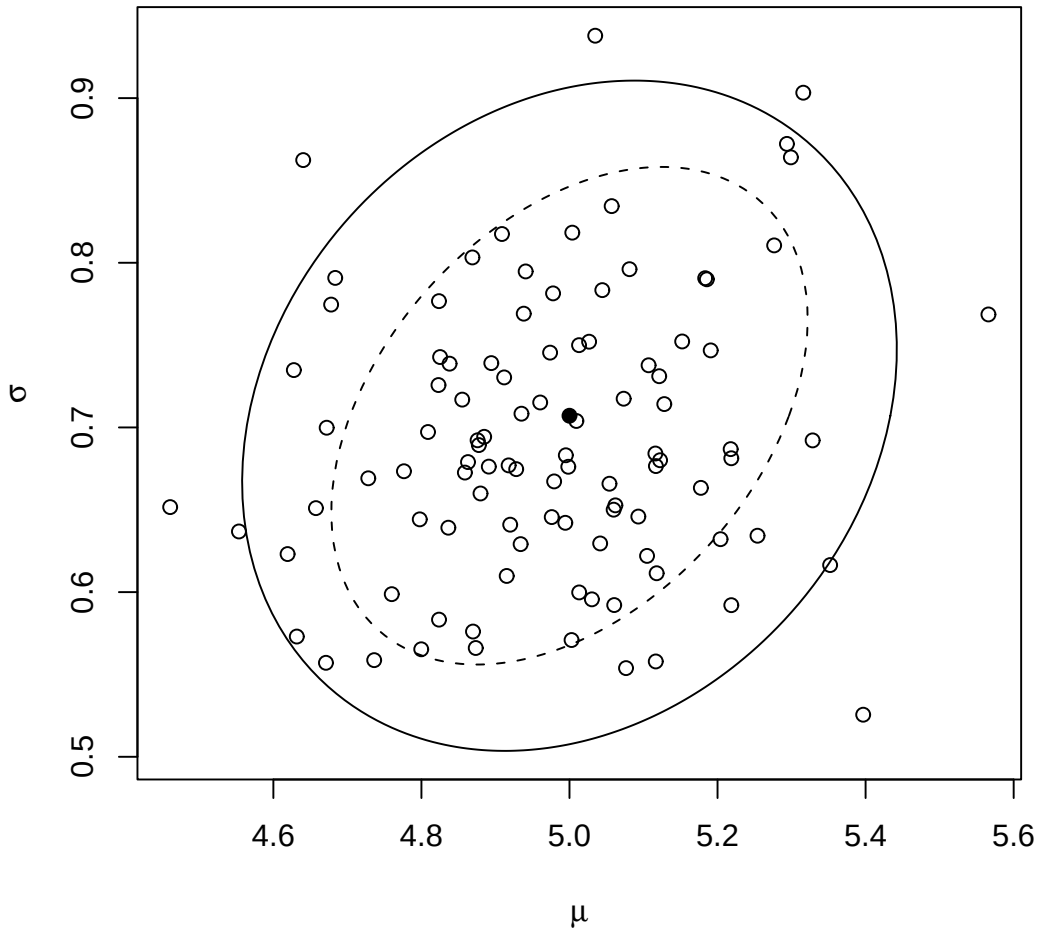



Figure 2: Simulated MLE with asymptotic 95% coverage ellipse (solid curve). The solid dot is the "simulation truth" parameter value (see text). Hollow dots are the MCMLE's for 100 simulated data sets, using sample sizes $n = 500$ and $m = 100$. The dashed curve is what the 95% coverage ellipse would be if we set $m$ to infinity.

Table 1: The MCMLE's and Standard Errors for the Salamander Data Set, using Sample Size $m = 10^4$ and $m = 10^7$. The MLE from Booth and Hobert (1999) is provided for comparison.

| | | $\beta_{R/R}$ | $\beta_{R/W}$ | $\beta_{W/R}$ | $\beta_{W/W}$ | $\sigma_f$ | $\sigma_m$ |
|---|---|---|---|---|---|---|---|
| MCMLE ($m = 10^4$) | est | 0.98 | 0.19 | −1.90 | 0.49 | 0.84 | 0.86 |
| | SE | 0.29 | 0.32 | 0.33 | 0.28 | 0.15 | 0.18 |
| MCMLE ($m = 10^7$) | est | 1.00 | 0.53 | −1.78 | 1.27 | 1.10 | 1.17 |
| | SE | 0.35 | 0.33 | 0.36 | 0.53 | 0.20 | 0.28 |
| MLE | est | 1.03 | 0.32 | −1.95 | 0.99 | 1.18 | 1.12 |

We included this example, even though our method does not perform well, because it is important for users to know that there is a level of complexity that our method does not handle as well as MCMC. That having been said, we want to point out virtues of our method. It is ordinary (independent sample) Monte Carlo, thus simpler to implement and easier to understand than MCMC. Also, it always provides accurate standard errors, unlike MCMC where convergence proofs are very difficult and rarely done in practice. Furthermore, this example has been considered difficult to analyze because its likelihood involves a 20-dimensional integral. For many less complicated data sets, our method would work.

**4. Discussion.** We have described a Monte Carlo method to approximate the MLE, when there are missing data and the observed data likelihood is not available in closed form. The MLE converges to the minimizer $\theta^*$ of the Kullback-Leibler information, which is the true parameter value when the model is correctly specified. We have proved that our estimate MCMLE is a consistent and asymptotically normal estimate of $\theta^*$ as both Monte Carlo and observed data sample sizes go to infinity simultaneously. Plug-in estimates of the asymptotic variance are provided in (11) for constructing confidence regions for $\theta^*$. We applied our method to Logit-Normal

GLMM examples.

In practice, a statistical model $f_\theta$ is often chosen only for mathematical convenience and may contain simplistic and unrealistic assumptions. However, it is usually possible to simulate i. i. d. data $Y$'s from a more realistic model $g$. We have presented the theory so that it can be used for the study of model misspecification in missing data models. The theory applies whether the $Y$'s are a Monte Carlo sample or real data. In either case we can estimate $\theta^*$ using $\hat{\theta}_{m,n}$ and know what accuracy we have. By comparing $f_{\hat{\theta}_{m,n}}$ (an estimate of $f_{\theta^*}$, the "best" approximation to $g$ in the model) with $g$, we can assess model validity as whether the particular model is reasonable for approximating the truth or how its simplifying assumptions influence scientific conclusions.

In our scheme, when the observed data are i. i. d., a whole Monte Carlo sample $X$'s of missing data is used $n$ times—for each $f_{\theta,m}(y_j)$, $1 \le j \le n$. Instead suppose the Monte Carlo sample $X$ is used only once, split into $n$ groups, each group used for approximating one $f_\theta(y_j)$ (making the asymptotics simpler). Then the resulting estimate has the same form of asymptotic variance as in (8) (or equivalently (4)), with $W$ replaced by

$$\widetilde{W} = Q \operatorname{var}_h \left\{ \frac{\nabla f_{\theta^*}(X|Y)}{h(X)} \right\}.$$

By Jensen's inequality, $\widetilde{W} \ge W$. Thus using the $X$'s $n$ times makes $\hat{\theta}_{m,n}$ more accurate.

