Estimability and Likelihood Inference for Generalized Linear Mixed Models Using Data Cloning

Subhash R. LELE, Khurram NADEEM, and Byron SCHMULAND

Maximum likelihood estimation for Generalized Linear Mixed Models (GLMM), an important class of statistical models with substantial applications in epidemiology, medical statistics, and many other fields, poses significant computational difficulties. In this article, we use data cloning, a simple computational method that exploits advances in Bayesian computation, in particular the Markov Chain Monte Carlo method, to obtain maximum likelihood estimators of the parameters in these models. This method also leads to a simple estimator of the asymptotic variance of the maximum likelihood estimators. Determining estimability of the parameters in a mixed model is, in general, a very difficult problem. Data cloning provides a simple graphical test to not only check if the full set of parameters is estimable but also, and perhaps more importantly, if a specified function of the parameters is estimable. One of the goals of mixed models is to predict random effects. We suggest a frequentist method to obtain prediction intervals for random effects. We illustrate data cloning in the GLMM context by analyzing the Logistic–Normal model for over-dispersed binary data, and the Poisson–Normal model for repeated and spatial counts data. We consider Normal–Normal and Binary–Normal mixture models to show how data cloning can be used to study estimability of various parameters. We contend that whenever hierarchical models are used, estimability of the parameters should be checked before drawing scientific inferences or making management decisions. Data cloning facilitates such a check on hierarchical models.

KEY WORDS: Bayesian computation; Hierarchical models; Random effects.

1. INTRODUCTION

Linear mixed models (LMM) (Searle, Casella, and McCulloch 1992) and their extension to generalized linear mixed models (GLMM) (McCulloch and Searle 2001) consist of some of the most useful models in statistics. They are widely used in various fields, for example, longitudinal data analysis (Diggle, Liang, and Zeger 1994), epidemiology (Clayton and Kaldor 1987) and ecology and environmental sciences (Clark and Gelfand 2006; Royle and Dorazio 2009). For theoretical discussion of LMM and GLMM, see McCulloch and Searle (2001). Most popular approaches to analyze these models are Bayesian, based on the Markov Chain Monte Carlo (MCMC) algorithm and noninformative priors. (Gilks, Richardson, and Spiegelhalter 1996; Spigelhalter et al. 2004). However, likelihood analysis for these models is difficult (McCulloch 1997; McCulloch and Searle 2001). Likelihood analysis, if used, is usually conducted using approximate likelihood (Breslow and Clayton 1993) or Monte Carlo estimation of the likelihood function (e.g., McCulloch 1997; deValpine 2004).

Recently, Lele, Dennis, and Lutscher (2007) reviewed the difficulties associated with Bayesian and likelihood based approaches and proposed an alternative approach, called data cloning, to compute maximum likelihood estimates and their standard errors for general hierarchical models. See also Doucet, Godsill, and Robert (2002), Kuk (2003), and Jacquier, Johannes, and Polson (2007) for methods similar to data cloning. This approach is based on Bayesian ideas, uses well-known MCMC methodology and can be easily implemented in standard software such as WinBUGS. Data cloning is applicable in most situations where the problem can be formulated as a Bayesian problem and where MCMC can be used to obtain random variates from the posterior distribution. Similar to the Bayesian methodology, data cloning avoids high-dimensional numerical integration and requires neither maximization nor differentiation of a function. It is based only on the computation of the means and the variances. Although data cloning uses a Bayesian formulation and computational techniques, the inferences are based on the classical frequentist paradigm. Unlike the Bayesian inference, these inferences do not depend on the choice of the prior distributions used in the implementation of the MCMC algorithm. The goals of this article are: (1) to use data cloning to analyze GLMM; (2) to provide a simple graphical procedure to determine an adequate number of clones; (3) to provide an algorithm to obtain prediction intervals for random effects; and, most importantly, (4) to provide a simple graphical procedure to determine estimability of the parameters in hierarchical models.

2. NOTATION AND STATISTICAL SET–UP

Let $\mathbf{y}_{(n)} = (y_1, y_2, \ldots, y_n)$ be the data vector where n denotes the sample size. We consider the following general hierarchical model set-up:

Hierarchy 1: $\mathbf{y}_{(n)}|\mathbf{X} = \mathbf{x} \sim h(\mathbf{y}_{(n)}; \mathbf{X} = \mathbf{x}, \boldsymbol{\theta}_1)$.
Hierarchy 2: $\mathbf{X} \sim g(\mathbf{x}; \boldsymbol{\theta}_2)$.

We observe $\mathbf{y}_{(n)}$ whereas $\mathbf{x}$ are unobserved. The parameters of interest are $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

The goal of the analysis is to estimate the parameters $\boldsymbol{\theta}$ and predict the unobserved states $\mathbf{x}$. The likelihood function for this hierarchical model set-up is $L(\boldsymbol{\theta}; \mathbf{y}_{(n)}) = \int h(\mathbf{y}_{(n)}|\mathbf{x}; \boldsymbol{\theta}_1) g(\mathbf{x}; \boldsymbol{\theta}_2)\, d\mathbf{x}$. The difficulties associated with using this function for statistical inference are mainly computational: (1) calculation of the likelihood function generally involves high-dimensional integration; (2) obtaining the location of the maximum using numerical search procedures is difficult because of the stochastic nature of the estimated likelihood; and (3) computing standard errors of the resultant estimators involves further difficulties in numerical computation of the second derivatives of the

Subhash R. Lele (E-mail: *slele@ualberta.ca*) and Byron Schmuland are Professors of Statistics, and Khurram Nadeem is Graduate Student, Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada. This work was supported in part by funding from NSERC, Canada. The comments by those involved in the editorial process improved the article substantially and are greatly appreciated.

Journal of the American Statistical Association
December 2010, Vol. 105, No. 492, Theory and Methods
DOI: 10.1198/jasa.2010.tm09757

log-likelihood function. Data cloning methodology described below circumvents all these difficulties in a remarkably simple fashion.

3. DESCRIPTION OF THE DATA–CLONING METHOD

Following is a brief description of the data-cloning method. For details and pedagogical description, see Lele, Dennis, and Lutscher (2007). Let us start with the standard Bayesian approach to inference for hierarchical models. Let the prior distribution on the parameter space be denoted by $\pi(\boldsymbol{\theta})$. Then, the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}_{(n)})$ is

$$\pi\big(\boldsymbol{\theta}|\mathbf{y}_{(n)}\big) = \frac{\{\int h(\mathbf{y}_{(n)}|\mathbf{x};\boldsymbol{\theta}_1)g(\mathbf{x};\boldsymbol{\theta}_2)\,d\mathbf{x}\}\pi(\boldsymbol{\theta})}{C(\mathbf{y}_{(n)})} = \frac{L(\boldsymbol{\theta};\mathbf{y}_{(n)})\pi(\boldsymbol{\theta})}{C(\mathbf{y}_{(n)})},$$

where $C(\mathbf{y}_{(n)}) = \int L(\theta;\mathbf{y}_{(n)})\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$ is the normalizing constant. The MCMC algorithms (Gilks, Richardson, and Spiegelhalter 1996; Spiegelhalter et al. 2004) are computational tools that facilitate generation of random variates from the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y}_{(n)})$ without ever actually computing the integrals in the numerator or the denominator.

