Monte Carlo EM Estimation for Time Series Models Involving Counts

Author(s): K. S. Chan and Johannes Ledolter

Source: *Journal of the American Statistical Association*, Mar., 1995, Vol. 90, No. 429 (Mar., 1995), pp. 242-252

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2291149

Monte Carlo EM Estimation for Time Series Models Involving Counts

K. S. CHAN and Johannes LEDOLTER*

The observations in parameter-driven models for time series of counts are generated from latent unobservable processes that characterize the correlation structure. These models result in very complex likelihoods, and even the EM algorithm, which is usually well suited for problems of this type, involves high-dimensional integration. In this article we discuss a Monte Carlo EM (MCEM) algorithm that uses a Markov chain sampling technique in the calculation of the expectation in the E step of the EM algorithm. We propose a stopping criterion for the algorithm and provide rules for selecting the appropriate Monte Carlo sample size. We show that under suitable regularity conditions, an MCEM algorithm will, with high probability, get close to a maximizer of the likelihood of the observed data. We also discuss the asymptotic efficiency of the procedure. We illustrate our Monte Carlo estimation method on a time series involving small counts: the polio incidence time series previously analyzed by Zeger.

KEY WORDS: Asymptotic efficiency; Gibbs sampler; Latent process; Markov chain techniques; Parameter-driven models; Polio incidence series.

1. INTRODUCTION

Let $\{Y_t\}_{t=1,2,\ldots,n}$ be a time series of count data and let $\{U_t\}_{t=1,2,\ldots,n}$ be an observed vector-valued covariate process. We are interested in modeling the contemporaneous dependence of Y_t on the covariates in U_t as well as its dependence on past values. For count data, two classes of models have been considered in the literature: parameter-driven models and observation-driven models (see Cox 1981). Conceptually, the parameter-driven model is more elegant, but its main disadvantage is of a numerical nature. The parameter-driven model specifies an unobserved latent process and ordinarily results in a complex likelihood. But if one could observe the latent process, then the likelihood would be simplified greatly. Hence one may want to use the EM algorithm to maximize the likelihood. There is a complication, however. For count data, the E step is ordinarily intractable, because the conditional distribution of the latent process given the count data is complicated. Computing the E step by brute force (i.e., by numerical quadrature) seems to be out of the question. In this article we propose to estimate the conditional expectation via Markov chain sampling techniques, such as the Gibbs sampler and/or the Metropolis algorithm. A Markov chain sampling technique is used to carry out the E step. The resulting modified scheme, called the Monte Carlo EM (MCEM) algorithm, was first proposed by Wei and Tanner (1990). Guo and Thompson (1992) used the MCEM algorithm in the context of a complex genetic model. (For recent surveys on Markov chain sampling methods, see Besag and Green 1993; Geyer 1991; Smith and Roberts 1993; and Tierney 1994.)

In the exact EM algorithm, the E step imputes the unobserved log-likelihood of the complete data, consisting of the observed data and the latent process, by the conditional expectation of the complete-data log-likelihood given the observed data. In the MCEM algorithm the conditional expectation of the log-likelihood of the complete data is estimated by averaging the conditional log-likelihoods of simulated sets of complete data. An important property of the EM algorithm is that the likelihood of the observed data always increases along an EM sequence. (For other convergence properties of the EM algorithm, see Dempster, Laird, and Rubin 1977 and Wu 1983.) For the MCEM algorithm, this property is lost. But it is shown in this article that under suitable regularity conditions, an MCEM sequence will, with high probability, get close to a maximizer of the likelihood of the observed data.

The article is organized as follows. Section 2 consists of three subsections. In the first subsection we describe the MCEM algorithm; in the second subsection we summarize formulas for estimating the change of the log-likelihood of the observed data, the score statistic, and the Fisher information; and in the third subsection we describe a stopping rule. We discuss the application of the MCEM algorithm to a first-order parameter-driven model for series of count data in Section 3, using Zeger's polio incidence series as an example. Finally, we discuss convergence properties of the MCEM algorithm in Section 4. Proofs are collected in the Appendix.

2. THE MCEM ALGORITHM

2.1 Description of the Algorithm

Let $\mathbf{Y} = \mathbf{y}$ be the observed incomplete data and let $\mathbf{X}$ be the unobserved complete data of which $\mathbf{Y}$ is a measurable function. (Capital letters are used for random variables and corresponding small letters for observed values.) The parameter space is denoted by Ω and two arbitrary elements are denoted by θ and θ'. The EM algorithm consists of two steps:

$$E \text{ step:} \quad \text{form } Q(\theta' \mid \theta) = E_\theta(l_{\mathbf{X}}(\theta') \mid \mathbf{y});$$

$$M \text{ step:} \quad \text{maximize } Q(\cdot \mid \theta).$$

$E_\theta(\cdot \mid \mathbf{y})$ denotes the conditional expectation given $\mathbf{Y} = \mathbf{y}$, where θ is the true parameter and $l_{\mathbf{X}}(\theta')$ is the log-likelihood

* K. S. Chan is Associate Professor, Department of Statistics and Actuarial Science, and Johannes Ledolter is Professor, Department of Statistics and Actuarial Science and the Department of Management Sciences, University of Iowa, Iowa City, IA 52242. The research of K. S. Chan was partially supported by National Science Foundation Grant DMS 91-18626 and the Old Gold summer fellowship from the University of Iowa. The research of J. Ledolter was supported in part by a grant from the Midwest Transportation Center. The authors gratefully acknowledge the constructive comments of two editors, an associate editor, and the referees.

of $\mathbf{X}$. Suppose that for all θ, $Q(\cdot\,|\,\theta)$ has a unique global maximizer at $\mathbf{M}(\theta)$ and that $\mathbf{M}(\theta)$ is continuous in θ. Then an EM sequence $\{\theta_k\}$ is obtained from $\theta_{k+1} = \mathbf{M}(\theta_k)$, and $\{\theta_k\}$ is a Markov chain with a deterministic transition.

When the E step cannot be carried out explicitly, we obtain a Monte Carlo estimate as follows. For a fixed $\theta \in \Omega$, let $\{\mathbf{X}_j\}$ be an ergodic Markov chain whose invariant distribution is the conditional distribution of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$. To emphasize the dependence of $\mathbf{X}_j$ on θ, we sometimes write $\mathbf{X}_j(\theta)$. Then, form

$$\text{MCE step:}\quad Q^{(m)}(\theta'\,|\,\theta) = \sum_{1\le j\le m} l_{\mathbf{X}_j(\theta)}(\theta')/m.$$

Ergodicity implies that as $m \to \infty$, $Q^{(m)}(\theta'\,|\,\theta) \to Q(\theta'\,|\,\theta)$ a.s. The MCEM algorithm is the same as the EM algorithm, except that the E step is replaced by the MCE step.

We assume that for any fixed θ, the random objective function $Q^{(m)}(\cdot\,|\,\theta)$ attains a unique global maximum at $\mathcal{M}_m(\theta)$. When $m = \infty$, so that the E step is carried out exactly, the random variable $\mathcal{M}_m(\theta)$ becomes $\mathbf{M}(\theta)$.