The Monte Carlo method we analyzed in this paper can be extended in two ways. First, the importance sampling density $h$ can be allowed to depend on the observed data. Second, the simulated missing data can be allowed to be a Markov chain. In practice, both of these extensions are commonly used together. For example, the conditional density $f_{\theta_0}(x|y)$ of $X$ given the observed data $y$ is used as an importance sampling, from which simulating i. i. d. sample is often impossible and MCMC is necessary (Lange and Sobel, 1991; Thompson and Guo, 1991; Thompson, 2003). Providing theory for either of these extensions is an important open question.

# APPENDIX

**A. Proof of Theorem 2.2.** We first prove the following lemmas, then prove Theorem 2.2.

LEMMA A.1. *Under condition (5) of Theorem 2.2, as $m \to \infty$ and $n \to \infty$*

$$K_{m,n}(\theta) \to K(\theta)$$

*with probability one for each $\theta$.*

PROOF. Since

$$\frac{f_{\theta,m}(y)}{f_\theta(y)} - 1 = (\mathbb{P}_m - P)\frac{f_\theta(\cdot, y)}{h(\cdot)f_\theta(y)} = (\mathbb{P}_m - P)\frac{f_\theta(\cdot|y)}{h(\cdot)},$$

by condition (5)

$$\left\| \frac{f_{\theta,m}(\cdot)}{f_\theta(\cdot)} - 1 \right\|_{\mathcal{Y}} \xrightarrow{au}_m 0$$

by Lemma 1.9.2 in van der Vaart and Wellner (1996). That is, for every $\epsilon > 0$, there exists a measurable set $A$ such that $\Pr(A) \geq 1 - \epsilon$ and

$$\left\| \frac{f_{\theta,m}(\cdot)}{f_\theta(\cdot)} - 1 \right\|_{\mathcal{Y}} \longrightarrow_m 0$$

uniformly on $A$. So for every $\epsilon_1 > 0$ there exists $M \in \mathbb{N}$ such that

$$1 - \epsilon_1 \leq \frac{f_{\theta,m}(y)}{f_\theta(y)} \leq 1 + \epsilon_1, \qquad \text{for } m \geq M \text{ and all } y \in \mathcal{Y}$$

uniformly on $A$. Thus

$$\log(1 - \epsilon_1) \leq \log \frac{f_{\theta,m}(y)}{f_\theta(y)} \leq \log(1 + \epsilon_1)$$

for $m \geq M$ and all $y \in \mathcal{Y}$ uniformly on $A$. So

$$\left\| \log \frac{f_{\theta,m}(\cdot)}{f_\theta(\cdot)} \right\|_{\mathcal{Y}} \longrightarrow_m 0$$

uniformly on $A$. That is, for any $\epsilon_2 > 0$ there exists $M_1 \in \mathbb{N}$ such that

$$\sup_{y \in \mathcal{Y}} \left| \log \frac{f_{\theta,m}(y)}{f_\theta(y)} \right| \leq \epsilon_2, \qquad \omega \in A, m \geq M_1$$

and hence

$$|K_{m,n}(\theta) - K_n(\theta)| = \left| \frac{1}{n} \sum_{j=1}^{n} \log \frac{f_\theta(Y_j)}{f_{\theta,m}(Y_j)} \right| \leq \epsilon_2,$$

for $\omega \in A$, $m \geq M_1$, and all $n \in \mathbb{N}$. Hence

$$\sup_{n \in \mathbb{N}} |K_{m,n}(\theta) - K_n(\theta)| \xrightarrow{au}_m 0. \tag{15}$$

Since

$$K_n(\theta) \xrightarrow{as}_n K(\theta) \tag{16}$$

by the strong law of large numbers, the result follows from applying the triangle inequality to (15) and (16). □

LEMMA A.2. *Under conditions (3) and (4) of Theorem 2.2,*

$$\liminf_{(m,n) \to (\infty,\infty)} \inf_{\phi \in B} K_{m,n}(\phi) \geq -Q \log P \sup_{\phi \in B} \frac{f_\phi(X,Y)}{h(X)g(Y)} \tag{17}$$

*with probability one for each subset $B$ of $B_\theta \cap C_\theta$.*

PROOF. By condition (3) the right hand side of (17) is not $-\infty$.

$$\begin{aligned}
\inf_{\phi \in B} K_{m,n}(\phi) &= \inf_{\phi \in B} -\mathbb{Q}_n \log \mathbb{P}_m \frac{f_\phi(X,Y)}{h(X)g(Y)} \\
&= -\sup_{\phi \in B} \mathbb{Q}_n \log \mathbb{P}_m \frac{f_\phi(X,Y)}{h(X)g(Y)} \\
&\geq -\mathbb{Q}_n \log \mathbb{P}_m \sup_{\phi \in B} \frac{f_\phi(X,Y)}{h(X)g(Y)},
\end{aligned}$$

where the inequality follows from the logarithm function being increasing and the supremum operation being superadditive. By condition (4)

$$\left\| (\mathbb{P}_m - P) \sup_{\phi \in B} \frac{f_\phi(\cdot, y)}{h(\cdot)g(y)} \right\|_{\mathcal{Y}} \xrightarrow{au}_m 0,$$

that is, for every $\epsilon_1 > 0$ and $\epsilon_2 > 0$, there exist a measurable set $A$ and an $M \in \mathbb{N}$ such that $\Pr(A) \geq 1 - \epsilon_1$ and

$$\mathbb{P}_m \sup_{\phi \in B} \frac{f_\phi(\cdot, y)}{h(\cdot)g(y)} \leq P \sup_{\phi \in B} \frac{f_\phi(\cdot, y)}{h(\cdot)g(y)} + \epsilon_2$$

for all $m \geq M$, $y \in \mathcal{Y}$, and $\omega \in A$. So

$$\inf_{\phi \in B} K_{m,n}(\phi) \geq -\mathbb{Q}_n \log \left\{ P \sup_{\phi \in B} \frac{f_\phi(X, Y)}{h(X)g(Y)} + \epsilon_2 \right\}$$

for all $m \geq M$, $n \in \mathbb{N}$, and $\omega \in A$.

Now applying the strong law of large numbers to the right hand side, there exist a measurable set $A_2$ and an $N \in \mathbb{N}$ such that $\Pr(A_2) \geq 1 - \epsilon_3$ and

$$-\mathbb{Q}_n \log \left\{ P \sup_{\phi \in B} \frac{f_\phi(X, Y)}{h(X)g(Y)} + \epsilon_2 \right\} \geq -Q \log \left\{ P \sup_{\phi \in B} \frac{f_\phi(X, Y)}{h(X)g(Y)} + \epsilon_2 \right\} - \epsilon_4$$

for all $n \geq N$, and $\omega \in A_2$. Hence

$$\inf_{\phi \in B} K_{m,n}(\phi) \geq -Q \log \left\{ P \sup_{\phi \in B} \frac{f_\phi(X, Y)}{h(X)g(Y)} + \epsilon_2 \right\} - \epsilon_4$$

for all $m \geq M$, $n \geq N$, and $\omega \in A \cap A_2$. Since $\epsilon$'s were arbitrary, (17) holds almost uniformly. □

LEMMA A.3. *Under conditions (2) and (3) of Theorem 2.2, $K$ is lower semi-continuous.*

PROOF. Let $\theta$ be a point of $\Theta$ and $\{\theta_k\}$ a sequence in $\Theta$ converging to $\theta$.