To understand the idea behind the data-cloning algorithm, imagine a hypothetical situation where the statistical experiment underlying the observations $\mathbf{y}_{(n)}$ is repeated independently by K different individuals and by happenstance all these individuals obtain exactly the same set of observations $\mathbf{y}_{(n)}$. Let us denote these data by $\mathbf{y}^{(K)} = (\mathbf{y}_{(n)}, \mathbf{y}_{(n)}, \ldots, \mathbf{y}_{(n)})$. The likelihood function based on the combination of the data from these K independent experiments is given by $[L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^K$. Notice two important features of this likelihood function: (a) the location of the maximum of this function is exactly equal to the location of the maximum of $L(\boldsymbol{\theta};\mathbf{y}_{(n)})$, and (b) the Fisher information matrix based on this likelihood is K times the Fisher information matrix based on $L(\boldsymbol{\theta};\mathbf{y}_{(n)})$. In the following, we denote the maximum likelihood estimator by $\hat{\boldsymbol{\theta}}_{(n)}$ and the Fisher information matrix based on $L(\boldsymbol{\theta};\mathbf{y}_{(n)})$ by $\mathbf{I}(\hat{\boldsymbol{\theta}}_{(n)})$. We assume that the parameters are identifiable and that there is a unique mode (but possibly multiple smaller peaks) to the likelihood function. It is easy to see that the posterior distribution of $\boldsymbol{\theta}$ conditional on the data $\mathbf{y}^{(K)} = (\mathbf{y}_{(n)}, \mathbf{y}_{(n)}, \ldots, \mathbf{y}_{(n)})$ is given by

$$\pi_K\big(\boldsymbol{\theta}|\mathbf{y}_{(n)}\big) = \frac{[\int h(\mathbf{y}_{(n)}|\mathbf{x};\boldsymbol{\theta}_1)g(\mathbf{x};\boldsymbol{\theta}_2)\,d\mathbf{x}]^K\pi(\boldsymbol{\theta})}{C(K;\mathbf{y}_{(n)})} = \frac{[L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^K\pi(\boldsymbol{\theta})}{C(K;\mathbf{y}_{(n)})},$$

where

$$C\big(K,\mathbf{y}_{(n)}\big) = \int\left[\int h\big(\mathbf{y}_{(n)}|\mathbf{x};\boldsymbol{\theta}_1\big)g(\mathbf{x};\boldsymbol{\theta}_2)\,d\mathbf{x}\right]^K\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}$$

is the normalizing constant. Furthermore, it follows from the standard result regarding the asymptotic behavior of the posterior distributions (e.g., Walker 1969) that, under regularity conditions, if K is large, then $\pi_K(\boldsymbol{\theta}|\mathbf{y}^{(K)})$ is approximately Normal with mean $\hat{\boldsymbol{\theta}}_{(n)}$ and variance equal to $\frac{1}{K}\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{(n)})$. Hence, when K is large, this distribution is nearly degenerate at the MLE $\hat{\boldsymbol{\theta}}_{(n)}$. Furthermore, the mean of this posterior distribution is the MLE and K times the posterior variance is the corresponding asymptotic variance of the MLE $\hat{\boldsymbol{\theta}}_{(n)}$.

Of course, in reality, we do not have data from K such independent experiments. But, suppose, instead of looking at the distribution $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)})$ as the posterior distribution of $\boldsymbol{\theta}$ given the observations from K independent experiments, we look upon it as just another distribution, defined over the parameter space Θ, with probability function $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)}) = [L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^K\pi(\boldsymbol{\theta})/C(K,\mathbf{y}_{(n)})$. This distribution is simply a function of the single set of observations $\mathbf{y}_{(n)}$ and the model components $h(\cdot)$, $g(\cdot)$, and $\pi(\cdot)$. Because we do not have K independent experiments, results on the asymptotic behavior of the posterior distribution by Walker (1969) are not directly applicable. In the Appendix, we prove directly that, under regularity conditions, as K becomes large, this distribution is nearly degenerate at the MLE $\hat{\boldsymbol{\theta}}$ and the mean of the probability distribution $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)}) = [L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^K\pi(\boldsymbol{\theta})/C(K,\mathbf{y}_{(n)})$ converges to $\hat{\boldsymbol{\theta}}_{(n)}$, and for continuous parameters, its variance is approximately $K^{-1}\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{(n)})$. These are deterministic convergences of a sequence of functions and not the probabilistic convergences used in Walker (1969). It follows then that if we can generate random variates $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_B$ from the $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)}) = [L(\boldsymbol{\theta};\mathbf{y}_{(n)})]^K\pi(\boldsymbol{\theta})/C(K,\mathbf{y}_{(n)})$ distribution, then we can use their mean and variance to obtain the MLE $\hat{\boldsymbol{\theta}}_{(n)}$ and its asymptotic variance.

Fortunately, such generation of random variates from $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)})$ is quite easy using the MCMC technique. Essentially we conduct the thought experiment described above using computers. We create the K-cloned dataset, $\mathbf{y}^{(K)} = (\mathbf{y}_{(n)}, \mathbf{y}_{(n)}, \ldots, \mathbf{y}_{(n)})$, by repeating the observed data vector K times. We pretend that these data were obtained from K independent experiments and use the standard MCMC method to generate random variates from the posterior distribution $\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)})$. If K is large, the MLE of the parameter $\boldsymbol{\theta}$ is simply the mean of these random variates. Furthermore, if the parameter space is continuous, K times the variance (or, variance–covariance matrix for the multiparameter case) of these random variates is the variance of the MLE, the inverse of the Fisher information, based on the original data.

Remarkably, this procedure avoids: (1) analytical or numerical evaluation of the high-dimensional integral which is a major computational hurdle for maximum likelihood estimation for GLMM; (2) numerical optimization of a function; and (3) numerical computation of the curvature of the likelihood function. The number of clones to be used in the procedure is completely under the control of the analyst. It can be made as large as necessary to achieve the desired accuracy of the resultant estimates. Furthermore, as long as the prior distribution is not degenerate and the model satisfies some regularity conditions, the results do not depend on the choice of the prior distribution. Nevertheless, a prior that has large probability mass near the true MLE requires fewer clones to achieve the desired accuracy.

Determining Adequate Number of Clones

Determination of an adequate number of clones is the same as determining when the posterior distribution is nearly degenerate. A plot of the largest eigenvalue of the posterior variance as a function of the number of clones K is a simple, graphical way to determine if the posterior distribution has become

nearly degenerate. In fact, we also know that the largest eigenvalue of the posterior distribution converges to zero at the same rate as $1/K$. Hence, we divide the largest eigenvalue of the posterior variance for K clones, λ_K, by the largest eigenvalue of the posterior variance for a single clone, λ_1. We call this the standardized largest eigenvalue and denote it by λ_K^S. We plot λ_K^S against K and compare it with the expected value plot of $1/K$. We choose the number of clones so that λ_K^S is below a specified threshold. We know that as we increase the clones, $(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T \mathbf{V}^{-1} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \sim \chi_p^2$, where $\mathbf{V}$ is the variance of the posterior distribution, is approximately true. We compute two different statistics: (a) $\omega = \frac{1}{B} \sum_{q=1}^{B} (O_q - E_q)^2$, where $O_q = (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T \mathbf{V}^{-1} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})$ and E_q are the quantiles for χ_p^2 random variable, and (b) $\tilde{r}^2 = 1 - \text{corr}^2(O_q, E_q)$. If these statistics are close to zero, it indicates that the $(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})^T \mathbf{V}^{-1} (\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \sim \chi_p^2$ approximation is reasonable (Johnson and Wichern 2007).

We want to emphasize here that data cloning is simply a computational algorithm to compute the MLE and the inverse of the Fisher information. Although the heuristic explanation alludes to it, the mathematical proof of the algorithm does not depend on and in no way assumes that the K clones are independent of each other. The data-cloning idea is used only as a means to coax MCMC into generating random variates from the distribution $[L(\boldsymbol{\theta}; \mathbf{y}_{(n)})]^K \pi(\boldsymbol{\theta}) / C(K, \mathbf{y}_{(n)})$. Furthermore, as the number of cloned copies increases, the algorithm provides a better and better approximation of the true location of the MLE and true inverse of the Fisher information for the observed data. The statistical accuracy of the estimator is a function of the sample size and not of the number of cloned copies one uses. Data cloning does not improve the statistical efficiency of the estimator by artificially increasing the sample size.