2.2 Relative Likelihood, Score, and Standard Errors

It will be shown in Section 4 that with suitable starting values, an MCEM sequence will, with high probability, get close to the maximizer of the likelihood of the observed data. But due to the noisy character of the simulated log-likelihood of the complete data, the MCEM sequence may subsequently drift elsewhere. Hence it is important to monitor the change of the likelihood along the MCEM sequence. The values of the score and the computation of standard errors for the estimates are also of importance. The Markov chain sample can be used to supply estimates for these quantities. Here we assume that all functions are sufficiently regular to admit the required differentiation and allow the interchange of integration and differentiation.

Let $f_\theta(\mathbf{y})$ and $f_\theta(\mathbf{x})$ be the pdf's of $\mathbf{Y}$ and $\mathbf{X}$. Let $f_\theta(\mathbf{x}\,|\,\mathbf{y})$ be the conditional pdf of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$. Then

$$f_\theta(\mathbf{y})/f_{\theta'}(\mathbf{y}) = f_\theta(\mathbf{x})f_{\theta'}(\mathbf{x}\,|\,\mathbf{y})/(f_{\theta'}(\mathbf{x})f_\theta(\mathbf{x}\,|\,\mathbf{y})).$$

Hence multiplying $f_\theta(\mathbf{x}\,|\,\mathbf{y})$ on both sides and integrating out $\mathbf{x}$, we get

$$f_\theta(\mathbf{y})/f_{\theta'}(\mathbf{y}) = E_{\theta'}(f_\theta(\mathbf{X})/f_{\theta'}(\mathbf{X})\,|\,\mathbf{y}). \tag{1}$$

Let $\mathbf{D}$ be the differentiation operator with respect to θ. Louis (1982) has shown that

$$\mathbf{D}l_\mathbf{Y}(\theta) = E_\theta(\mathbf{D}l_\mathbf{X}(\theta)\,|\,\mathbf{y}) \tag{2}$$

$$-\mathbf{D}^2 l_\mathbf{Y}(\theta) = E_\theta(-\mathbf{D}^2 l_\mathbf{X}(\theta)\,|\,\mathbf{y}) - E_\theta(\mathbf{D}l_\mathbf{X}(\theta)\mathbf{D}^t l_\mathbf{X}(\theta)\,|\,\mathbf{y}) + \mathbf{D}l_\mathbf{Y}(\theta)\mathbf{D}^t l_\mathbf{Y}(\theta), \tag{3}$$

where the superscript t denotes the transpose.

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ be a Markov chain sample whose invariant distribution is the conditional distribution of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$, using θ_{j+1} as the parameter. Using Equation (1), we can estimate $\Delta l_\mathbf{Y}(\theta_j, \theta_{j+1}) = l_\mathbf{Y}(\theta_{j+1}) - l_\mathbf{Y}(\theta_j)$ by

$$\tilde{\Delta} l_\mathbf{Y}(\theta_j, \theta_{j+1}) = -\log\left(\left(\sum_{k=1}^m f_{\theta_j}(\mathbf{X}_k)/f_{\theta_{j+1}}(\mathbf{X}_k)\right)\Big/ m\right). \tag{4}$$

The relative log-likelihood $l_\mathbf{Y}(\theta_j) - l_\mathbf{Y}(\theta_0)$ can be estimated by $\sum_{k=0}^{j-1} \tilde{\Delta} l_\mathbf{Y}(\theta_k, \theta_{k+1})$. The formulas in (2) and (3) can be used to estimate the score and the observed Fisher information from the Markov chain sample.

2.3 Stopping Criterion for the MCEM Algorithm and Selection of the Monte Carlo Sample Size

Our objective is to maximize the log-likelihood $l_\mathbf{Y}(\cdot)$. In the implementation of the MCEM algorithm we need to specify m, the Monte Carlo sample size of the Gibbs sampler, and a stopping rule. Assume for the moment that the change in the log-likelihood $\Delta l_\mathbf{Y}(\theta, \tau) = l_\mathbf{Y}(\tau) - l_\mathbf{Y}(\theta)$ can be computed exactly. Under this scenario and assuming m fixed, an MCEM sequence $(\theta_j^{(m)})$ is stopped at the Kth iteration if the random variable $\Delta l_\mathbf{Y}(\theta, \mathcal{M}_m(\theta))$, where $\theta = \theta_K^{(m)}$, is sufficiently small stochastically. That is, the algorithm is stopped when the next iteration is unlikely to change the log-likelihood.

Once we have reached the maximum, changes in the log-likelihood from subsequent iterations, $\Delta l_\mathbf{Y}(\theta, \mathcal{M}_m(\theta))$, should vary around zero. Assume that $\Delta l_\mathbf{Y}(\theta, \mathcal{M}_m(\theta))$ has mean $\mu = \mu(m, \theta)$ and variance $\sigma^2 = \sigma^2(m, \theta)$. Then the interval $(\mu - L\sigma, \mu + L\sigma)$, where L is an integer such as 4, describes the *effective range* of $\Delta l_\mathbf{Y}(\theta, \mathcal{M}_m(\theta))$; Chebychev's inequality implies that, with probability no less than $100(1 - 1/L^2)\%$, $\Delta l_\mathbf{Y}(\theta, \mathcal{M}_m(\theta))$ lies within that interval. Thus a stopping criterion for judging the smallness of the random variable requires that (a) $\sigma \le \delta$, a predetermined level of tolerance, and (b) the effective range includes zero; that is, $\Delta l_\mathbf{Y}(\theta, \mathcal{M}_m(\theta))$ differs from zero by less than $2L\sigma$. The requirement that $\sigma \le \delta$ determines the Monte Carlo sample size m needed to satisfy the required level of tolerance.

In our case it is difficult to calculate $\Delta l_\mathbf{Y}(\theta, \tau)$ precisely. But, as discussed in Section 2.2, it can be estimated by

$$\tilde{\Delta} l_\mathbf{Y}(\theta, \tau) = -\log\left(\left(\sum_{k=1}^m f_\theta(\mathbf{X}_k)/f_\tau(\mathbf{X}_k)\right)\Big/ m\right),$$

where $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ is a stationary ergodic Gibbs sampler with invariant distribution equal to that of $\mathbf{X}$ given $\mathbf{Y} = \mathbf{y}$, with τ being the true parameter. The preceding stopping rule is modified by replacing $\Delta l_\mathbf{Y}(\theta, \tau)$ with $\tilde{\Delta} l_\mathbf{Y}(\theta, \tau)$.

To determine the appropriate value of m, we suggest that a pilot study be carried out with m set as a moderately large value, say m_1. Suppose that the log-likelihood of the MCEM iterates of θ attains the global maximum at $\theta_K^{(m_1)}$. Select a small stretch of iterates $\theta_j^{(m_1)}$, $K_1 \le j \le K_2$ around $\theta_K^{(m_1)}$. For example, take the 10 iterates that are immediately after $\theta_K^{(m_1)}$. For each θ equal to $\theta_j^{(m_1)}$, $K_1 \le j \le K_2$, replicate the MCEM algorithm for a single iteration l times to obtain l independent replicates of $\mathcal{M}_{m_1}(\theta)$ and l independent replicates of $\tilde{\Delta} l_\mathbf{Y}(\theta, \mathcal{M}_{m_1}(\theta))$. In each replication an independent Gibbs sampler is used to form $\tilde{\Delta} l_\mathbf{Y}(\theta, \mathcal{M}_{m_1}(\theta))$; recall that the Gibbs sampler is simulated with $\mathcal{M}_{m_1}(\theta)$ as the true parameter.