$$\begin{aligned}
\limsup_{k \to \infty} Q \log \frac{f_{\theta_k}(\cdot)}{g(\cdot)} &= \lim_{n \to \infty} \sup_{k \geq n} Q \log P \frac{f_{\theta_k}(X, Y)}{h(X)g(Y)} \\
&\leq \lim_{n \to \infty} Q \log P \sup_{k \geq n} \frac{f_{\theta_k}(X, Y)}{h(X)g(Y)} \\
&= Q \log P \limsup_{k \to \infty} \frac{f_{\theta_k}(X, Y)}{h(X)g(Y)},
\end{aligned}$$

where the second equality follows from the monotone convergence theorem by condition (3).

$$\begin{aligned}\liminf_{k\to\infty} K(\theta_k) &= -\limsup_{k\to\infty} Q \log \frac{f_{\theta_k}(\cdot)}{g(\cdot)} \\ &\geq -Q \log P \limsup_{k\to\infty} \frac{f_{\theta_k}(X,Y)}{h(X)g(Y)} \\ &\geq -Q \log P \frac{f_\theta(X,Y)}{h(X)g(Y)} = K(\theta),\end{aligned}$$

where the last inequality follows from condition (2). □

Proof of Theorem 2.2. Let $(m_k, n_k)$ be a subsequence of $(m, n)$. We need to show

$$K \leq \text{e-lim inf}_k K_{m_k,n_k} \leq \text{e-lim sup}_k K_{m_k,n_k} \leq K,$$

which is equivalent to

$$K(\theta) \leq \sup_{B\in\mathcal{N}(\theta)} \liminf_{k\to\infty} \inf_{\phi\in B} K_{m_k,n_k}(\phi), \tag{18}$$

$$K(\theta) \geq \sup_{B\in\mathcal{N}(\theta)} \limsup_{k\to\infty} \inf_{\phi\in B} K_{m_k,n_k}(\phi), \tag{19}$$

where $\mathcal{N}(\theta)$ is the set of neighborhoods of the point $\theta$.

By condition (1) there is a countable basis $\mathcal{B} = \{B_1, B_2, \dots\}$ for the topology of $\Theta$. Choose a particular countable dense subset $\Theta_c = \{\theta_1, \theta_2, \dots\}$ by choosing $\theta_n \in B_n$ to satisfy

$$K(\theta_n) \leq \inf_{\phi\in B_n} K(\phi) + 1/n.$$

Let

$$\mathcal{N}_c(\theta) = \{B \in \mathcal{B} \cap \mathcal{N}(\theta) : B \subset B_\theta \cap C_\theta\}$$

where $B_\theta$ is given by condition (3) and $C_\theta$ is given by condition (4). The set $\mathcal{N}_c(\theta)$ is a countable neighborhood basis for each $\theta$. So the suprema over the uncountable set $\mathcal{N}(\theta)$ in (18) and (19) can be replaced by suprema over the countable set $\mathcal{N}_c(\theta)$.

We shall need

$$\limsup_{k\to\infty} K_{m_k,n_k}(\theta) \le K(\theta), \tag{20}$$

$$\liminf_{k\to\infty} \inf_{\phi\in B} K_{m_k,n_k}(\phi) \ge -Q \log P \sup_{\phi\in B} \frac{f_\phi(X,Y)}{h(X)g(Y)} \tag{21}$$

to hold simultaneously for all $\theta \in \Theta_c$ and all $B \in \bigcup_{\theta\in\Theta} \mathcal{N}_c(\theta)$ with probability one. Inequality (20) holds at each point $\theta$ with probability one by Lemma A.1, and inequality (21) holds at each subset $B \in \mathcal{N}_c(\theta)$ with probability one by Lemma A.2. Since $\Theta_c$ and $\bigcup_{\theta\in\Theta} \mathcal{N}_c(\theta)$ (which is a subset of $\mathcal{B}$) are countable and since a countable union of null sets (one exception set for each limit) is still a null set, we have (20) and (21) simultaneously on $\Theta_c$ and $\bigcup_{\theta\in\Theta} \mathcal{N}_c(\theta)$, respectively, with probability one.

First we establish (19). If $B \in \mathcal{B}$ and $\theta \in B \cap \Theta_c$, then by (20)

$$K(\theta) \ge \limsup_{k} K_{m_k,n_k}(\theta) \ge \limsup_{k\to\infty} \inf_{\phi\in B} K_{m_k,n_k}(\phi)$$

So

$$\inf_{\phi\in B\cap\Theta_c} K(\phi) \ge \limsup_{k\to\infty} \inf_{\phi\in B} K_{m_k,n_k}(\phi)$$

and

$$\sup_{B\in\mathcal{N}_c(\theta)} \inf_{\phi\in B\cap\Theta_c} K(\phi) \ge \sup_{B\in\mathcal{N}_c(\theta)} \limsup_{k\to\infty} \inf_{\phi\in B} K_{m_k,n_k}(\phi).$$

The left-hand side is equal to $K(\theta)$ by lower semicontinuity of $K$ (Lemma A.3) and by the construction of $\Theta_c$.

Now

$$\begin{aligned}
\sup_{B\in\mathcal{N}_c(\theta)} \liminf_{k\to\infty} \inf_{\phi\in B} K_{m_k,n_k}(\phi) &\ge \sup_{B\in\mathcal{N}_c(\theta)} -Q \log P \sup_{\phi\in B} \frac{f_\phi(X,Y)}{h(X)g(Y)} \\
&= -Q \log P \inf_{B\in\mathcal{N}_c(\theta)} \sup_{\phi\in B} \frac{f_\phi(X,Y)}{h(X)g(Y)} \\
&= -Q \log P \frac{f_\theta(X,Y)}{h(X)g(Y)} = K(\theta),
\end{aligned}$$

where the first inequality follows from (21), the first equality from the monotone convergence theorem, and the second equality from condition (2). So we have (18). □

**B. Proof of Theorem 2.3.** We break up the proof into several lemmas. Using Lemma B.3 and Lemma B.4 below, we shall find the asymptotics of the two terms on the right hand side in (9) to be

$$\begin{pmatrix} \sqrt{n}\,\mathbb{Q}_n \nabla \log f_{\theta^*}(Y) \\ \sqrt{m}\,\mathbb{Q}_n \nabla \log \mathbb{P}_m f_{\theta^*}(X|Y)/h(X) \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \begin{pmatrix} V & 0 \\ 0 & W \end{pmatrix}\right). \tag{22}$$

The asymptotics for the first term follows from the central limit theorem

$$\sqrt{n}\,\mathbb{Q}_n \nabla \log f_{\theta^*}(Y) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V). \tag{23}$$

The second term is asymptotically independent of the first by Lemma B.3 and asymptotically normally distributed by Lemma B.4. Lemma B.3 and Lemma B.4 are original. The others use standard arguments.

LEMMA B.1. *Let $Q$ denote a separable probability measure on a metric space and $\mathbb{Q}_n$ the empirical measure for an i. i. d. sample from $Q$. Then $\mathbb{Q}_n \xrightarrow{\mathcal{L}} Q$ almost surely.*

PROOF. Let $\mathcal{B}$ be a countable basis for the metric space, and let $\mathcal{A}$ be the set of all finite intersections of elements of $\mathcal{B}$. The latter is countable because $\mathcal{B}^n$ is countable for each $n$, and hence $\bigcup_{n=1}^{\infty} \mathcal{B}^n$ is countable.