Prediction of Random Effects

An important inferential component to many hierarchical models is prediction of random effects. One can use MCMC along with data cloning to obtain point prediction and prediction intervals for the random effects. The method is based on the results of Harris (1989) where it is shown that if one uses the bootstrap distribution of the parameters as the 'prior,' the posterior distribution of the random effects is the best approximation, in Kullback–Leibler divergence, to the true distribution. We suggest replacing the bootstrap distribution by the Normal approximation obtained by data cloning. This may also be looked upon as the prior invariant component of the posterior distribution. Thus, prediction inference on random effects is obtained by using

$$\pi\left(\mathbf{x}|\mathbf{y}_{(n)}\right) = \frac{\int h(\mathbf{y}_{(n)}|\mathbf{x}, \boldsymbol{\theta}_1) g(\mathbf{x}|\boldsymbol{\theta}_2) \phi(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_{(n)}, \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{(n)}))\, d\boldsymbol{\theta}}{C(\mathbf{y}_{(n)})},$$

where $\phi(\cdot, \mu, \sigma^2)$ indicates the Normal density with mean μ and variance σ^2. The MCMC algorithm can be used to obtain the draws from this distribution without actually conducting the integration. We simply obtain the random numbers from

$$\pi\left(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}_{(n)}\right) = \frac{h(\mathbf{y}_{(n)}|\mathbf{x}, \boldsymbol{\theta}_1) g(\mathbf{x}|\boldsymbol{\theta}_2) \phi(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}_{(n)}, \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{(n)}))}{C(\mathbf{y}_{(n)})}$$

and utilize only the $\mathbf{x}$ component.

4. ILLUSTRATIVE EXAMPLES

In the following we apply data cloning to obtain maximum likelihood estimates and associated asymptotic standard errors for three important subclasses of Generalized Linear Mixed Models with wide applications in medical statistics and epidemiology. The detailed description of the scientific problems, statistical models, and the data is available in Breslow and Clayton (1993). The following descriptions are borrowed from Breslow and Clayton (1993, section 6).

(1) Logistic–Normal Mixed Model

Crowder (1978, table 3) presented data on the proportion of seeds that germinated on each of 21 plates arranged according to a 2×2 factorial layout by seed variety and type of root extract. He noted that the within-group variation exceeded that predicted by binomial sampling theory. A natural way to account for extraneous plate-to-plate variability in this situation is by means of the following GLMM:

Hierarchy 1: $Y_i|p_i \sim \text{Binomial}(n_i, p_i)$, where

$$\log \frac{p_i}{1 - p_i} = \alpha_0 + \alpha_{\text{seed}} SEED + \alpha_{\text{extract}} EXTRACT + \alpha_{\text{interaction}} SEED * EXTRACT + b_i.$$

Hierarchy 2: $b_i \sim N(0, \sigma_b^2)$.

Breslow and Clayton (1993) provide the exact ML estimates of the parameters along with their standard errors based on numerical integration. In Table 1, we provide the results based on the data-cloning algorithm with two different priors, a noninformative prior and prior based on the GLM estimates and compare them with those based on noninformative Bayes estimates. The data cloning based MLEs and their SEs are nearly identical to the exact ML estimators and are invariant to the choice of the priors. Figure 1(a) gives the data cloning convergence diagnostics and Figure 1(b) shows the data-cloning-based point predictions and prediction intervals for the probability of germination along with those based on noninformative priors. These match reasonably well with the ones obtained by using noninformative Bayes approach.

(2) Longitudinal Data

Thall and Vail (1990, table 2) presented data from a clinical trial of 59 epileptics who were randomized to a new drug ($Trt = 1$) or a placebo ($Trt = 0$) as an adjuvant to the standard chemotherapy. Baseline data available at entry into the trial included the number of epileptic seizures recorded in the preceding eight-week period and age in years. The logarithm of the fourth of the number of baseline seizures (*Base*) and the logarithm of age (*AGE*) were treated as covariates in the analysis. A multivariate response variable consisted of the counts of seizures during the two weeks before each of four clinic visits (Visit, coded -3, -1, 1, and 3). Preliminary analysis indicated that the counts were substantially lower during the fourth visit and a binary variable ($V4 = 1$ for fourth visit, 0 otherwise) was constructed to model such effects. Breslow and Clayton (1993) use the following GLMM for modeling these data:

Table 1. Maximum likelihood estimates and standard errors (SEs) using data cloning under two different priors and comparison with the estimates and variances using the noninformative Bayesian analysis

Parameters		Data Cloning 1	Data Cloning 2	Noninformative Bayes
Seeds data (Logistic Normal model)	α_0	−0.5484 (0.1693)	−0.5491 (0.1623)	−0.5488 (0.2129)
	α_1	0.0970 (0.2758)	0.0993 (0.2771)	0.0515 (0.3462)
	α_2	1.3372 (0.2403)	1.3378 (0.2357)	1.3583 (0.3076)
	α_{12}	−0.8113 (0.3837)	−0.8133 (0.3879)	−0.8181 (0.4762)
	σ	0.2376 (0.1069)	0.2361 (0.1061)	0.3546 (0.1469)
Lip Cancer data (Spatial Poisson Normal model)	α_0	−0.4381 (0.1693)	−0.4397 (0.1372)	−0.5581 (0.1496)
	α_1	0.6078 (0.0901)	0.6084 (0.1181)	0.6560 (0.0893)
	σ	1.2888 (0.2112)	1.2890 (0.1992)	1.4468 (0.2214)
	γ	0.1770 (0.0111)	0.1770 (0.0101)	0.1429 (0.0388)
Epilepsy data (Poisson Normal model)	α_0	−1.3934 (1.1965)	−1.4070 (1.2343)	−1.4165 (1.2537)
	α_{Base}	0.8782 (0.1318)	0.8822 (0.1180)	0.8824 (0.1293)
	α_{Trt}	−0.9493 (0.3827)	−0.9448 (0.3959)	−0.9739 (0.3889)
	α_{BT}	0.3501 (0.1913)	0.3473 (0.1975)	0.3632 (0.1980)
	α_{Age}	0.4852 (0.3519)	0.4872 (0.3715)	0.4883 (0.3700)
	α_{V4}	−0.1019 (0.0861)	−0.1016 (0.0872)	−0.1026 (0.0877)
	σ_b	0.3590 (0.0430)	0.3593 (0.0412)	0.3622 (0.0428)
	σ_{b1}	0.4623 (0.0622)	0.4621 (0.0635)	0.4934 (0.0697)

NOTE: For the Seeds data (Logistic Normal model), the exact MLEs and SEs are $\alpha_0 = -0.546$ (0.167), $\alpha_1 = 0.097$ (0.278), $\alpha_2 = 1.337$ (0.237), $\alpha_{12} = -0.811$ (0.385), $\sigma = 0.236$ (0.110).

Hierarchy 1: $Y_{ijk}|\mu_{jk} \sim \text{Poisson}(\mu_{jk})$, where

$$\log \mu_{jk} = \alpha_0 + \alpha_{\text{AGE}} AGE + \alpha_{\text{BASE}} BASE + \alpha_{\text{Trt}} Trt + \alpha_{\text{BT}}(BASE * Trt) + \alpha_{\text{V4}} V4 + b_j + b_{jk}.$$

Hierarchy 2: $b_j \sim N(0, \sigma_b^2)$ and $b_{jk} \sim N(0, \sigma_{b1}^2)$.

In Table 1, we present the MLEs obtained using data-cloning procedure. The results again do not depend on the choice of the priors. In Figure 1(c), we show the convergence diagnostic plots and Figure 1(d) shows the data cloning based-point predictions and prediction intervals for subject effects. These match reasonably well with the ones obtained using noninformative priors.

(3) Spatial Smoothing of Disease Maps

One of the most common applications of GLMM is in the context of spatial smoothing of disease maps (Clayton and Kaldor 1987; Diggle, Tawn, and Moyeed 1998). We consider the data reported in Clayton and Kaldor (1987) on the number of lip cancer cases in the 56 counties of Scotland. Clayton and Kaldor (1987) proposed an empirical Bayes estimation of the county specific SMRs using several alternative assumptions about the distribution of the random effects. These data subsequently were analyzed by Breslow and Clayton (1993) using the PQL. In the following analysis, we use a proper, conditionally specified autoregression (CAR) model. A full discussion of these different analyses along with the Bayesian implementation is available in WinBUGS (Spiegelhalter et al. 2004, maps section). The model we use is as follows:

Hierarchy 1: $Y_i|\mu_i \sim \text{Poisson}(\mu_i)$.