Let $\tilde{\Delta} l_{ij}$ be the jth replicate of $\tilde{\Delta} l_\mathbf{Y}(\theta_i^{(m_1)}, \mathcal{M}_{m_1}(\theta_i^{(m_1)}))$. Compute the pooled variance

$$s^2 = \frac{1}{(l-1)(K_2 - K_1 + 1)} \sum_{i=K_1}^{K_2} \sum_{j=1}^{l} (\tilde{\Delta} l_{ij} - \tilde{\Delta} l_{i\cdot})^2, \quad (5)$$

where $\tilde{\Delta} l_{i\cdot}$ is the average of the replicates of $\tilde{\Delta} l_{\mathbf{Y}}(\theta_i^{(m_1)}, \mathcal{M}_{m_1}(\theta_i^{(m_1)}))$. The pooled standard deviation s measures the "average" variability of $\tilde{\Delta} l_{\mathbf{Y}}(\theta, \mathcal{M}_{m_1}(\theta))$ for a fixed θ within a "neighborhood" of the "maximizer" $\theta_K^{(m_1)}$.

It is shown in the Appendix that asymptotically, s is inversely proportional to m. The "flatness" of the log-likelihood around a maximizer is the main reason for the $1/m$ asymptotics, which is different from the usual $1/\sqrt{m}$ asymptotics for s. The foregoing result can be used to determine the Monte Carlo sample size m necessary to satisfy $\sigma \le \delta$. Suppose that the Monte Carlo sample size m_1 leads to $s = s_1$. Let m be the required sample size so that $\sigma \le \delta$. Because $m\sigma \approx m_1 s_1$, we require $m_1 s_1 / m \le \delta$ and hence $m \ge m_1 s_1 / \delta$. We then restart the MCEM algorithm with $\theta_K^{(m_1)}$ as the starting value and m equal to the smallest integer larger than $m_1 s_1 / \delta$. The iterations are stopped at the Kth iterate when $\tilde{\Delta} l_{\mathbf{Y}}(\theta_K^{(m)}, \mathcal{M}_m(\theta_K^{(m)}))$ differs from zero by less than $2L\sigma$ in magnitude. Note that $\sigma \approx m_1 s_1 / m$ is known when m is fixed.

Problems with this strategy could arise if (a) the $1/m$ asymptotics have not yet taken hold for the Monte Carlo sample size adopted in the pilot study, and/or (b) the unlikely event that $\tilde{\Delta} l_{\mathbf{Y}}(\theta_K^{(m)}, \mathcal{M}_m(\theta_K^{(m)}))$ falls outside the effective range occurs. To guard against these two possibilities, we suggest continuing the algorithm after the Kth iterate for a few more iterations, repeating the procedure for computing s, and confirming that s is not much larger than δ. A finding that s is much larger than δ would indicate that the $1/m$ asymptotics may not yet be applicable. In this case we would recommend that a new larger sample size be calculated, assuming that the $1/m$ asymptotics now hold for the sample size not less than that of the second run of the MCEM algorithm. If necessary, the procedure can be repeated until s is smaller than the stated tolerance level. The implementation of the preceding scheme and some empirical evidence for the $1/m$ asymptotics are illustrated in the next section.

3. PARAMETER-DRIVEN MODELS FOR COUNT DATA

3.1 The Model and the Application of the MCEM Algorithm

We now discuss a generalized regression model for a time series of count data. To account for the dependence of the count data on its past values, we use a simple parameter-driven model. We assume that $\{W_t\}$ is a stationary Gaussian AR(1) latent process; that is, $W_t = \rho W_{t-1} + \varepsilon_t$, where $\{\varepsilon_t\}$ is iid $N(0, \sigma_\varepsilon^2)$. Given this latent process $\{W_t\}$, the observations $\{Y_t\}$ are generated independently from a Poisson distribution with mean λ_t satisfying

$$\log \lambda_t = \alpha' U_t + W_t. \quad (6)$$

The parameter vector consists of the coefficients in (6) and the parameters of the latent process; that is, $\theta = (\alpha', \rho, \sigma_\varepsilon^2)'$. Our analysis is carried out conditional on the covariate process, which is assumed to be deterministic. The likelihood of the observed data does not have a simple closed form, and maximum likelihood (ML) estimation is intractable. Zeger (1988) proposed an estimating equation approach for the estimation of the parameters. His approach assumes a stationary autoregressive latent process but does not specify any distributional assumption on the latent process. An advantage of Zeger's approach is that only first- and second-order moments need to be specified. But for higher-order autoregressive cases, this method need not yield admissible parameter estimates.

Let $X_i = (Y_i, W_i)$, $\mathbf{Y} = (Y_1, \ldots, Y_n)$, $\mathbf{W} = (W_1, \ldots, W_n)$, and $\mathbf{X} = (X_1, \ldots, X_n)$. Here we apply the MCEM algorithm to obtain approximate maximum likelihood estimates. The MCE step estimates the conditional expectation $E_\theta(l_{\mathbf{X}}(\theta') | \mathbf{y})$. Apart from constants and terms independent of θ', the log-likelihood of $\mathbf{X}$ (conditional on the initial latent variable W_1) is given by

$$\begin{aligned} l_{\mathbf{X}}(\theta) &= \sum_{i=1}^{n} \{-\exp(W_i + \alpha' U_i) + Y_i W_i + Y_i \alpha' U_i\} \\ &\quad - (n-1)/2 \log(\sigma_\varepsilon^2) \\ &\quad - \sum_{i=1}^{n-1} (W_{i+1} - \rho W_i)^2 / (2\sigma_\varepsilon^2) \\ &= \sum_{i=1}^{n} \{-A_i \exp(\alpha' U_i) + Y_i \alpha' U_i\} \\ &\quad + E - (n-1)/2 \log(\sigma_\varepsilon^2) \\ &\quad - (B - 2\rho D + C\rho^2)/(2\sigma_\varepsilon^2), \end{aligned} \quad (7)$$

where $A_i = \exp(W_i)$, $B = \sum_1^{n-1} W_{i+1}^2$, $C = \sum_1^{n-1} W_i^2$, $D = \sum_1^{n-1} W_i W_{i+1}$, and $E = \sum_1^n W_i Y_i$. Clearly, $l_{\mathbf{X}}(\theta)$ is linear in A_i, B, C, D, and E, which are measurable functions of $\mathbf{X}$. Furthermore,

$$\begin{aligned} Q^{(m)}(\cdot | \theta) &= \sum_{i=1}^{n} \{-\bar{A}_i \exp(\alpha' U_i) + Y_i \alpha' U_i\} \\ &\quad + \bar{E} - (n-1)/2 \log(\sigma_\varepsilon^2) \\ &\quad - (\bar{B} - 2\rho\bar{D} + \bar{C}\rho^2)/(2\sigma_\varepsilon^2), \end{aligned}$$

where, for example, $\bar{A}_i = \sum_{j=1}^m A_i(\mathbf{X}_j)/m$ and $\{\mathbf{X}_j\}$ is a Markov chain sample of $\mathbf{X}$ given $\mathbf{y}$ and with θ as the true parameter. Note that $Q^{(m)}(\theta' | \theta)$ is concave in θ'. This is the reason for conditioning on W_1, because without conditioning, $Q^{(m)}(\cdot | \theta)$ is no longer concave. For fixed θ, the ρ and the σ_ε^2 component of the maximizer of $Q^{(m)}(\cdot | \theta)$ are given by $\hat{\rho} = \bar{D}/\bar{C}$ and $\hat{\sigma}_\varepsilon^2 = (\bar{B} - \bar{D}^2/\bar{C})/(n-1)$. The α component of the maximizer is obtained numerically via the Davidon–Fletcher–Powell algorithm, subroutine DFPMIN (see Press, Flannery, Teukolsky, and Vetterling 1986, p. 310).