Now for each $A \in \mathcal{A}$ we have $\mathbb{Q}_n(A) \to Q(A)$ by the strong law of large numbers. Hence, a countable union of null sets being a null set, this implies

$$\mathbb{Q}_n(A) \to Q(A), \qquad A \in \mathcal{A} \tag{24}$$

almost surely (the null set not depending on $A$). By Theorem 2.2 in Billingsley (1999), $\mathcal{A}$ is a convergence determining class, meaning (24) implies $\mathbb{Q}_n \xrightarrow{\mathcal{L}} Q$. Hence $\mathbb{Q}_n \xrightarrow{\mathcal{L}} Q$ almost surely. □

LEMMA B.2. *Under conditions (8) through (10) of Theorem 2.3, the function $y \mapsto k(\omega, y)$ is bounded and continuous for almost all $\omega$, where*

$$k(\omega, y) = \mathbb{G}_P(\omega)\left\{\frac{\nabla f_{\theta^*}(X|y)}{h(X)}\right\} \tag{25}$$

*and $\mathbb{G}_P$ is the tight Gaussian process in $l^\infty(\mathcal{F}_3)$ with zero mean and covariance function $E(\mathbb{G}_P f \cdot \mathbb{G}_P g) = Pfg - PfPg$.*

PROOF. From $\mathbb{G}_P$ being a random element in $l^\infty(\mathcal{F}_3)$, for each $\omega$

$$\|\mathbb{G}_P(\omega)\|_{\mathcal{F}_3} = \left\| \mathbb{G}_P(\omega) \frac{\nabla f_{\theta^*}(X|y)}{h(X)} \right\|_{\mathcal{Y}} = \|k(\omega, y)\|_{\mathcal{Y}} < \infty,$$

thus the function $y \mapsto k(\omega, y)$ is bounded.

The function $y \mapsto k(\omega, y)$ is continuous if and only if each $y \mapsto k_i(\omega, y)$ is continuous, where $k_i(\omega, y)$ is the $i$-th coordinate of $k(\omega, y)$. Let

$$\mathcal{F}_{3i} = \left\{ f_i : f = (f_1, \ldots, f_d) \in \mathcal{F}_3 \right\}.$$

From $\mathbb{G}_P$ being a tight Gaussian process, for almost all $\omega$ the function sample path $f \mapsto \mathbb{G}_P(\omega) f$ is $\rho_i$-continuous on $\mathcal{F}_{3i}$ (van der Vaart and Wellner, 1996, Section 1.5), where

$$\rho_i(f, g) = \left\{ E(\mathbb{G}_P f - \mathbb{G}_P g)^2 \right\}^{1/2} = \left\{ P(f - g)^2 \right\}^{1/2}, \qquad f, g \in \mathcal{F}_{3i}.$$

Let $\left[\nabla f_{\theta^*}(\,\cdot\,|y)/h(\,\cdot\,)\right]_i$ denote the $i$-th coordinate of $\nabla f_{\theta^*}(\,\cdot\,|y)/h(\,\cdot\,)$. If $y_n \to y$ in $\mathcal{Y}$ then

$$\begin{aligned} &\rho_i \left( \left[\nabla f_{\theta^*}(\,\cdot\,|y_n)/h(\,\cdot\,)\right]_i, \left[\nabla f_{\theta^*}(\,\cdot\,|y)/h(\,\cdot\,)\right]_i \right)^2 \\ &\qquad = P \left( \left[\nabla f_{\theta^*}(\,\cdot\,|y_n)/h(\,\cdot\,)\right]_i - \left[\nabla f_{\theta^*}(\,\cdot\,|y)/h(\,\cdot\,)\right]_i \right)^2 \le 4P(F_i^2) \end{aligned}$$

where $F_i$ is the $i$-th coordinate of $F$ in condition (8). So by the dominated convergence theorem and conditions (8) and (10)

$$\rho_i \left( \left\{ \nabla f_{\theta^*}(\,\cdot\,|y_n)/h(\,\cdot\,) \right\}_i, \left\{ \nabla f_{\theta^*}(\,\cdot\,|y)/h(\,\cdot\,) \right\}_i \right) \to 0,$$

and this shows the function $y \mapsto \left[\nabla f_{\theta^*}(\,\cdot\,|y)/h(\,\cdot\,)\right]_i$ from $\mathcal{Y}$ to $(\mathcal{F}_{3i}, \rho_i)$ is continuous. Then as the composition of the two continuous functions, the function $y \mapsto k_i(\omega, y)$ is continuous for almost all $\omega \in \Omega$. □

LEMMA B.3. *Under conditions (6), (7), and (9) of Theorem 2.3,*

$$\sqrt{m}\,\mathbb{Q}_n \nabla \log \mathbb{P}_m f_{\theta^*}(X|Y)/h(X) \xrightarrow{\mathcal{L}}_{m,n} Q\,\mathbb{G}_P \nabla f_{\theta^*}(X|Y)/h(X), \tag{26}$$

*and this limit is independent of the first term on the right hand side in (9), where $\mathbb{G}_P$ is a tight Gaussian process in $l^\infty(\mathcal{F}_3)$ with zero mean and covariance function $E(\mathbb{G}_P f \cdot \mathbb{G}_P g) = Pfg - PfPg$.*

PROOF. Conditions (6) and (7) say

$$\begin{aligned} \mathbb{P}_m &\xrightarrow{au} P \quad \text{in} \quad l^\infty(\mathcal{F}_2), \\ \mathbb{G}_m &\xrightarrow{\mathcal{L}^*} \mathbb{G}_P \quad \text{in} \quad l^\infty(\mathcal{F}_3), \end{aligned}$$

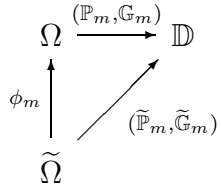
where $\mathbb{G}_m = \sqrt{m}(\mathbb{P}_m - P)$. By Slutsky's theorem (van der Vaart and Wellner, 1996, Example 1.4.7)

$$(\mathbb{P}_m, \mathbb{G}_m) \xrightarrow{\mathcal{L}^*} (P, \mathbb{G}_P) \quad \text{in} \quad \mathbb{D},$$

where

$$\mathbb{D} = l^\infty(\mathcal{F}_2) \times l^\infty(\mathcal{F}_3).$$

By the almost sure representation theorem (van der Vaart and Wellner, 1996, Theorem 1.10.4 and Addendum 1.10.5), letting $(\Omega, \mathcal{A}, \Pr)$ denote the probability space on which the $X_n$ and hence the $\mathbb{P}_n$ are defined ($\Pr$ can be taken to be $P^\infty$), there exist measurable and perfect functions $\phi_m$ on some probability space $(\widetilde{\Omega}, \widetilde{\mathcal{A}}, \widetilde{\Pr})$ such that the following diagram commutes




and $\Pr = \widetilde{\Pr} \circ \phi_m^{-1}$ and

$$(\widetilde{\mathbb{P}}_m, \widetilde{\mathbb{G}}_m) \xrightarrow{au} (\widetilde{\mathbb{P}}_\infty, \widetilde{\mathbb{G}}_\infty) \quad \text{in} \quad \mathbb{D}, \tag{27}$$

where

$$(\mathbb{P}_\infty, \mathbb{G}_\infty) = (P, \mathbb{G}_P),$$

where on the right hand side $P$ denotes a constant random element of $l^\infty(\mathcal{F}_2)$. We write also $\widetilde{P}$ instead of $\widetilde{\mathbb{P}}_\infty$ and $\widetilde{\mathbb{G}}_P$ instead of $\widetilde{\mathbb{G}}_\infty$.