Hierarchy 2: $\log \mu_i = \log e_i + \alpha_0 + \alpha_1 \frac{x_i}{10} + b_i$, where $e_i =$ expected count and $x_i = \%$ of workforce employed in agriculture, fishing, and forestry.

Hierarchy 3: $\underline{\mathbf{b}} \sim \text{MVN}(\underline{\mathbf{0}}, \mathbf{V})$, where $\mathbf{V} = \sigma^2(I - \gamma \mathbf{C})^{-1}\mathbf{M}$, $M_{ij} = 1/e_i$, the inverse of the expected count in the ith area, and $C_{ij} = (e_i/e_j)^{1/2}$. The spatial association parameter $\gamma \in (\gamma_{\min}, \gamma_{\max})$, where $\gamma_{\min}^{-1}$ and $\gamma_{\max}^{-1}$ are the smallest and largest eigenvalues of $\mathbf{M}^{-1/2}\mathbf{C}\mathbf{M}^{1/2}$.

This ensures that the distribution of the random effects is a proper distribution. The maximum likelihood estimates and standard errors of the parameters are provided in Table 1. Convergence diagnostics are shown in Figure 1(e) and predicted random effects and associated prediction intervals for counties are shown in Figure 1(f). They again match well with the ones based on noninformative priors.

5. ESTIMABILITY DIAGNOSTICS

Many hierarchical models have nonidentifiable parameters. For example, in the standard measurement error model $Y_i|\mu_i \sim N(\mu_i, \sigma^2)$ and $\mu_i \sim N(\mu, \tau^2)$ where $i = 1, 2, \ldots, n$, the parameters $(\mu, \sigma^2 + \tau^2)$ are identifiable but parameters (μ, σ^2, τ^2) are not identifiable. It is known that (McCulloch and Searle 2001) for the Logistic–Normal model (Example 1, Section 4), if only one observation per stratum is available, the variance parameter σ^2 is confounded with the intercept parameter β_0. The analytical proof of this result, however, is difficult. In most practical applications, models are substantially more complex (Royle and Dorazio 2009; Clark and Gelfand 2006), making analytical proofs for identifiability of the parameters extremely difficult and are rarely attempted. Analysis is usually carried out as if the parameters are, in fact, identifiable (Lele 2010).

Data cloning provides a simple solution to this important problem. We prove (the Appendix) that if the parameters are nonestimable, as we increase the number of clones, the posterior distribution converges to a truncated prior distribution, truncated over the space of nonestimable parameter values.

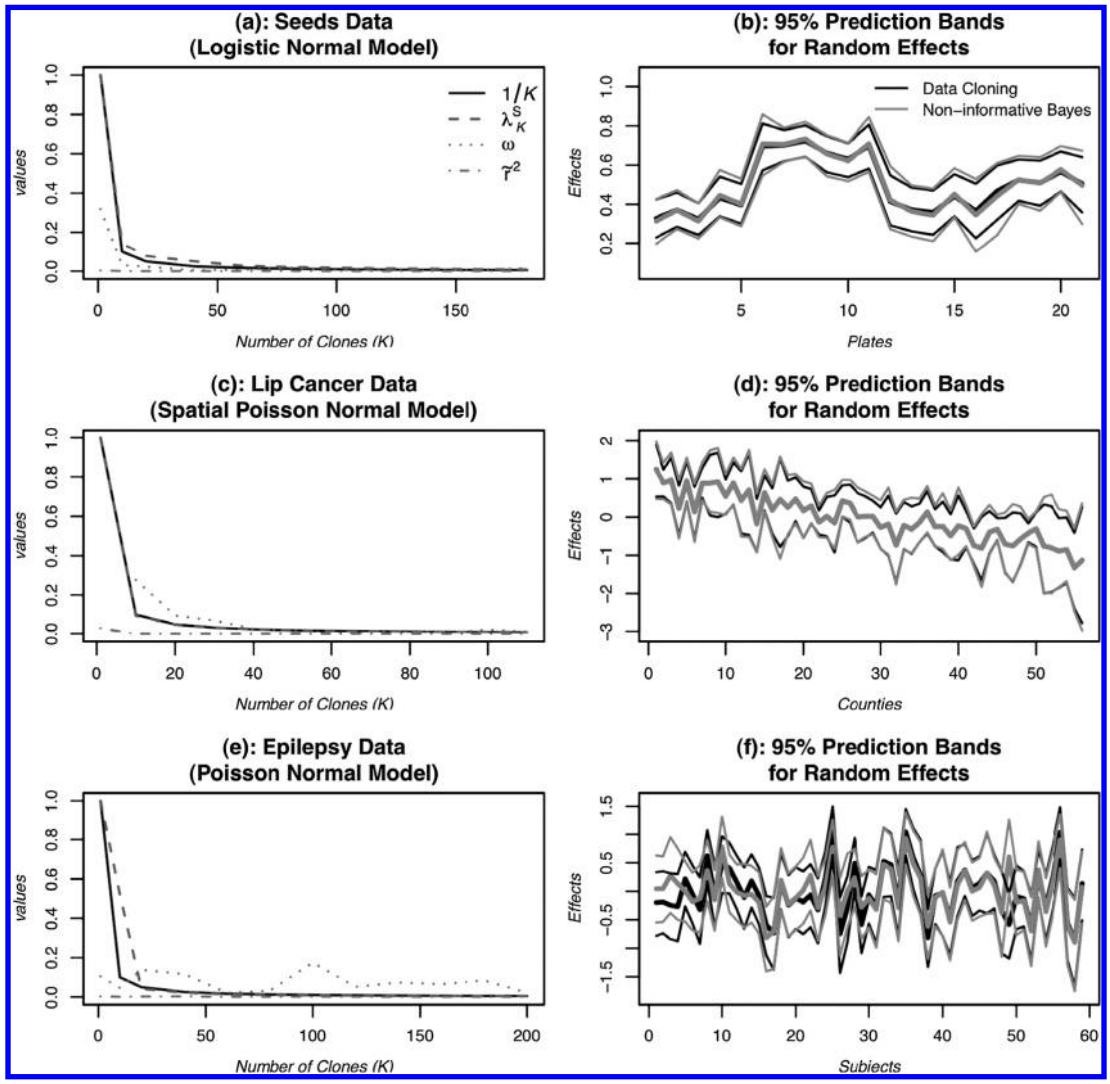

Figure 1. Data cloning convergence diagnostics and prediction of random effects for the three examples. The standardized eigenvalues converge to zero at the expected rate for all three cases. Data-cloning-based prediction intervals for random effects are quite similar to the ones obtained using noninformative priors.

Consequently, the largest eigenvalue of the posterior variance matrix does not converge to zero. This result can be used to study lack of identifiability of the parameters in the hierarchical model as a whole.

In practice, one may be interested in finding out whether certain functions of parameters are estimable. For example, in linear regression if the covariate matrix is singular, the regression parameters are nonestimable; however, the mean responses or differences in the treatment effects are estimable. Similarly in the applications of hierarchical models, a researcher might be interested in knowing if a specific parameter or a function of the parameters is estimable or not. The result in the Appendix can be used to find out if a specific parameter or a function of the parameters is estimable. If the variance of the posterior distribution of the parameter of interest converges to zero, the parameter is estimable. Thus, data cloning not only alerts the researcher about nonestimability of the parameters in the model but also helps him/her in deciding if certain parameter(s) of interest are estimable or not. In the following, we illustrate the use of this technique.