We must sample from the conditional distribution of $\mathbf{X}$. It suffices to sample from the conditional distribution of $\mathbf{W}$ and use the Gibbs sampler to generate the required Markov chain. By letting $\mathbf{W}^{-i}$ be $\mathbf{W}$ with W_i being omitted, we have

$$\begin{aligned} f_\theta(w_i | \mathbf{w}^{-i}, \mathbf{y}) &\propto f_\theta(\mathbf{x}) \\ &= f_\theta(\mathbf{y} | \mathbf{w}) f_\theta(\mathbf{w}) \propto f_\theta(y_i | w_i) f_\theta(w_i | \mathbf{w}^{-i}). \end{aligned} \quad (8)$$

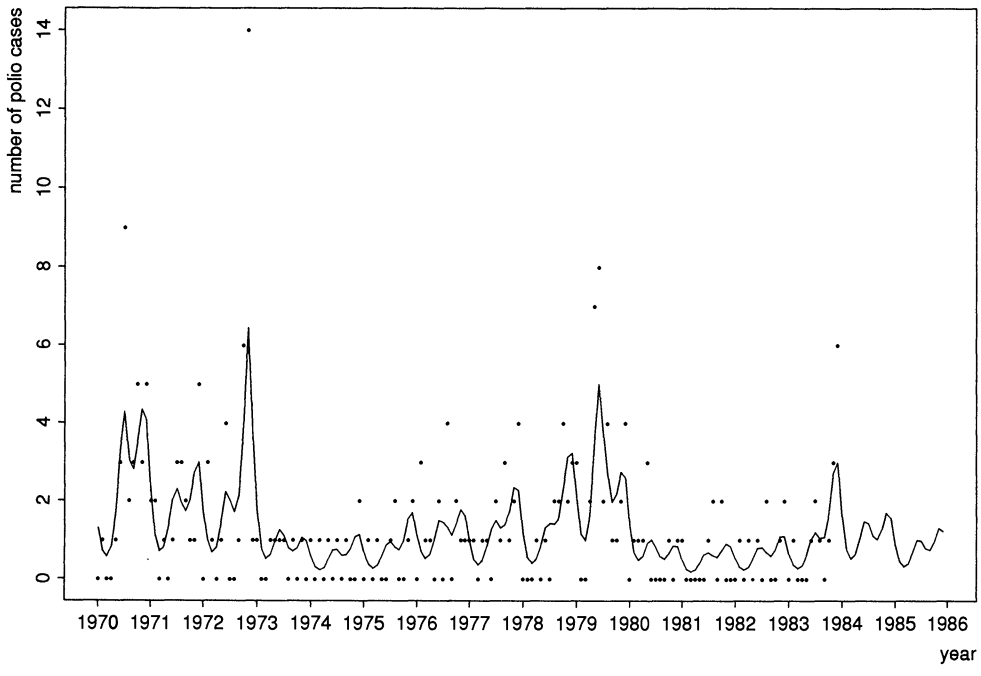

Figure 1. Monthly Numbers of Polio Cases. Dots represent observations. The solid line connects the smoothed means for 1970 to 1983 and the forecasts for 1984 and 1985.

The Gaussianity of $\mathbf{W}$ implies that the conditional distribution of W_i, given the rest of $\mathbf{W}$, is $N(\mu_i, \sigma_i^2)$ for some μ_i and σ_i^2. It is more convenient to sample $V_i = W_i + \alpha' U_i$. Up to an additive constant,

$$\log f_\theta(v_i \mid \mathbf{w}^{-i}, \mathbf{y}) = -\exp(v_i) + v_i y_i - (v_i - \gamma_i)^2/(2\sigma_i^2), \quad (9)$$

where $\gamma_i = \mu_i + \alpha' U_i$. The sampling of V_i from the foregoing log-concave density can be done via a universal rejection algorithm (see Devroye 1984).

The moments μ_i and σ_i^2 can be derived as follows. Routine algebraic calculations show that for $i \neq 1, n$, the conditional distribution of W_i, given $\mathbf{W}^{-i}$, is $N(\rho(w_{i-1} + w_{i+1})/(1 + \rho^2), \sigma_\varepsilon^2/(1 + \rho^2))$. For $i = n$, it is $N(\rho w_{n-1}, \sigma_\varepsilon^2)$. For $i = 1$, it is $N(\rho w_2, \sigma_\varepsilon^2)$.

The MCEM algorithm of Section 2 can be readily applied. Furthermore, because $E_\theta(\lambda_t \mid \mathbf{y}) = \exp(\alpha' U_t) E_\theta(\exp(W_t) \mid \mathbf{y})$ for all t, smoothing and prediction are easily carried out. The l-step prediction can be obtained from

$$E_\theta(\lambda_{n+l} \mid \mathbf{y}) = \exp(\alpha' U_{n+l} + .5\sigma_\varepsilon^2(1 - \rho^{2l})/(1 - \rho^2)) \times E_\theta(\exp(\rho^l W_n \mid \mathbf{y}). \quad (10)$$

The conditional expectations can be estimated via Markov chain sampling. This approach is easily modified to handle missing values and irregular sampling intervals.

The generalization of the AR(1) model to higher-order autoregressive models is straightforward. Indeed, a richer generalization is possible. The latent process $\{W_t\}$ can be represented as a noisy observation of a latent process that forms an inhomogeneous Markov chain; that is,

$$Z_{t+1} = F_t Z_t + G_t R_t \quad (11)$$

and

$$W_t = H_t Z_t + V_t. \quad (12)$$

Here it is assumed that the coefficient matrices F_t, H_t, and G_t are known up to a parameter θ, which includes α. $\{R_t\}$ and $\{V_t\}$ are independent Gaussian iid sequences. We say that W_t is a perfect observation of Z_t if V_t is identically zero. It is well known that any autoregressive integrated moving average (ARIMA) model can be put into such a Markovian representation.

Gibbs sampling can be carried out readily as

$$f_\theta(w_i \mid \mathbf{w}^{-i}, \mathbf{z}, \mathbf{y}) \propto f_\theta(y_t \mid w_i) f_\theta(w_i \mid z_i) \quad (13)$$

and

$$f_\theta(z_i \mid \mathbf{z}^{-i}, \mathbf{w}, \mathbf{y}) \propto f_\theta(w_i \mid z_i) f_\theta(z_i \mid \mathbf{z}^{-i}). \quad (14)$$

The fixed-point smoothing formulas of Anderson and Moore (1979, pp. 172–173) can be applied to derive $f_\theta(z_i \mid \mathbf{z}^{-i})$.