The almost uniform convergence in (27) implies almost sure convergence. There exists $\widetilde{A}$ such that $\widetilde{\Pr}(\widetilde{A}) = 1$,

$$\|(\widetilde{\mathbb{P}}_m, \widetilde{\mathbb{G}}_m) - (\widetilde{P}, \widetilde{\mathbb{G}}_P)\|_{\mathbb{D}} \longrightarrow_m 0$$

on $\widetilde{A}$, and $\tilde{k}(\tilde{\omega}, \cdot\,) = k(\phi_\infty(\tilde{\omega}), \cdot\,)$ is bounded and continuous for $\tilde{\omega} \in \widetilde{A}$ (Lemma B.2), where $k$ is defined by (25). That is, on $\widetilde{A}$,

$$\begin{aligned}\|(\widetilde{\mathbb{P}}_m - \widetilde{P}) f_{\theta^*}(\cdot|y)/h(\cdot)\|_{\mathcal{Y}} &\longrightarrow_m 0 \\ \|(\widetilde{\mathbb{G}}_m - \widetilde{\mathbb{G}}_P) \nabla f_{\theta^*}(\cdot|y)/h(\cdot)\|_{\mathcal{Y}} &\longrightarrow_m 0\end{aligned}$$

Since the function $(s,t) \mapsto t/s$ is uniformly continuous on $[s_0, \infty) \times \mathbb{R}$ with $s_0 > 0$ and since $\widetilde{P} f_{\theta^*}(\,\cdot\,|y)/h(\,\cdot\,) = 1$,

$$\left\| \frac{\widetilde{\mathbb{G}}_m \nabla f_{\theta^*}(\cdot|y)/h(\cdot)}{\widetilde{\mathbb{P}}_m f_{\theta^*}(\cdot|y)/h(\cdot)} - \widetilde{\mathbb{G}}_P \nabla f_{\theta^*}(\cdot|y)/h(\cdot) \right\|_{\mathcal{Y}} \longrightarrow_m 0$$

on $\widetilde{A}$. Thus for every $\tilde{\omega} \in \widetilde{A}$ and every sequence $\{Y_1, Y_2, \dots\}$

$$\mathbb{Q}_n \frac{\widetilde{\mathbb{G}}_m(\tilde{\omega}) \nabla f_{\theta^*}(X|Y)/h(X)}{\widetilde{\mathbb{P}}_m(\tilde{\omega}) f_{\theta^*}(X|Y)/h(X)} \longrightarrow_m \mathbb{Q}_n \widetilde{\mathbb{G}}_P(\tilde{\omega}) \left\{ \frac{\nabla f_{\theta^*}(X|Y)}{h(X)} \right\} \tag{28}$$

uniformly in $n$.

Note that the right hand side of (28) is $\mathbb{Q}_n \tilde{k}(\tilde{\omega}, \cdot\,)$. Since $\tilde{k}(\tilde{\omega}, \cdot\,)$ is bounded and continuous, if $(H, \mathcal{B}, \mathrm{Qr})$ is the probability space on which the $Y_n$ are defined (Qr can be taken to be $Q^\infty$) by Lemma B.1, there exists a $B$ such that $\mathrm{Qr}(B) = 1$ and

$$\mathbb{Q}_n(\eta) \widetilde{\mathbb{G}}_P(\tilde{\omega}) \left\{ \frac{\nabla f_{\theta^*}(X|Y)}{h(X)} \right\} \longrightarrow_n Q\, \widetilde{\mathbb{G}}_P(\tilde{\omega}) \left\{ \frac{\nabla f_{\theta^*}(X|Y)}{h(X)} \right\} \tag{29}$$

for all $\tilde{\omega} \in \widetilde{A}$ and all $\eta \in B$.

Now combining (28) and (29)

$$\mathbb{Q}_n(\eta)\frac{\widetilde{\mathbb{G}}_m(\tilde{\omega})\nabla f_{\theta^*}(X|Y)/h(X)}{\widetilde{\mathbb{P}}_m(\tilde{\omega})f_{\theta^*}(X|Y)/h(X)} \longrightarrow_{m,n} Q\,\widetilde{\mathbb{G}}_P(\tilde{\omega})\left\{\frac{\nabla f_{\theta^*}(X|Y)}{h(X)}\right\} \tag{30}$$

for all $\tilde{\omega} \in \widetilde{A}$ and all $\eta \in B$. Even though we first let $m \to \infty$ and then $n \to \infty$, the limit would be the same no matter how $m$ and $n$ go to $\infty$ because of the uniformity in (28) (this can be shown by a triangle inequality). Hence

$$\mathbb{Q}_n\frac{\widetilde{\mathbb{G}}_m\nabla f_{\theta^*}(X|Y)/h(X)}{\widetilde{\mathbb{P}}_m f_{\theta^*}(X|Y)/h(X)} \xrightarrow{as}_{m,n} Q\,\widetilde{\mathbb{G}}_P\left\{\frac{\nabla f_{\theta^*}(X|Y)}{h(X)}\right\}$$

and this implies convergence in distribution

$$\mathbb{Q}_n\frac{\widetilde{\mathbb{G}}_m\nabla f_{\theta^*}(X|Y)/h(X)}{\widetilde{\mathbb{P}}_m f_{\theta^*}(X|Y)/h(X)} \xrightarrow{\mathcal{L}}_{m,n} Q\,\widetilde{\mathbb{G}}_P\left\{\frac{\nabla f_{\theta^*}(X|Y)}{h(X)}\right\}.$$

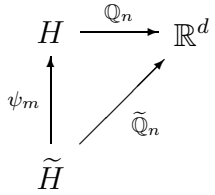
Now because $\Pr = \widetilde{\Pr} \circ \phi_m^{-1}$

$$\mathbb{Q}_n\frac{\mathbb{G}_m\nabla f_{\theta^*}(X|Y)/h(X)}{\mathbb{P}_m f_{\theta^*}(X|Y)/h(X)} \xrightarrow{\mathcal{L}}_{m,n} Q\,\mathbb{G}_P\left\{\frac{\nabla f_{\theta^*}(X|Y)}{h(X)}\right\}$$

and this implies (26) because $P\{\nabla f_{\theta^*}(X|y)/h(X)\} = 0$ by condition (9).

Since the limit does not contain any randomness associated with $Y$'s, it is independent of the first term on the right in (9).