We start with a model where identifiabilty of various parameters is well established. Let $Y_i|\mu_i \sim N(\mu_i, \sigma^2)$ and $\mu_i \sim N(\mu, \tau^2)$ for $i = 1, 2, \ldots, n$. We generated a single realization from this model and used data cloning to estimate the parameters. We plot the largest eigenvalue of the posterior variance, λ_K^S, as a function of K. We also plot the posterior variance for various parameters that are of interest. In Figure 2(a), it is clear that λ_K^S does not converge to zero as the number of clones is increased, indicating nonestimability for the full model. On the other hand, the posterior variance for μ and $\gamma = \sigma^2 + \tau^2$ converges to zero as the number of clones increases, indicating their estimability. This shows that in the Normal–Normal model, μ and $\sigma^2 + \tau^2$ are estimable whereas σ^2 and τ^2 individually are not. Now we consider the classic Kalman filter model $Y_i|\mu_i \sim N(\mu_i, \sigma^2)$ and $\mu_i|\mu_{i-1} \sim N(a + c\mu_{i-1}, \tau^2)$ for $i = 1, 2, \ldots, n$. The Normal–Normal model above is a particular case of this model. However, introduction of correlation makes the parameters (a, c, σ^2, τ^2) identifiable as long as $c \neq 0$. In Figure 2(b), the plot of λ_K^S for the Kalman filter model clearly shows that the parameters are estimable.

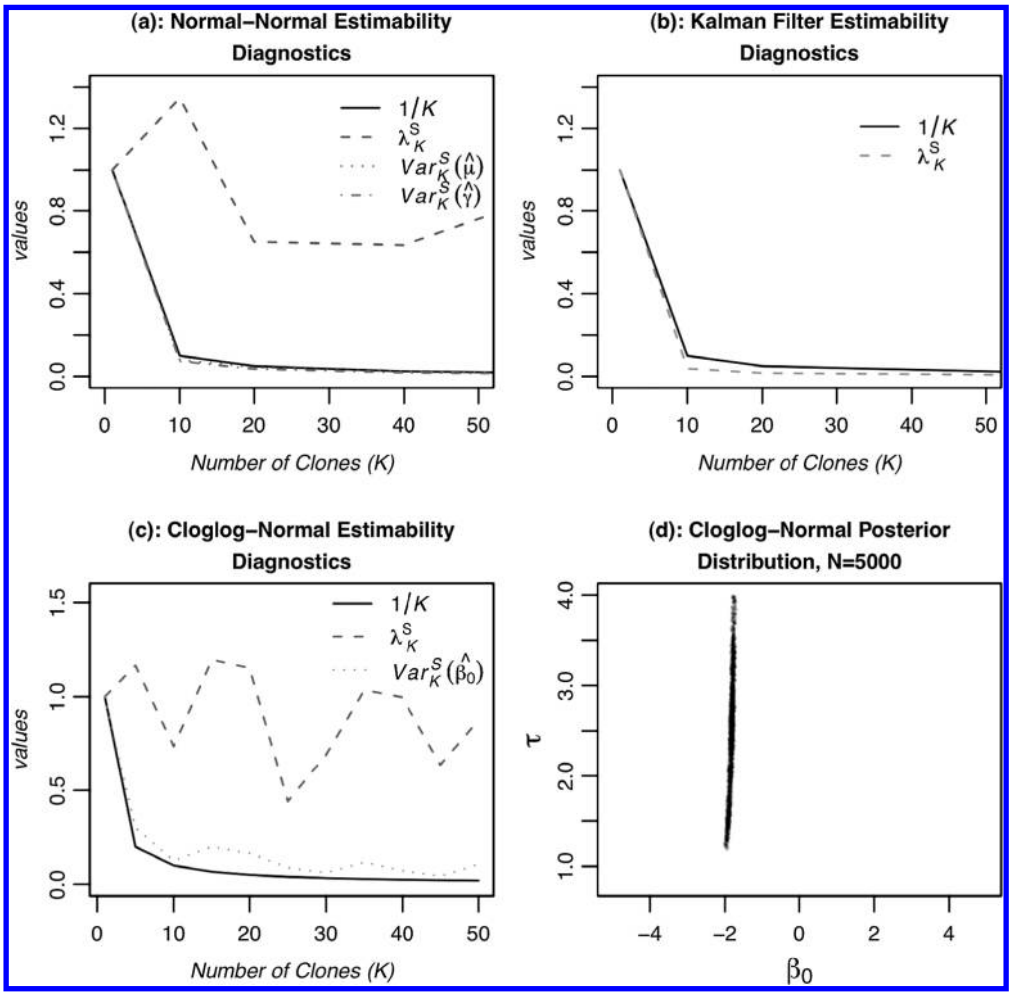

Figure 2. Estimability diagnostics using data cloning. In part (a), we consider Normal–Normal mixture. It is clear that λ_K^S does not converge to zero as K increases indicating nonestimability. However, the variance for μ does converge to zero indicating estimability. In part (b), we consider Kalman filter model. All parameters are estimable because λ_K^S does converge to zero as K increases. In part (c), we consider Binary–Normal mixture with complementary log–log link. It is clear that the model is nonestimable. Part (d) shows the posterior distribution is a truncated version of the prior distribution on a nondegenerate set supporting the nonestimability result.

Next we consider mixed Binary regression model. The analytical proof for the identifiability of various parameters in this model is difficult to establish (McCulloch and Searle 2001). Let $Y_i|p_i \sim \text{Bernoulli}(p_i), p_i = 1 - \exp(-\exp(\beta_0 + \varepsilon_i))$ and $\varepsilon_i \sim N(0, \sigma^2)$ for $i = 1, 2, \ldots, n$. We considered $n = 100$ and the number of clones $1, 5, 10, \ldots, 50$. In Figure 2(c), we plot λ_K^S against K. It is obvious that the parameters in this model are nonestimable. To check this result, we also plotted in Figure 2(d) the posterior distribution based on 5000 observations and uniform priors. It is quite clear that the posterior distribution is nondegenerate even for such a large sample size and informative priors of Uniform$(-5, 5)$ and Uniform$(0.8, 4)$. The marginal posterior distribution plot of β_0 as well as the data-cloning plot for its variance as a function of the number of clones indicates that this parameter may be estimable. However, the rate at which the variance for β_0 converges to zero is not close to the theoretical rate of $1/K$ as was the case when the parameters are consistently estimable. Convergence may not necessarily indicate that the estimator is consistent for the true value. The posterior mean for β_0 was -1.82 (true value $= -2$) indicating possible inconsistency of this estimator.

The Bayesian perspective on identifiability is discussed in various articles (see, e.g., Gelfand and Sahu 1999 or Eberly and Carlin 2000). Both these articles note that sometimes the identifiability problems are subtly apparent in the convergence diagnostics for the MCMC or in the sensitivity of the posterior to the choice of the prior. They also discuss the concept of Bayesian learning when prior distribution is changed due to the data. They seem to indicate that existence of Bayesian learning implies there are likely to be no problems with estimability. In the Binary–Normal example discussed above, the posterior distribution for the precision parameter $\tau = 1/\sigma^2$ was different than the prior distribution indicating some 'Bayesian learning' but clearly the parameter is nonestimable. Thus, some Bayesian learning is feasible even when the parameter is nonestimable. See also Lele (2010) for another example. This is concurrent with our result in Theorem A.2 that the posterior distribution in the nonestimable parameter case is a truncated version of the prior distribution, not necessarily the prior distribution itself. Similarly, we obtained good mixing and convergence (Gelman–Rubin statistics of 1.06 and 1.12, respectively). These results also indicate that good mixing and convergence of the MCMC

or evidence of Bayesian learning, although necessary, is not sufficient for estimability of the parameters.