3.2 Example: Polio Incidence

We illustrate the MCEM algorithm on an example taken from Zeger (1988), who analyzed the monthly number of cases of poliomyelitis from January 1970–December 1983. A time sequence plot of the polio data is shown in Figure 1. The solid line connects the smoothed means from the fitted model that is discussed herein. A central question studied by Zeger (1988) is whether the data follow a decreasing time trend. There is seasonality in the data, and it is modeled with trigonometric components involving the first two harmonics. The covariate vector in (6) is given by $U_t = (1, t/1{,}000, \cos(2\pi t/12), \sin(2\pi t/12), \cos(2\pi t/6), \sin(2\pi t/6))'$. Zeger's model is slightly different from our model in Section 3. Zeger assumed that $\exp(W_t)$ follows a stationary AR(1) process, but did not assume that W_t is Gaussian. Furthermore, Zeger used the estimating equation method to estimate the parameters.

Figure 2. Log-Likelihood for Iterations 10 Through 300.

We adopt the latent AR(1) model described in the previous section. The MCEM algorithm is used to carry out the estimation. The starting values of α are obtained by fitting a log-linear model to the data, assuming no temporal dependence. The starting values for ρ and σ_ε^2 are 0 and 1.0. The Markov chain sample size, m, is initially set to 200. The Gibbs sampling in the first MCE step starts with all W's set to zero. A run of 400 **W**'s is simulated by the Gibbs sampler. The initial 200 **W**'s are discarded as transient values. The last 200 **W**'s are retained to compute $Q^{(m)}(\cdot \mid \theta)$. The final value of **W** from the previous run is then used as the starting value of the Gibbs sampler in the next MCE step, and a run of 220 **W**'s is generated. The first 20 **W**'s are discarded as transient values. This procedure is then repeated in subsequent MCE steps. The time sequence plot of the estimated log-likelihood (of the observed data) for 300 iterations is shown in Figure 2. The log-likelihood is maximized at the 219th iteration and subsequently moves into a slightly lower likelihood region. Changes for selected coefficients are shown in Figure 3. After about 200 iterations, most coefficients stabilize, although the intercept and the slope display trend movements that may be of complementary nature. The pooled standard deviation s defined by formula (5) is estimated as .0621. Details of the computation and empirical evidence for the $1/m$ asymptotics are reported in Section 3.3. The parameter values at the 219th iteration are used as starting values for a new MCEM iteration. The Markov chain sample is increased to 2,000, which, according to our theory, should reduce s to one-tenth of its value. Simulation results in Section 3.3 confirm this reduction, as s for $m = 2{,}000$ is estimated as .0058. This implies that the MCEM algorithm maximizes the likelihood within a tolerance of about .5%. After nine iterations, the algorithm is stopped. All estimated changes of the log-likelihood are smaller than .02 in magnitude, which is well within the effective range for such changes. The values of the parameters change only slightly; the first two significant digits are virtually unchanged. The approximate ML estimates are given by

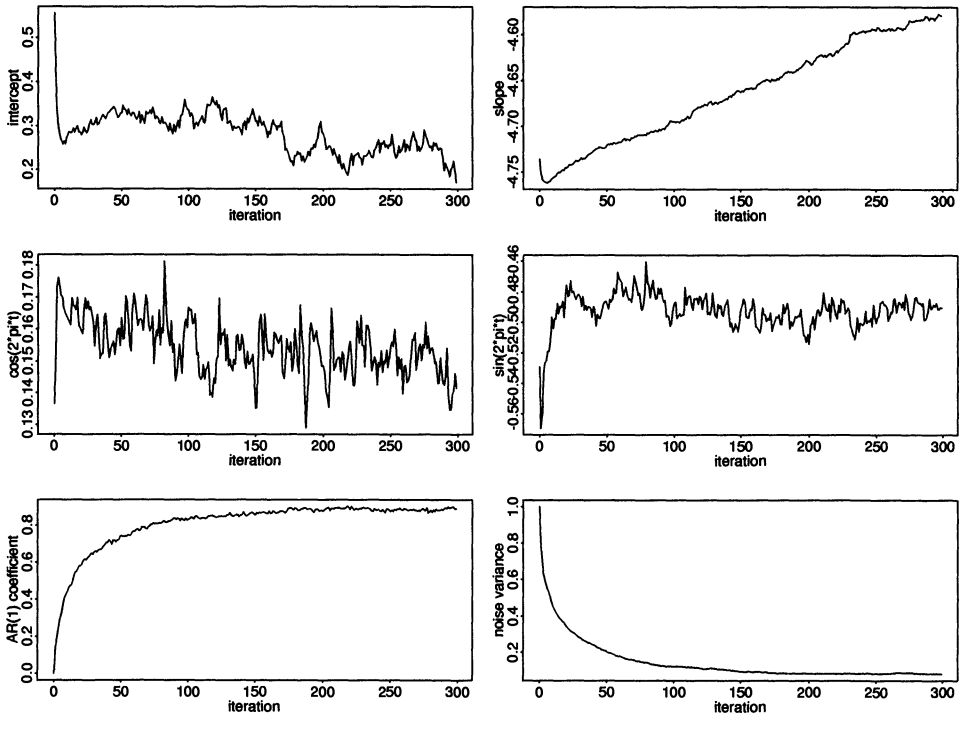

Figure 3. Changes for Selected Coefficient Estimates.

$$\hat{\alpha}' = (.211 \quad -4.62 \quad .149 \quad -.495 \quad .439 \quad -.0418)$$
$$(.125) \quad (1.38) \quad (.0899) \quad (.116) \quad (.102) \quad (.0986)$$
$$\hat{\rho} = .894 \quad (.0355) \qquad \hat{\sigma}_\varepsilon^2 = .0824.$$

Because the observation in November 1972 appears to be an outlier, we augment U_t by the indicator function $I_{\text{Nov},1972}$, which is equal to 1 in November 1972 and to zero elsewhere. The approximate ML estimates, with the estimate of the indicator coefficient added last, are

$$\hat{\alpha}' = (-.0262 \quad -3.67 \quad .0859 \quad -.456 \quad .402 \quad .0213 \quad 1.95)$$
$$(.134) \quad (1.41) \quad (.0930) \quad (.117) \quad (.104) \quad (.102) \quad (.308)$$
$$\hat{\rho} = .958 \quad (.0236) \qquad \hat{\sigma}_\varepsilon^2 = .0336.$$

The figures within parentheses are standard errors, obtained from Equation (3). The last term on the right side of (3) is zero at the ML estimate and is dropped in the calculation.

Because Zeger (1988) modeled $\xi_t = \exp(W_t)/E(\exp(W_t))$, his and our results are not directly comparable. But the following calculations shed light on their relationships. Because $E(\exp(W_t)) = \exp(.5\sigma_W^2)$, where $\sigma_W^2 = \sigma_\varepsilon^2/(1 - \rho^2)$, and because ξ_t in Zeger's model is normalized to have mean 1, we need to subtract $.5\hat{\sigma}_W^2$ from our intercept estimate to make them comparable. Letting σ_ξ^2 and $\rho_\xi(1)$ be the variance and the first lag autocorrelation of ξ_t, we obtain $\sigma_\xi^2 = \exp(\sigma_W^2) - 1$ and $\rho_\xi(1) = (\exp(\rho\sigma_W^2) - 1)/(\exp(\sigma_W^2) - 1)$. Based on these expressions, the ML estimate for σ_ξ^2 and $\rho_\xi(1)$ can be obtained. Our results, as well as those reported by Zeger, are displayed in Table 1. In general, the estimates are fairly similar; however, a few differences should be pointed out:

1. Our ML estimate of the autoregressive coefficient is closer to 1 than Zeger's estimate.
2. Adjusting for the outlier reduces the intercept and increases the slope.
3. Our slope estimate in the model that adjusts for the outlier is smaller, in absolute value, than the one reported by Zeger, but significantly different from zero. Zeger reported a t ratio of 1.62, which is not significant at the .05 level.
4. Standard errors for the ML estimates are uniformly smaller than the ones obtained by Zeger. In particular, the standard error for the slope parameter is reduced by 50%.