More precisely, there is an almost sure representation for (23) with commutative diagram




and $\mathrm{Qr} = \widetilde{\mathrm{Qr}} \circ \psi_n^{-1}$ and if we combine (30) with this we get

$$\begin{pmatrix} \sqrt{n}\,\widetilde{\mathbb{Q}}_n(\tilde{\eta})\nabla \log f_{\theta^*}(Y) \\ \widetilde{\mathbb{Q}}_n(\tilde{\eta})\frac{\widetilde{\mathbb{G}}_m(\tilde{\omega})\nabla f_{\theta^*}(X|Y)/h(X)}{\widetilde{\mathbb{P}}_m(\tilde{\omega})f_{\theta^*}(X|Y)/h(X)} \end{pmatrix} = \begin{pmatrix} \sqrt{n}\,\widetilde{\mathbb{Q}}_n(\tilde{\eta})\nabla \log f_{\theta^*}(Y) \\ \sqrt{m}\,\widetilde{\mathbb{Q}}_n(\tilde{\eta})\nabla \log \widetilde{\mathbb{P}}_m(\tilde{\omega})\left\{\frac{\nabla f_{\theta^*}(X|Y)}{h(X)}\right\} \end{pmatrix}$$

$$\longrightarrow_{m,n} \begin{pmatrix} Z(\tilde{\eta}) \\ Q\,\widetilde{\mathbb{G}}_P(\widetilde{\omega})\left\{\frac{\nabla f_{\theta^*}(X|Y)}{h(X)}\right\} \end{pmatrix}$$

holding for almost all $\tilde{\eta}$ and $\tilde{\omega}$, where $Z(\tilde{\eta})$ is $\mathcal{N}(0, V)$. In this representation, it is clear that the two terms on the right hand side, being functions of independent random variables, are independent. This almost sure convergence implies weak convergence, and undoing the almost sure representation gives the result. □

LEMMA B.4. *Under conditions (5) and (8) through (10) of Theorem 2.3, $W$ is finite and*

$$Q\,\mathbb{G}_P \nabla f_{\theta^*}(X|Y)/h(X) \sim \mathcal{N}(0, W). \tag{31}$$

PROOF. Let $T$ denote the left hand side of (31):

$$T(\omega) = \int k(\omega, y)\, dQ(y),$$

where $k$ is defined by (25). By condition (5) there is a sequence $\{Q_i\}$ of probability measures with finite support such that $Q_i \xrightarrow{\mathcal{L}} Q$ (Aliprantis and Border, 1999, Theorem 14.10 and Theorem 14.12). Let $T_i$ be defined by

$$T_i(\omega) = \int k(\omega, y)\, dQ_i(y)$$

(this integral is a finite sum by $Q_i$ having a finite support). Since the function $y \mapsto k(\omega, y)$ is bounded and continuous for almost all $\omega \in \Omega$ (Lemma B.2), $T_i(\omega) \to T(\omega)$ for almost all $\omega \in \Omega$, that is, $T_i \xrightarrow{as} T$.

From $\mathbb{G}_P$ being a Gaussian process, $T_i$ is normally distributed with mean

$$ET_i = E\left(\int k(\cdot, y)\, dQ_i(y)\right) = \int E\left\{k(\cdot, y)\right\}\, dQ_i(y) = 0$$

and variance

$$\operatorname{var} T_i = E\left(T_i T_i^T\right) = \iint E\left\{k(\cdot, y)k(\cdot, s)^T\right\}\, dQ_i(y)\, dQ_i(s).$$

Note

$$E\left\{k(\cdot, y)k(\cdot, s)^T\right\} = P\left\{\frac{\nabla f_{\theta^*}(\cdot|y)}{h(\cdot)} \frac{\nabla f_{\theta^*}(\cdot|s)^T}{h(\cdot)}\right\}$$

and for all $y$ and $s$

$$P\left|\frac{\nabla f_{\theta^*}(\cdot|y)}{h(\cdot)}\frac{\nabla f_{\theta^*}(\cdot|s)^T}{h(\cdot)}\right|_{ij} \leq \left(PFF^T\right)_{ij}.$$

Since $PFF^T$ is finite by condition (8), the function $(y,s) \mapsto E\{k(\cdot,y)k(\cdot,s)^T\}$ is bounded and continuous by the dominated convergence theorem. So

$$\begin{aligned}
\operatorname{var} T_i &\to \iint E\{k(\cdot,y)k(\cdot,s)^T\}\, dQ(y)\, dQ(s) \\
&= \iint P\left\{\frac{\nabla f_{\theta^*}(\cdot|y)}{h(\cdot)}\frac{\nabla f_{\theta^*}(\cdot|s)^T}{h(\cdot)}\right\} dQ(y)\, dQ(s) \\
&= P\left\{\iint \frac{\nabla f_{\theta^*}(\cdot|y)}{h(\cdot)}\frac{\nabla f_{\theta^*}(\cdot|s)^T}{h(\cdot)}\, dQ(y)\, dQ(s)\right\} \\
&= P\left\{\int \frac{\nabla f_{\theta^*}(\cdot|y)}{h(\cdot)}\, dQ(y)\right\}\left\{\int \frac{\nabla f_{\theta^*}(\cdot|s)}{h(\cdot)}\, dQ(s)\right\}^T \\
&= \operatorname{var}_P Q\frac{\nabla f_{\theta^*}(X|Y)}{h(X)} \\
&= W,
\end{aligned}$$

where $W$ is defined by (7). The second equality follows from Fubini. From

$$\begin{aligned}
\left|\iint E\{k(\cdot,y)k(\cdot,s)^T\}\, dQ(y)\, dQ(s)\right|_{ij} &\leq \iint \left(PFF^T\right)_{ij} dQ(y)\, dQ(s) \\
&= \left(PFF^T\right)_{ij},
\end{aligned}$$

$W$ is finite and the application of Fubini is justified.

Now for any $t \in \mathbb{R}^d$

$$\exp(-t^T(\operatorname{var} T_i)t/2) \to \exp(-t^T W t/2).$$

The left hand side is the characteristic function of $T_i$. By the equivalence between the convergence in distribution and the convergence of characteristic functions

$$T_i \xrightarrow{\mathcal{L}} \mathcal{N}(0, W).$$

Since $T_i \xrightarrow{\mathcal{L}} T$ from $T_i \xrightarrow{as} T$, we have $T \sim \mathcal{N}(0, W)$. □

LEMMA B.5. *Under conditions (2) and (5) through (10) of Theorem 2.3,*

$$\left(\frac{V}{n}+\frac{W}{m}\right)^{-1/2}\nabla K_{m,n}(\theta^*)\xrightarrow{\mathcal{L}}\mathcal{N}(0,I). \tag{32}$$

PROOF. We start by proving (32) under the additional condition

$$\frac{n}{m+n}\to\alpha. \tag{33}$$

Necessarily $0\le\alpha\le 1$. First we do the case $0<\alpha<1$. Then

$$(m+n)\left(\frac{V}{n}+\frac{W}{m}\right)\longrightarrow\frac{V}{\alpha}+\frac{W}{1-\alpha}. \tag{34}$$

Also by the continuous mapping theorem, (9), and (22),

$$\begin{aligned}\sqrt{m+n}\,\nabla K_{m,n}(\theta^*) &= -\sqrt{\frac{m+n}{n}}\sqrt{n}\,\mathbb{Q}_n\nabla\log f_{\theta^*}(Y)\\ &\quad-\sqrt{\frac{m+n}{m}}\sqrt{m}\,\mathbb{Q}_n\nabla\log\mathbb{P}_m f_{\theta^*}(X|Y)/h(X)\\ &\xrightarrow{\mathcal{L}}\frac{V^{1/2}Z_1}{\sqrt{\alpha}}+\frac{W^{1/2}Z_2}{\sqrt{1-\alpha}},\end{aligned} \tag{35}$$

where $Z_1$ and $Z_2$ are independent normal random vectors with mean 0 and variance $I$. The right hand side of (35) has variance which is the right hand side of (34), so by Slutsky's theorem we have (32).