Convergence problems with MCMC and sensitivity to the choice of the prior can arise for various reasons. Aside from the possibility of nonestimability, they can also arise when the likelihood is relatively, but not exactly, flat or has multiple but unequal modes. These problems do not necessarily imply that the parameters are nonestimable. In data cloning, the information content of the sample is increased through cloning. By doing so, we eliminate the possibility of small information content affecting the convergence of MCMC and sensitivity to the choice of the prior. Thus, data-cloning-based test is clear and unambiguous. Of course, we consider this test as an additional tool to check for possible problems with the model and not a replacement of the checks proposed by Eberly and Carlin (2000) and others. Furthermore, in practice, published articles based on MCMC methodology seldom provide information on whether such checks were, in fact, conducted. Data-cloning methodology forces researchers to think about estimability issue carefully and to conduct such checks.

Hierarchical models are easy to construct and, thanks to MCMC, are easy to analyze. As a general principle, complexity of the model should not exceed the information content in the data (Lele 2010). Data cloning alerts the researcher to the potential pitfalls of the model such as nonestimability and points out any mismatch between the desired complexity of the model and what is feasible given the data.

6. DISCUSSION

In this article, we show the applicability of data cloning for conducting likelihood inference for GLMM and for predicting random effects. It is well known (e.g., Natarajan and Kass 2000) that the choice of the noninformative prior is crucial when applying MCMC to conduct the noninformative Bayesian inference. Improper priors can lead to improper posteriors. The data-cloning algorithm, because it is invariant to the choice of the prior distribution, can utilize a prior distribution that is computationally convenient and proper. Thus avoiding the possibility of improper posterior distributions. The inverse of the Fisher information matrix is not always a good approximation to the variance of the estimator, especially for smaller sample sizes. One can always use bootstrapping as an alternative to the inverse of the Fisher information to estimate the variance and to obtain confidence intervals. One of the appealing features of data cloning is the test for estimability of parameters in hierarchical models. Understanding estimability of the parameters is extremely important in practice, where models are complex and analytical results are sparse. Any valid scientific inference can only be based on identifiable parameters. Thus, checking for estimability is critical for good scientific practice. Although not illustrated here, further inference procedures such as model selection using information criteria, profile likelihood for inference in the presence of nuisance parameters etc. are also possible using data cloning (Ponciano et al. 2009).

APPENDIX

A.1 Proof of Convergence

Let Θ denote the parameter space. This is subset of a p-dimensional Euclidean space.

Let $f(\mathbf{y};\boldsymbol{\theta})$ denote the joint probability density function of the data vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$. We assume that this is a bounded function as a function of $\boldsymbol{\theta}$. Let $\pi(\boldsymbol{\theta})$ denote the prior distribution, a probability density function, defined on the parameter space.

Let $\pi_K(\boldsymbol{\theta}|\mathbf{y}) = f^K(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/c(K)$, where $c(K) = \int f^K(\mathbf{y}|\boldsymbol{\theta}) \times \pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$. We are suppressing the dependence of $c(K)$ on $\mathbf{y}$ for notational simplicity.

Assumption A.1. The function $f(\cdot)$, as a function of $\boldsymbol{\theta}$, has a local maximum at $\boldsymbol{\theta}_\infty$ and $f(\boldsymbol{\theta}_\infty) > 0$ and $\pi(\boldsymbol{\theta}_\infty) > 0$. The maximum likelihood estimator is, by definition, denoted by $\boldsymbol{\theta}_\infty$.

Assumption A.2. The function $\pi(\cdot)$ is continuous at $\boldsymbol{\theta}_\infty$, the function $f(\cdot)$ has continuous second derivatives in a neighborhood of $\boldsymbol{\theta}_\infty$ and $D^2 f(\boldsymbol{\theta}_\infty)$ is strictly negative definite.

Assumption A.3. For any $\delta > 0$, we have $\gamma(\delta) := \sup\{f(\boldsymbol{\theta}) : \|\boldsymbol{\theta} - \boldsymbol{\theta}_\infty\| > \delta\} < f(\boldsymbol{\theta}_\infty)$.

Definition A.1 (Neighborhood). Let $\boldsymbol{\Sigma} = \{-D^2 f(\boldsymbol{\theta}_\infty)\}^{-1/2}$ and for $\delta > 0$ define $N(\delta) := \{\boldsymbol{\theta} : \|\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty)\| < \delta\}$. Because $\boldsymbol{\Sigma}$ is positive definite, this defines a system of neighborhoods of $\boldsymbol{\theta}_\infty$.

Definition A.2. Let $\boldsymbol{\Theta}_K$ be a random variable on $\Re^p$ with density function $\pi_K(\cdot)$ and define the standardized variable $\Psi_K = \sqrt{K}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Theta}_K - \boldsymbol{\theta}_\infty)$ that has density function $g_K(\boldsymbol{\theta}) = \frac{|\boldsymbol{\Sigma}|}{K^{p/2}}\pi_K(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta})$.

Without loss of generality, we can assume that $f(\boldsymbol{\theta}_\infty) = 1$. This is simply a standardized likelihood function and the computation of the posterior distribution $\pi_K(\boldsymbol{\theta}|\mathbf{y})$ is invariant to such standardizations. Thus, $\boldsymbol{\Sigma} = \{-D^2 f(\boldsymbol{\theta}_\infty)\}^{-1/2}$ corresponds to the square root of the inverse of the Fisher information matrix because $D^2 f(\boldsymbol{\theta}_\infty) = D^2 \log f(\boldsymbol{\theta}_\infty)$.

Lemma A.1. Under Assumptions A.1 and A.2, $f^K(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta})$ converges to $\exp(-\|\boldsymbol{\theta}\|^2/2)$ uniformly on bounded sets of $\boldsymbol{\theta}$ as $K \to \infty$.

Proof. Fix $\delta_0 > 0$ so small that $D^2 f(\boldsymbol{\theta})$ is continuous on the neighborhood $N(\delta_0)$. For every $\boldsymbol{\theta}$ in this neighborhood, Taylor's theorem says that there is some $\boldsymbol{\theta}^+$ on the line segment joining $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_\infty$ so that

$$\begin{aligned} f(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}_\infty) + Df(\boldsymbol{\theta}_\infty)(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty) \\ &\quad + \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty)^T (D^2 f(\boldsymbol{\theta}^+))(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty) \\ &= 1 - \tfrac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty)^T (-D^2 f(\boldsymbol{\theta}^+))(\boldsymbol{\theta} - \boldsymbol{\theta}_\infty). \end{aligned} \tag{A.1}$$

For any $\boldsymbol{\theta} \in \Re^p$, when K is large, the vector $\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}$ is in $N(\delta_0)$ and we have

$$f\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right) = 1 - \frac{\boldsymbol{\theta}^T \boldsymbol{\Sigma}^T \{-D^2 f(\boldsymbol{\theta}_K)\}\boldsymbol{\Sigma}\boldsymbol{\theta}}{2K},$$

for some $\boldsymbol{\theta}_K$ on the line segment joining $\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}$ and $\boldsymbol{\theta}_\infty$.

For $\varepsilon > 0$, choose $\delta(\varepsilon) < \delta_0$ so small that for $\boldsymbol{\theta} \in N(\delta(\varepsilon))$, we have $D^2 f(\boldsymbol{\theta})$ is negative definite and $\|\boldsymbol{\Sigma}^T(-D^2 f(\boldsymbol{\theta}))\boldsymbol{\Sigma} - \mathbf{I}\| \le \varepsilon$. Now, for $0 \le x, y \le K$, we have

$$\begin{aligned} &\left|\left(1 - \frac{x}{K}\right)^K - \left(1 - \frac{y}{K}\right)^K\right| \le |x - y| \qquad \text{and} \\ &\left|\left(1 - \frac{y}{K}\right)^K - \exp(-y)\right| \le \frac{y^2}{K}. \end{aligned} \tag{A.2}$$

Fix $M > 1$ and $0 < \varepsilon < 1$ and let $K \ge \max((M/\delta(\varepsilon))^2, M^2)$. Then for $\|\boldsymbol{\theta}\| < M$ we have $\boldsymbol{\theta}_K \in N(\delta(\varepsilon))$, so using (A.2) with $x =$

$\frac{1}{2}\boldsymbol{\theta}^T\boldsymbol{\Sigma}^T(-D^2 f(\boldsymbol{\theta}_K))\boldsymbol{\Sigma}\boldsymbol{\theta}$ and $y = \|\boldsymbol{\theta}\|^2/2$ we get

$$\left| f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right) - \exp\left(-\frac{\|\boldsymbol{\theta}\|^2}{2}\right)\right| \le \frac{\varepsilon M^2}{2} + \frac{M^4}{4K}.$$

Because ε is arbitrary, this gives the result.