The solid line in Figure 1 connects the smoothed means computed from the estimated model that includes an indicator for the outlier. The solid curve beyond the data range represents the predicted mean values of Y_t.

In the preceding analysis we assume that the latent process is Gaussian and that a first-order autoregression is adequate to capture the temporal dependence of the data. These assumptions need to be checked. Apart from a recent paper by Harvey and Koopman (1992), the literature gives little guidance on how to check adequacy of assumptions made for latent processes. More work is needed, and we are currently developing the relevant diagnostic tools.

3.3 Empirical Evidence for the $1/m$ Asymptotics

We discuss in Section 2.3 and show in the Appendix that in a small neighborhood of the maximizer of the log-likelihood, the pooled standard deviation of the estimated change in the log-likelihood from one iteration to the next is inversely proportional to the Monte Carlo sample size m. This result is used in constructing the stopping rule. Here we provide empirical evidence for this result.

We take the parameter values at the 219th iteration, $\hat{\theta}$, and replicate the MCEM algorithm for the next $l = 10$ times. This results in 10 independent replicates of $\mathcal{M}_m(\hat{\theta})$, the estimate at the next iteration, and $\tilde{\Delta}l_Y(\hat{\theta}, \mathcal{M}_m(\hat{\theta}))$, the estimated change in the log-likelihood. We repeat this procedure for several different values in the neighborhood around the

Table 1. Coefficients and Standard Errors

	(a) $\hat{\theta}$	Standard error	(b) $\hat{\theta}$	Standard error	(c) $\hat{\theta}$	Standard error
Intercept (log rate in January 1976)	.17	.13	.42		.20	
Trend	−4.35	2.68	−4.62	1.38	−3.67	1.41
$\cos(2\pi t/12)$	−.11	.16	.15	.09	.09	.09
$\sin(2\pi t/12)$	−.48	.17	−.50	.12	−.46	.12
$\cos(2\pi t/6)$	.20	.14	.44	.10	.40	.10
$\sin(2\pi t/6)$	−.41	.14	−.04	.10	.02	.10
$I_{\text{Nov},1972}$					1.95	.31
$\hat{\sigma}_\xi^2$	.77		.54		.51	
$\hat{\rho}_\xi(1)$	.77		.88		.95	

NOTE: (a) Zeger's parameter-driven model using the method of estimating equations; (b) parameter-driven model in Section 3 using a MCEM algorithm; and (c) same as (b) but with outlier adjustment. Note that the intercept estimates for models (b) and (c) are adjusted to make them comparable with Zeger's results.

maximizer. This is carried out by selecting one of the iterates $\mathcal{M}_m(\hat{\theta})$ at random and replicating the MCEM 10 times. This procedure is repeated until we have a total of 10 independent replications for 10 parameter values in the neighborhood of the maximizer of the log-likelihood. This procedure is carried out for several different Monte Carlo sample sizes. The pooled standard deviations are given in Table 2.

Table 2 shows that $1/s$ is roughly proportional to the Monte Carlo sample size m. Doubling the Monte Carlo sample size will cut the standard deviation in half. In our simulations we also checked whether or not successive changes in the log-likelihood were centered at zero. Our tests confirmed that in the neighborhood of the maximizer, fluctuations in log-likelihood changes are centered around zero.

Table 2. Pooled Standard Deviations of Changes in the Log-Likelihood Function From Successive Iterations in the Neighborhood of the Maximizer

Monte Carlo sample size m	Standard deviation s	Inverse of standard deviation $1/s$
100	.14920	6.7
200	.06207	16.1
600	.01651	60.6
1,000	.01216	82.2
2,000	.00578	173.2

4. CONVERGENCE PROPERTIES OF THE MCEM ALGORITHM

4.1 Convergence to the Maximizer

We need additional assumptions to establish the convergence properties of the MCEM algorithm. We assume that for all θ', $l_{\mathbf{X}}(\theta') = q(\mathbf{Z}, \theta')$, where $\mathbf{Z}$ is a measurable vector function of $\mathbf{X}$ and q is linear in $\mathbf{Z}$. The dimension of $\mathbf{Z}$ is not assumed constant and may grow with that of $\mathbf{X}$. This assumption is satisfied for the model defined by Equation (6), with $\{W_t\}$ being a Gaussian autoregressive moving average (ARMA) process. The linearity assumption is adopted to simplify the analysis that follows; see the remark at the end of this section for a more general condition. Let Ξ be the convex hull of the support of $\mathbf{Z}$, which is assumed to be the same for all θ. It follows from the linearity of q in $\mathbf{Z}$ that $Q^{(m)}(\theta' \mid \theta) = q(\bar{\mathbf{Z}}_m, \theta')$ and $Q(\theta' \mid \theta) = q(E_\theta(\mathbf{Z}), \theta')$. We also assume that for all $\mathbf{Z} \in \Xi$, $q(\mathbf{Z}, \cdot)$ attains the unique global maximum at $\mathcal{M}(\mathbf{Z})$ and $\mathcal{M}(\mathbf{Z})$ is continuous in $\mathbf{Z}$. Here $\bar{\mathbf{Z}}_m = \bar{\mathbf{Z}}_m(\theta)$ is the sample mean of $\mathbf{Z}$'s computed from the Markov chain sample $\mathbf{X}_1(\theta), \mathbf{X}_2(\theta), \ldots, \mathbf{X}_m(\theta)$. For a fixed θ, ergodicity implies that $\bar{\mathbf{Z}}_m \to E_\theta(\mathbf{Z})$ a.s. as $m \to \infty$. But a somewhat stronger convergence will be needed in our result, and we assume that the convergence, in probability, of $\bar{\mathbf{Z}}_m \to E_\theta(\mathbf{Z})$ is uniform over compact subsets of Ω. Under these assumptions, the MCEM sequence $\{\theta_k^{(m)}\}$ obtained from $\theta_{k+1}^{(m)} = \mathcal{M}(\bar{\mathbf{Z}}_m(\theta_k^{(m)}))$ forms a Markov chain. Given $\theta_k^{(m)} = \theta$,

$$\theta_{k+1}^{(m)} = \mathbf{M}(\theta) + (\mathcal{M}(\bar{\mathbf{Z}}_m(\theta)) - \mathbf{M}(\theta)). \qquad (15)$$

The last term on the right side of (15) suggests that the transition of the MCEM sequence is a random perturbation of that of the EM algorithm. This is useful in understanding the behavior of the MCEM algorithm. Because $\mathcal{M}(E_\theta(\mathbf{Z})) = \mathbf{M}(\theta)$, the perturbation is ordinarily stochastically small for large m. We shall assume that $l_{\mathbf{Y}}(\theta)$ is continuous in θ. It is known that $l_{\mathbf{Y}}(\mathbf{M}(\theta)) \geq l_{\mathbf{Y}}(\theta)$, where $l_{\mathbf{Y}}(\theta)$ is the log-likelihood function of $\mathbf{Y}$. Wu (1983) showed that under suitable regularity conditions, an EM sequence converges to a stationary point of $l_{\mathbf{Y}}(\theta)$. The limit of an EM sequence, if it exists, is a fixed point of $\mathbf{M}(\theta)$; that is, $\mathbf{M}(\theta) = \theta$. Let $\mathbf{M}^k$ stand for the composition of $\mathbf{M}$ with itself for k times. If a fixed point is in the interior of Ω and $Q(\theta' \mid \theta)$ is differentiable in θ', then it is also a stationary point.