Now consider the case $\alpha=0$. Then $n=o(m)$ and we essentially have the first term.

$$n\left(\frac{V}{n}+\frac{W}{m}\right)\longrightarrow V,$$

and

$$\sqrt{n}\,\nabla K_{m,n}(\theta^*)\xrightarrow{\mathcal{L}}\mathcal{N}(0,V),$$

so again we have (32). The case $\alpha=1$ is similar.

Now the subsequence principle gives us (32) even without the assumption (33). Since $[0,1]$ is compact, it is always possible to choose a subsequence such that (33) holds. Since the limiting distribution does not depend on the subsequence chosen, the whole sequence converges to the same distribution. □

LEMMA B.6. *Under conditions (3), (6), (7), and (11) of Theorem 2.3,*

$$\sup_{\theta \in S_\rho} \left\| \nabla^2 K_{m,n}(\theta) - C(\theta) \right\|_\infty \xrightarrow{au} 0,$$

*where*

$$C(\theta) = -Q \left\{ \nabla^2 \log f_\theta(Y) \right\},$$

*and $C$ is continuous on $S_\rho$.*

PROOF. From

$$\nabla^2 K_{m,n}(\theta) = -\mathbb{Q}_n \nabla^2 \log f_\theta(Y) - \mathbb{Q}_n W_m(\theta, Y), \tag{36}$$

where

$$W_m(\theta, y) = \frac{\mathbb{P}_m \nabla^2 f_\theta(\cdot|y)/h(\cdot)}{\mathbb{P}_m f_\theta(\cdot|y)/h(\cdot)} - \frac{\{\mathbb{P}_m \nabla f_\theta(\cdot|y)/h(\cdot)\}\{\mathbb{P}_m \nabla f_\theta(\cdot|y)/h(\cdot)\}^T}{\{\mathbb{P}_m f_\theta(\cdot|y)/h(\cdot)\}^2},$$

we have

$$\left\| \nabla^2 K_{m,n}(\theta) - C(\theta) \right\|_\infty \leq \left\| -\mathbb{Q}_n \nabla^2 \log f_\theta(Y) - C(\theta) \right\|_\infty + \left\| -\mathbb{Q}_n W_m(\theta, Y) \right\|_\infty .$$

Condition (3) says that if $(\nabla^2 \log f_\theta)_{k,l}$ are the components of $\nabla^2 \log f_\theta$ and $C_{k,l}$ those of $C$,

$$\max_{k,l} \sup_{\theta \in S_\rho} \left| -\mathbb{Q}_n (\nabla^2 \log f_\theta(Y))_{k,l} - C_{k,l}(\theta) \right| \xrightarrow{au} 0.$$

The order of the max and sup can be interchanged (either way it is the max over both) and this proves

$$\sup_{\theta \in S_\rho} \left\| -\mathbb{Q}_n \nabla^2 \log f_\theta(Y) - C(\theta) \right\|_\infty \xrightarrow{au} 0.$$

The continuity of $C$ is established in the proof of (Ferguson, 1996, Theorem 16(a)). So we are done if we can show

$$\sup_{\theta \in S_\rho} \left\| -\mathbb{Q}_n W_m(\theta, Y) \right\|_\infty \xrightarrow{au} 0. \tag{37}$$

By condition (11)

$$\sup_{\theta\in S_\rho}\sup_{y\in\mathcal{Y}}\left\|\mathbb{P}_m\nabla^2 f_\theta(\cdot|y)/h(\cdot)\right\|_\infty \xrightarrow{au} 0. \tag{38}$$

From the fundamental theorem of calculus,

$$\nabla f_\theta(x|y) - \nabla f_{\theta^*}(x|y) = \int_0^1 \nabla^2 f_{\theta^*+s(\theta-\theta^*)}(x|y)(\theta-\theta^*)\,ds.$$

So

$$\begin{aligned}\mathbb{P}_m\nabla f_\theta(\cdot|y)/h(\cdot) = \mathbb{P}_m\nabla f_{\theta^*}(\cdot|y)/h(\cdot)& \\ + \int_0^1 \mathbb{P}_m\nabla^2 f_{\theta^*+s(\theta-\theta^*)}(\cdot|y)/h(\cdot)(\theta-\theta^*)\,ds,&\end{aligned}$$

and this implies, for any $\theta\in S_\rho$,

$$\begin{aligned}\sup_{y\in\mathcal{Y}}\left\|\mathbb{P}_m\nabla f_\theta(\cdot|y)/h(\cdot)\right\|_\infty \le \sup_{y\in\mathcal{Y}}&\left\|\mathbb{P}_m\nabla f_{\theta^*}(\cdot|y)/h(\cdot)\right\|_\infty \\ &+ \sup_{\theta\in S_\rho}\sup_{y\in\mathcal{Y}}\left\|\mathbb{P}_m\nabla^2 f_\theta(\cdot|y)/h(\cdot)\right\|_\infty \rho.\end{aligned}$$

The second term on the right converges almost uniformly to zero by (38). The first term on the right also converges almost uniformly to zero because $\mathcal{F}_3$ is $P$-Glivenko-Cantelli from being $P$-Donsker (from condition (7)). Thus

$$\sup_{\theta\in S_\rho}\sup_{y\in\mathcal{Y}}\left\|\mathbb{P}_m\nabla f_\theta(\cdot|y)/h(\cdot)\right\|_\infty \xrightarrow{au} 0. \tag{39}$$

A similar argument with condition (6) and (39) implies

$$\sup_{\theta\in S_\rho}\sup_{y\in\mathcal{Y}}\left|\mathbb{P}_m f_\theta(\cdot|y)/h(\cdot) - 1\right| \xrightarrow{au} 0. \tag{40}$$

Now (38), (39), and (40) imply

$$\sup_{\theta\in S_\rho}\sup_{y\in\mathcal{Y}}\left\|W_m(\theta,y)\right\|_\infty \xrightarrow{au} 0,$$

and we have (37) because

$$\sup_{\theta\in S_\rho}\left\|\mathbb{Q}_n W_m(\theta,Y)\right\|_\infty \le \sup_{\theta\in S_\rho}\sup_{y\in\mathcal{Y}}\left\|W_m(\theta,y)\right\|_\infty.$$

□

Another application of the fundamental theorem of calculus with $\theta \in S_\rho$ gives

$$\nabla K_{m,n}(\theta) - \nabla K_{m,n}(\theta^*) = \int_0^1 \nabla^2 K_{m,n}\big(\theta^* + s(\theta - \theta^*)\big)(\theta - \theta^*)\, ds. \tag{41}$$

Defining

$$D_{m,n}(\theta) = \int_0^1 \nabla^2 K_{m,n}\big(\theta^* + s(\theta - \theta^*)\big)\, ds, \tag{42}$$

(41) can be rewritten as

$$\nabla K_{m,n}(\theta) - \nabla K_{m,n}(\theta^*) = D_{m,n}(\theta)\,(\theta - \theta^*)\,. \tag{43}$$