Corollaries.

(1) By the continuity of π at $\boldsymbol{\theta}_\infty$ and Lemma A.1, $\pi(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta})f^K(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta})$ converges to $\pi(\boldsymbol{\theta}_\infty)\exp(-\|\boldsymbol{\theta}\|^2/2)$ uniformly on bounded sets.

(2) Lemma A.1 and Fatou's lemma give us $\pi(\boldsymbol{\theta}_\infty)|\boldsymbol{\Sigma}|(2\pi)^{p/2} \le \liminf_K c(K)K^{p/2}$. In particular, there is a constant $C > 0$ so that $\frac{1}{c(K)} \le CK^{p/2}$.

Lemma A.2. Under Assumptions A.1 and A.2, the following three are equivalent.

(a) $\Psi_K \Rightarrow N(\mathbf{0}, \mathbf{I}_p)$ (convergence in distribution to a Normal random variable).

(b) The density g_k converges pointwise to a multivariate standard normal density function. That is, $c(K)K^{p/2} \to \pi(\boldsymbol{\theta}_\infty)|\boldsymbol{\Sigma}|(2\pi)^{p/2}$.

(c) $\boldsymbol{\Theta}_K \Rightarrow \delta_{\boldsymbol{\theta}_\infty}$ where $\delta_{\boldsymbol{\theta}_\infty}$ indicates a degenerate distribution at $\boldsymbol{\theta}_\infty$.

Proof. To show (a) $\Rightarrow$ (b). The density $g_K(\cdot)$ can be written as

$$g_K(\boldsymbol{\theta}) = \frac{|\boldsymbol{\Sigma}|}{K^{p/2}c(K)}\pi\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right).$$

Let B be a bounded Borel set with positive Lebesgue measure. From the convergence in (a), we have

$$\frac{1}{(2\pi)^{p/2}}\int_B \exp(-\|\boldsymbol{\theta}\|^2/2)\,d\boldsymbol{\theta} = \lim_K \frac{|\boldsymbol{\Sigma}|}{c(K)K^{p/2}}\int_B \pi\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)d\boldsymbol{\theta}.$$

On the other hand, the uniform convergence from Lemma A.1 gives

$$\lim_K \int_B \pi\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)f^K\left(\boldsymbol{\theta}_\infty + \frac{1}{\sqrt{K}}\boldsymbol{\Sigma}\boldsymbol{\theta}\right)d\boldsymbol{\theta} = \pi(\boldsymbol{\theta}_\infty)\int_B \exp(-\|\boldsymbol{\theta}\|^2/2)\,d\boldsymbol{\theta}.$$

Hence we can conclude that $c(K)K^{p/2} \to \pi(\boldsymbol{\theta}_\infty)|\boldsymbol{\Sigma}|(2\pi)^{p/2}$ as K converges to infinity. Combined with convergence in Lemma A.1, this gives $g_K(\boldsymbol{\theta}) \to \frac{1}{(2\pi)^{p/2}}\exp(-\|\boldsymbol{\theta}\|^2/2)$.

(b) $\Rightarrow$ (a) follows from Scheffe's theorem.

(a) $\Rightarrow$ (c) is obvious.

To show (c) $\Rightarrow$ (b). Because $D^2 f$ and π are continuous at $\boldsymbol{\theta}_\infty$ and $\boldsymbol{\Sigma}$ is strictly positive definite, from (A.1) we see that for any $\varepsilon > 0$ we can find $\delta > 0$ so that $\boldsymbol{\theta} \in N(\delta)$ implies

$$\begin{aligned} &f(\boldsymbol{\theta}) < 1 - \tfrac{1}{2}(1-\varepsilon)(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)^T\boldsymbol{\Sigma}^{-2}(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty) \qquad \text{and} \\ &\pi(\boldsymbol{\theta}) \le (1+\varepsilon)\pi(\boldsymbol{\theta}_\infty). \end{aligned} \tag{A.3}$$

Also, by assumption (c), we may assume that K is so large that $1 - \varepsilon \le \int_{N(\delta)} \pi_K(\boldsymbol{\theta})\,d\boldsymbol{\theta}$. Multiplying this inequality by $c(K)K^{p/2}(1-\varepsilon)^{-1}$ and using (A.3) gives

$$\begin{aligned} c(K)K^{p/2} &\le (1-\varepsilon)^{-1}K^{p/2}\int_{N(\delta)} \pi(\boldsymbol{\theta})f^K(\boldsymbol{\theta})\,d\boldsymbol{\theta} \\ &\le (1-\varepsilon)^{-1}K^{p/2}(\pi(\boldsymbol{\theta}_\infty)+\varepsilon) \\ &\quad\times \int_{N(\delta)}\left[1 - \frac{1}{2}(1-\varepsilon)(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)^T\boldsymbol{\Sigma}^{-2}(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)\right]^K d\boldsymbol{\theta} \\ &\le (1-\varepsilon)^{-1}K^{p/2}(\pi(\boldsymbol{\theta}_\infty)+\varepsilon) \\ &\quad\times \int_{N(\delta)}\left[-\frac{K}{2}(1-\varepsilon)(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)^T\boldsymbol{\Sigma}^{-2}(\boldsymbol{\theta}-\boldsymbol{\theta}_\infty)\right]d\boldsymbol{\theta} \\ &= (1-\varepsilon)^{-1}(\pi(\boldsymbol{\theta}_\infty)+\varepsilon)|\boldsymbol{\Sigma}|(2\pi)^{p/2}. \end{aligned}$$

By letting $K \to \infty$ and then $\varepsilon \to 0$, we get $\limsup_K c(K)K^{p/2} \le \pi(\boldsymbol{\theta}_\infty)|\boldsymbol{\Sigma}|(2\pi)^{p/2}$. The other half comes from the inequality in Corollary A.1.

Corollary to Lemma A.2. Under Assumptions A.1, A.2, and A.3, $\boldsymbol{\Theta}_K \Rightarrow \delta_{\boldsymbol{\theta}_\infty}$.

Proof. Using Assumption A.3 and the second corollary to Lemma A.1, we see that for any $\delta > 0$,

$$\frac{1}{c(K)}\int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_\infty\|>\delta} \pi(x)f^K(x)\,dx \le CK^{p/2}\gamma(\delta)^K \to 0.$$

This implies $\boldsymbol{\Theta}_K \Rightarrow \delta_{\boldsymbol{\theta}_\infty}$.

The main result of the convergence of data-cloning algorithm, that under Assumptions A.1, A.2, and A.3, $\Psi_K \Rightarrow N(\mathbf{0}, \mathbf{I}_p)$, follows immediately.

Comment. The proof given in Jacquire, Johannes, and Polson (2007) assumes only Assumptions A.1 and A.2. The counter example below shows that they are not sufficient for convergence; Assumption A.3 is necessary. Let $\Theta = \Re$, $\pi(\theta) = \frac{1}{2.5}\min(1, \frac{1}{4\theta^2})$. Let the likelihood function be $f(\theta) = 1 - \frac{\theta^2}{2}$ when $|\theta| \le 1$ and $f(\theta) = 1 - \frac{1}{|\theta|^3}$ when $|\theta| > 1$. For this situation, $\sqrt{K}C(K) \to \infty$ and we do not get the convergence to a Normal distribution.