In our discussion we rely on definitions and results borrowed from the stability theory of dynamical systems (see Lasalle 1976). A fixed point, θ^*, of $\mathbf{M}$ is said to be asymptotically stable if (a) for any neighborhood V_1 of θ^*, there exists a neighborhood V_2 of θ^* such that for all k, $\mathbf{M}^k(\theta) \in V_1$ whenever $\theta \in V_2$; and (b) there exists a neighborhood V of θ^* such that $\mathbf{M}^k(\theta) \to \theta^*$ as $k \to \infty$ whenever $\theta \in V$. Condition (a) means that EM iterations stay close to θ^* whenever the starting value is sufficiently close to θ^*. Condition (b) means that an EM sequence converges to θ^* whenever the starting value is sufficiently close to θ^*. When Condition (a) holds, the fixed point is said to be stable. A stationary point of $l_{\mathbf{Y}}(\theta)$ is said to be isolated if it has a neighborhood in which it is the only stationary point.

Two results are stated next. The first result says that an isolated local maximizer of the log-likelihood surface is also an asymptotically stable fixed point of M. Hence the EM iterations converge to an isolated local maximizer if the starting value is close to the local maximizer. More importantly, asymptotic stability of a fixed point implies that the fixed point is stable under uniformly small additive perturbation of M (see Lasalle 1976, chap. 1, sec. 11). But the perturbation term on the right side of (15) ordinarily can assume large values, although the probability is small for large m. Nevertheless, Theorem 1 (which follows) says that for suitable starting values close to an isolated maximizer of the log-likelihood of the observations and with sufficiently large Markov chain sample size, an MCEM sequence will get close to the maximizer with high probability. Proofs of the following results are rather technical and are given in the Appendix.

Lemma 1. If θ^* is an isolated local maximizer of $l_{\mathbf{Y}}(\theta)$, then it is an asymptotically stable fixed point of $\mathbf{M}$.

Remark.

1. In discussing convergence of the EM algorithm, Wu (1983) did not give checkable conditions under which an EM sequence converges, except in the special case of a unimodal likelihood. Rather, he showed that the sequence of likelihoods converges.

2. In general, it is nontrivial to check whether every stationary point of the likelihood function is isolated and/or a local optimum. But for a particular stationary point, say θ^*, we can estimate the observed Fisher information matrix of the log-likelihood at θ^*; see Section 2.2. An examination of the eigenvalues of the information matrix reveals the nature

of the stationary point. If the eigenvalues are nonzero and of the same sign, then the stationary point is an isolated optimum.

Theorem 1. Suppose that the assumptions stated in the beginning of this subsection are satisfied. Let θ^* be an isolated local maximizer of $l_{\mathbf{Y}}(\theta)$. Then there exists a neighborhood, say V_1, of θ^* such that for starting values of the MCEM algorithm inside V_1 and for all $\varepsilon_1 > 0$, there exists a k_0 such that $\text{Prob}(|\theta_k^{(m)} - \theta^*| < \varepsilon_1 \text{ for some } k \le k_0) \to 1$ as $m \to \infty$.

Remark. As can be seen from the proof of Theorem 1 in the Appendix, the assumptions that $l_{\mathbf{X}}(\theta') = q(\mathbf{Z}, \theta')$ is linear in $\mathbf{Z}$ and that $\bar{\mathbf{Z}}_m \to E_\theta(\mathbf{Z})$, uniformly for θ over compact sets in probability, are needed only to guarantee that $\mathcal{M}_m(\theta)$ converges in probability to $\mathbf{M}(\theta)$ uniformly for θ over compact sets. Hence the former two assumptions can be replaced by the latter assumption on the convergence of $\mathcal{M}_m(\theta)$ to $\mathbf{M}(\theta)$.

4.2 Asymptotic Efficiency

The purpose of this subsection is to show that under suitable regularity conditions and in a small neighborhood of the ML estimate, an MCEM sequence can be approximated by an asymptotically stationary AR(1) process. Furthermore, the noise variance of the AR(1) process is inversely proportional to m, suggesting that for large values of m, the approximate ML estimate obtained from the MCEM algorithm is asymptotically as efficient as the ML estimate. We first derive the limiting distribution of $\mathcal{M}(\theta)$. In the following we write $\mathbf{D}_1$ for the partial derivative with respect to θ' and $\mathbf{D}_{11}$ for the second derivative with respect to θ'. If there is little risk of confusing the variable of differentiation, then we simply write $\mathbf{D}$ and $\mathbf{D}^2$ for the first and second derivatives. Expanding $\mathbf{D}_1 Q^{(m)}(\cdot\,|\theta)$ around $\mathbf{M}(\theta)$ and assuming that higher-order terms are negligible, we get

$$\sqrt{m}(\mathcal{M}_m(\theta) - \mathbf{M}(\theta)) = (-\mathbf{D}_{11}Q^{(m)}(\mathbf{M}(\theta)|\theta))^{-1}\sqrt{m}\mathbf{D}_1 Q^{(m)}(\mathbf{M}(\theta)|\theta) + o_p(1). \quad (16)$$

Assuming the law of large numbers, $-\mathbf{D}_{11}Q^{(m)}(\mathbf{M}(\theta)|\theta) \to \mathbf{V}_\theta = E_\theta(-\mathbf{D}_{11}l_{\mathbf{X}}(\mathbf{M}(\theta))|\mathbf{y})$ a.s. Suppose it holds that

$$\mathbf{D}_1 E_\theta(l_{\mathbf{X}}(\theta')|\mathbf{y}) = E_\theta(\mathbf{D}_1 l_{\mathbf{X}}(\theta')|\mathbf{y}).$$

Then $E_\theta(\mathbf{D}_1 Q^{(m)}(\mathbf{M}(\theta)|\theta))) = \mathbf{D}_1 Q(\mathbf{M}(\theta)|\theta) = \mathbf{0}$, because $\mathbf{M}(\theta)$ is the global maximizer of $Q(\cdot\,|\theta)$. The variance of $\sqrt{m}\mathbf{D}_1 Q^{(m)}(\mathbf{M}(\theta)|\theta)$ is equal to $\sum_{i=-m+1}^{m-1}(1 - |i|/m)\boldsymbol{\Gamma}_{\theta,i}$ where $\boldsymbol{\Gamma}_{\theta,i} = \text{cov}_\theta(\mathbf{D}_1 l_{\mathbf{X}_1}(\mathbf{M}(\theta)), \mathbf{D}_1 l_{\mathbf{X}_{1+i}}(\mathbf{M}(\theta)))$. In particular, $\boldsymbol{\Gamma}_{\theta,0} = \text{var}_\theta(\mathbf{D}_1 l_{\mathbf{X}}(\mathbf{M}(\theta))|\mathbf{y})$ remains unchanged for any Markov chain sampling scheme. Under mild regularity conditions, the central limit theorem implies that $\sqrt{m}\mathbf{D}_1 Q^{(m)}(\mathbf{M}(\theta)|\theta)$ is asymptotically $N(\mathbf{0}, \boldsymbol{\Gamma}_\theta)$ as $m \to \infty$, where $\boldsymbol{\Gamma}_\theta = \sum_{i=-\infty}^{\infty}\boldsymbol{\Gamma}_{\theta,i}$. Hence $\sqrt{m}(\mathcal{M}_m(\theta) - \mathbf{M}(\theta))$ is asymptotically $N(\mathbf{0}, \mathbf{V}_\theta^{-1}\boldsymbol{\Gamma}_\theta\mathbf{V}_\theta^{-1})$.