LEMMA B.7. *Under conditions (3), (6), (7), (11), and (13) of Theorem 2.3,*

$$D_{m,n}(\hat{\theta}_{m,n}) \xrightarrow{P} J, \tag{44}$$

*where $D_{m,n}(\theta)$ is defined by* (42) *and $J$ in condition (2) of Theorem 2.3.*

PROOF. Note that with $C$ as defined in Lemma B.6, $C(\theta^*) = J$ in condition (2). Hence

$$\begin{aligned}
\|D_{m,n}(\hat{\theta}_{m,n}) - J\|_\infty &= \left\| \int_0^1 \nabla^2 K_{m,n}\big(\theta^* + s(\hat{\theta}_{m,n} - \theta^*)\big)\, ds - C(\theta^*) \right\|_\infty \\
&\le \int_0^1 \left\| \nabla^2 K_{m,n}\big(\theta^* + s(\hat{\theta}_{m,n} - \theta^*)\big) - C(\theta^*) \right\|_\infty ds \\
&\le \int_0^1 \left\| \nabla^2 K_{m,n}\big(\theta^* + s(\hat{\theta}_{m,n} - \theta^*)\big) - C\big(\theta^* + s(\hat{\theta}_{m,n} - \theta^*)\big) \right\|_\infty ds \\
&\quad + \sup_{0 \le s \le 1} \left\| C\big(\theta^* + s(\hat{\theta}_{m,n} - \theta^*)\big) - C(\theta^*) \right\|_\infty
\end{aligned}$$

The term on the bottom line converges in probability to zero by the continuity of $C$ and the weak consistency of $\hat{\theta}_{m,n}$. The term on the next to the bottom line also converges in probability to zero because for any $\epsilon > 0$

$$\begin{aligned}
&\Pr\left( \int_0^1 \left\| \nabla^2 K_{m,n}\big(\theta^* + s(\hat{\theta}_{m,n} - \theta^*)\big) - C\big(\theta^* + s(\hat{\theta}_{m,n} - \theta^*)\big) \right\|_\infty ds > \epsilon \right) \\
&\qquad \le \Pr\left( \hat{\theta}_{m,n} \notin S_\rho \right) + \Pr\left( \sup_{\theta \in S_\rho} \left\| \nabla^2 K_{m,n}(\theta) - C(\theta) \right\|_\infty > \epsilon \right),
\end{aligned}$$

the first term on the right going to zero by the consistency of $\hat{\theta}_{m,n}$ and the second term on the right going to zero by Lemma B.6. Hence we have proved (44). □

PROOF OF THEOREM 2.3. Condition (13) of the theorem and (43) imply

$$D_{m,n}(\hat{\theta}_{m,n})\left(\hat{\theta}_{m,n}-\theta^*\right) = -\nabla K_{m,n}(\theta^*) + \min(m,n)^{-1/2}o_p(1),$$

and this implies

$$\hat{\theta}_{m,n}-\theta^* = -D_{m,n}(\hat{\theta}_{m,n})^{-1}\nabla K_{m,n}(\theta^*) + \min(m,n)^{-1/2}o_p(1),$$

because $D_{m,n}(\hat{\theta}_{m,n})$ is invertible with probability converging to one. Multiplying both sides by the same matrix,

$$\begin{aligned}\left(\frac{V}{n}+\frac{W}{m}\right)^{-1/2} J\left(\hat{\theta}_{m,n}-\theta^*\right) \\ &= -\left(\frac{V}{n}+\frac{W}{m}\right)^{-1/2} JD_{m,n}(\hat{\theta}_{m,n})^{-1}\nabla K_{m,n}(\theta^*) \\ &\quad + \left\{\min(m,n)\left(\frac{V}{n}+\frac{W}{m}\right)\right\}^{-1/2} J\, o_p(1).\end{aligned}$$

Since

$$\min(m,n)\left(\frac{V}{n}+\frac{W}{m}\right) = O(1)$$

and by Slutsky's theorem and Lemma B.7

$$JD_{m,n}(\hat{\theta}_{m,n})^{-1} \xrightarrow{P} I,$$

we have (8) by Slutsky's theorem and Lemma B.5. □

# REFERENCES


ALIPRANTIS, C. D. and BORDER, K. C. (1999). *Infinite Dimensional Analysis*. Springer-Verlag, Berlin.

ATTOUCH, H. (1984). *Variational Convergence for Functions and Operators*. Pitman, Boston.

AUBIN, J.-P. and FRANKOWSKA, H. (1990). *Set-Valued Analysis*. Birkhäuser, Boston.

Billingsley, P. (1999). *Convergence of Probability Measures*. 2nd ed. Wiley, New York.

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

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–37.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. Chapman & Hall, London.

Gelfand, A. E. and Carlin, B. P. (1993). Maximum-likelihood estimation for constrained- or missing-data models. *Canad. J. Statist.* **21** 303–311.

Geyer, C. J. (1994a). On the asymptotics of constrained M-estimation. *Ann. Statist.* **22** 1993–2010.

Geyer, C. J. (1994b). On the convergence of Monte Carlo maximum likelihood calculations. *J. Roy. Statist. Soc. Ser. B* **56** 261–274.

Guo, S. W. and Thompson, E. A. (1992). Monte Carlo estimation of mixed models for large complex pedigrees. *Am. J. Hum. Genet.* **51** 1111–1126.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* 221–233. Univ. California Press, Berkeley, Calif.

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

Kong, A., Liu, J. S., and Wong, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.

Lange, K. and Sobel, E. (1991). A random walk method for computing genetic location scores. *Am. J. Hum. Genet.* **49** 1320–1334.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data-analysis using generalized linear-models. *Biometrika* **73** 13–22.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. Wiley, Hoboken.

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

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

Moyeed, R. A. and Baddeley, A. J. (1991). Stochastic approximation of the MLE for a spatial point pattern. *Scand. J. Statist.* **18** 39–50.

Ott, J. (1979). Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. *Am. J. Hum. Genet.* **31** 161–175.

Penttinen, A. (1984). Modelling interaction in spatial point patterns: Parameter estimation by the maximum likelihood method. *Jyväskylä Studies in Computer Science, Economics, and Statistics* **7**.

Rockafellar, R. T. and Wets, R. J.-B. (1998). *Variational Analysis*. Springer-Verlag, Berlin.

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

Thompson, E. A. (2003). Linkage analysis. In *Handbook of Statistical Genetics*. 2nd ed. (D. J. Balding, M. Bishop, and C. Cannings, eds.) 893–918. Wiley, Chichester.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag, New York.

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

Wei, G. C. G. and Tanner, M. A. (1990). A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.* **85** 699–704.

White, H. A. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–26.

Wijsman, R. A. (1964). Convergence of sequences of convex sets, cones and functions. *Bull. Amer. Math. Soc.* **70** 186–188.

Wijsman, R. A. (1966). Convergence of sequences of convex sets, cones and functions. II. *Trans. Amer. Math. Soc.* **123** 32–45.

Younes, L. (1988). Estimation and annealing for Gibbsian fields. *Ann. Inst. H. Poincaré Probab. Statist.* **24** 269–294.

Y. J. Sung
Division of Medical Genetics
University of Washington
Box 357720
Seattle, WA 98195-7720
E-mail: yunju@u.washington.edu

C. J. Geyer
School of Statistics
University of Minnesota
313 Ford Hall
224 Church St. S. E.
Minneapolis, MN 55455
E-mail: charlie@stat.umn.edu