A.2 Determining Estimability

Theorem A.2. Consider the set $N(\boldsymbol{\theta}) = \{\boldsymbol{\theta} \in \Theta : L(\boldsymbol{\theta}, \mathbf{y}_{(n)}) = L(\hat{\boldsymbol{\theta}}_{(n)}, \mathbf{y}_{(n)})\}$. Suppose this set is not a single point set, that is, the likelihood function is identical over the set $N(\boldsymbol{\theta})$. As $K \to \infty$, the posterior distribution converges to a distribution with density $\frac{\pi(\boldsymbol{\theta})}{\int_{N(\boldsymbol{\theta})}\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$ for $\boldsymbol{\theta} \in N(\boldsymbol{\theta})$. If the set $N(\boldsymbol{\theta})$ is not a single point set, $\sigma^2_{K,n}$, the largest eigenvalue of the posterior variance matrix, does not converge to 0.

Proof. Consider

$$\frac{\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)})}{\pi_K(\boldsymbol{\theta}_{(n)}|\mathbf{y}_{(n)})} = \frac{\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}_{(n)})}\frac{f^K(\mathbf{y}_{(n)}|\boldsymbol{\theta})}{f^K(\mathbf{y}_{(n)}|\boldsymbol{\theta}_{(n)})}.$$

It is obvious that for $\boldsymbol{\theta} \notin N(\boldsymbol{\theta})$,

$$\frac{\pi_K(\boldsymbol{\theta}|\mathbf{y}_{(n)})}{\pi_K(\boldsymbol{\theta}_{(n)}|\mathbf{y}_{(n)})} = \frac{\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{\theta}_{(n)})}\frac{f^K(\mathbf{y}_{(n)}|\boldsymbol{\theta})}{f^K(\mathbf{y}_{(n)}|\boldsymbol{\theta}_{(n)})} \to 0.$$

It is also equally obvious that for $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in N(\boldsymbol{\theta})$, $\frac{\pi_K(\boldsymbol{\theta}_1|\mathbf{y}_{(n)})}{\pi_K(\boldsymbol{\theta}_2|\mathbf{y}_{(n)})} = \frac{\pi(\boldsymbol{\theta}_1)}{\pi(\boldsymbol{\theta}_2)}\frac{f^K(\mathbf{y}_{(n)}|\boldsymbol{\theta}_1)}{f^K(\mathbf{y}_{(n)}|\boldsymbol{\theta}_2)} = \frac{\pi(\boldsymbol{\theta}_1)}{\pi(\boldsymbol{\theta}_2)}$. Hence the result follows.

Corollary A.2. Let $g(\boldsymbol{\theta})$ be a function of $\boldsymbol{\theta}$ such that it takes unique value on the set $N(\boldsymbol{\theta})$. Then $g(\boldsymbol{\theta})$ is estimable.

An immediate consequence of the corollary, the posterior variance of $g(\boldsymbol{\theta})$ converges to 0 as we increase the number of clones. Hence a simple way to check for estimability of a specific function of $\boldsymbol{\theta}$ is to plot the posterior variance [or, the largest eigenvalue of the posterior variance matrix if $g(\boldsymbol{\theta})$ is a vector valued function] as a function of the number of clones. If this converges to 0 as we increase the clones, the function is estimable.

Corollary A.2. Let $\pi_1(\boldsymbol{\theta})$ and $\pi_2(\boldsymbol{\theta})$ be two different prior distributions. Then, it follows that, as $K \to \infty$, the posterior distributions converge to $\frac{\pi_1(\boldsymbol{\theta})}{\int_{N(\boldsymbol{\theta})}\pi_1(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$ and $\frac{\pi_2(\boldsymbol{\theta})}{\int_{N(\boldsymbol{\theta})}\pi_2(\boldsymbol{\theta})\,d\boldsymbol{\theta}}$ respectively. Hence the largest eigenvalue of the limiting posterior distribution depends on the choice of the prior distribution.

[Received December 2009. Revised July 2010.]

REFERENCES

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25. [1617,1619,1620]

Clark, J. S., and Gelfand, A. E. (eds.) (2006), *Hierarchical Modelling for the Environmental Sciences: Statistical Methods and Applications*, Oxford, U.K.: Oxford University Press. [1617,1620]

Clayton, D., and Kaldor, J. (1987), "Empirical Bayes Estimates of Age Standardized Relative Risks for Use in Disease Mapping," *Biometrics*, 43, 671–681. [1617,1620]

Crowder, M. J. (1978), "Beta-Binomial ANOVA for Proportions," *Applied Statistics*, 27, 34–37. [1619]

deValpine, P. (2004), "Monte Carlo State-Space Likelihoods by Weighted Posterior Kernel Density Estimation," *Journal of the American Statistical Association*, 99, 523–536. [1617]

Diggle, P., Liang, K. Y., and Zeger, S. L. (1994), *Analysis of Longitudinal Data*, Oxford, U.K.: Oxford University Press. [1617]

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-Based Geostatistics," *Applied Statistics*, 47, 299–350. [1620]

Doucet, A., Godsill, S. J., and Robert, C. P. (2002), "Marginal Maximum a Posteriori Estimation Using Markov Chain Monte Carlo," *Statistics and Computing*, 12, 77–84. [1617]

Eberly, L. E., and Carlin, B. P. (2000), "Identifiability and Convergence Issues for Markov Chain Monte Carlo Fitting of Spatial Models," *Statistics in Medicine*, 19, 2279–2294. [1622,1623]

Gelfand, A. E., and Sahu, S. K. (1999), "Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models," *Journal of the American Statistical Association*, 94, 247–253. [1622]

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (eds.) (1996), *Markov Chain Monte Carlo in Practice*, London: Chapman & Hall. [1617,1618]

Harris, I. R. (1989), "Predictive Fit for Natural Exponential Families," *Biometrika*, 76 (4), 675–684. [1619]

Jacquier, E., Johannes, M., and Polson, N. (2007), "MCMC Maximum Likelihood for Latent State Models," *Journal of Econometrics*, 137, 615–640. [1617,1624]

Johnson, R. A., and Wichern, D. W. (2007), *Applied Multivariate Statistical Analysis* (6th ed.), New Jersey: Prentice Hall. [1619]

Kuk, A. Y. C. (2003), "Automatic Choice of Driving Values in Monte Carlo Likelihood Approximation via Posterior Simulations," *Statistics and Computing*, 13, 101–109. [1617]

Lele, S. R. (2010), "Model Complexity and Information in the Data: Could It Be a House Built on Sand?" *Ecology*, to appear. [1620,1622,1623]

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," *Ecology Letters*, 10, 551–563. [1617,1618]

McCulloch, C. E. (1997), "Maximum Likelihood Algorithms for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 92, 162–170. [1617]

McCulloch, C. E., and Searle, S. R. (2001), *Generalized, Linear, and Mixed Models*, New York: Wiley. [1617,1620,1622]

Natarajan, R., and Kass, R. E. (2000), "Reference Bayesian Methods for Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 95, 227–237. [1623]

Ponciano, J. M., Taper, M. L., Dennis, B., and Lele, S. R. (2009), "Hierarchical Models in Ecology: Confidence Intervals, Hypothesis Testing, and Model Selection Using Data Cloning," *Ecology*, 90, 356–362. [1623]

Royle, J. A., and Dorazio, R. M. (2009), *Hierarchical Models and Inference in Ecology: The Analysis of Data From Populations, Metapopulations and Communities*, London, U.K.: Elsevier. [1617,1620]

Searle, S. R., Casella, G., and McCulloch, C. E. (1992), *Variance Components*, New York: Wiley. [1617]

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2004), *WinBUGS Version 1.4 User Manual*, London: MRC Biostatistics Unit, Institute of Public Health. [1617,1618,1620]

Thall, P. F., and Vail, S. C. (1990), "Some Covariance Models for Longitudinal Count Data With Overdispersion," *Biometrics*, 46, 657–671. [1619]

Walker, A. M. (1969), "On the Asymptotic Behavior of Posterior Distributions," *Journal of the Royal Statistical Society, Ser. B*, 31, 80–88. [1618]