Remarks.

1. The preceding heuristic argument can be made rigorous under suitable regularity conditions. Consistency in the sense that $\mathcal{M}_m(\theta) \to \mathbf{M}(\theta)$ a.s. as $m \to \infty$ can be proved using conditions similar to those of Wald (1949) (see also Pollard 1984, chap. 2). The law of large numbers holds if the Gibbs sampler is ergodic. Sufficient conditions for ergodicity of Gibbs samplers were discussed by Chan (1993) and Tierney (1994). Equation (16) is valid under the classic assumption that the second derivative, $\mathbf{D}_{11}Q^{(m)}(\tau|\theta)$, is continuous and bounded by an integrable random variable when τ is close to $\mathbf{M}(\theta)$. The central limit theorem holds under very general conditions for the Gibbs samplers (see Chan and Geyer 1994 and Tierney 1994).

2. Note that $\mathbf{V}_\theta$ and $\boldsymbol{\Gamma}_{\theta,0}$ are the same no matter which Markov chain sampling scheme is used. In practice, the weaker the dependence of the Gibbs sampler, the smaller the asymptotic variance of $\mathcal{M}_m(\theta)$. (See Liu, Wong, and Kong 1995 and Yue and Chan 1992 for discussion of the correlation structure of Gibbs samplers.)

The asymptotic distribution of $\mathcal{M}_m(\theta)$ suggests that asymptotically the MCEM sequence is an inhomogeneous vector nonlinear AR(1) process,

$$\theta_{k+1}^{(m)} \approx \mathbf{M}(\theta_k^{(m)}) + \mathbf{a}_{k+1}/\sqrt{m},$$

where for given $\mathbf{a}_1, \ldots, \mathbf{a}_k, \theta_1^{(m)}, \ldots, \theta_k^{(m)} = \theta$, $\mathbf{a}_{k+1}$ is $N(\mathbf{0}, \mathbf{V}_\theta^{-1}\boldsymbol{\Gamma}_\theta\mathbf{V}_\theta^{-1})$. When $\theta_k^{(m)}$ is close to θ^*, the global maximizer of $l_{\mathbf{Y}}(\cdot)$, we have

$$\theta_{k+1}^{(m)} - \theta^* \approx \mathbf{DM}(\theta^*)(\theta_k^{(m)} - \theta^*) + \mathbf{e}_{k+1}/\sqrt{m}, \quad (17)$$

where $\{\mathbf{e}_k\}$ is iid $N(\mathbf{0}, \mathbf{V}_*^{-1}\boldsymbol{\Gamma}_*\mathbf{V}_*^{-1})$ with $\mathbf{V}_* = \mathbf{V}_{\theta^*}$ and $\boldsymbol{\Gamma}_* = \boldsymbol{\Gamma}_{\theta^*}$. If θ^* is an isolated global maximizer, then the norm of the matrix $\mathbf{DM}(\theta^*)$ is less than 1. We assume that θ^* is an isolated maximizer and that the MCEM sequence starts in the vicinity of θ^*. For large m, Equation (17) implies that $(\theta_k^{(m)})$ is an asymptotically stationary AR(1) process. Its stationary distribution is $N(\theta^*, \boldsymbol{\Sigma}_*/m)$, where

$$\boldsymbol{\Sigma}_* = \sum_{i=0}^{\infty}(\mathbf{DM}(\theta^*))^i\mathbf{V}_*^{-1}\boldsymbol{\Gamma}_*\mathbf{V}_*^{-1}(\mathbf{DM}^i(\theta^*))^i.$$

This result implies that for large K and m and starting values of the MCEM sequence in the vicinity of θ^*, $\theta_K^{(m)} = \theta_* + O_p(1/\sqrt{m})$, with $\sqrt{m}(\theta_K^{(m)} - \theta^*)$ being approximately $N(\mathbf{0}, \boldsymbol{\Sigma}_*)$. The validity of the preceding approximation depends on the validity of the AR(1) approximation in a neighborhood of θ^* and on the iterates remaining in the neighborhood. Let θ_0 be the unknown true parameter. Recall that the ML estimate θ^* depends on the sample size n and, under suitable regularity conditions, the ML estimate is asymptotically Normal and asymptotically efficient. When $n/m \to 0$, we have

$$\sqrt{n}(\theta_K^{(m)} - \theta_0) = \sqrt{n}(\theta^* - \theta_0) + \frac{\sqrt{n}}{\sqrt{m}}\sqrt{m}(\theta_K^{(m)} - \theta^*) = \sqrt{n}(\theta^* - \theta_0) + o_p(1).$$

Thus $\theta_K^{(m)}$ and θ^* are asymptotically equivalent in that they have the same asymptotic Normal distribution. It is in this sense that we can treat $\theta_K^{(m)}$ as the ML estimate. Note that the asymptotic variance of the AR(1) process and the rate of convergence to the stationary distribution

depends on $\mathbf{DM}(\theta^*)$. It can be shown (see Dempster et al. 1977) that

$$\mathbf{DM}(\theta^*) = \mathbf{I} - E_*(-\mathbf{D}^2 l_{\mathbf{X}}(\theta^*)|\mathbf{y}))^{-1}(-\mathbf{D}^2 l_{\mathbf{Y}}(\theta^*)),$$

where $E_*(\cdot)$ denotes the expectation under θ^*. The numerator of the second term on the right side of the foregoing equation is the observed Fisher information of the incomplete data, and the denominator is the expected Fisher information of the complete data, given the incomplete observation $\mathbf{Y} = \mathbf{y}$. Hence ideally we should choose the complete data $\mathbf{X}$ so that it contains a minimal amount of expected missing information and at the same time the Gibbs sampling of its conditional distribution, given the observed data, is easy to implement.

APPENDIX: JUSTIFICATION OF THE $1/m$ ASYMPTOTICS

We now justify the $1/m$ asymptotics for s claimed in Section 2.3. We first consider the limiting distribution of $\tilde{\Delta} l_{\mathbf{Y}}(\theta_{\mathbf{h},m}, \mathcal{M}_m(\theta_{\mathbf{h},m}))$, where $\theta_{\mathbf{h},m} = \theta^* + \mathbf{h}/\sqrt{m}$ with $\mathbf{h}$ fixed. In other words, we consider the distribution of the estimated change of the log-likelihood resulting from a one-step iteration of the MCEM algorithm with its initial value within an $O(1/\sqrt{m})$ distance from the ML estimate. We will show that $m\tilde{\Delta} l_{\mathbf{Y}}(\theta_{\mathbf{h},m}, \mathcal{M}_m(\theta_{\mathbf{h},m}))$ converges in distribution to a random variable, say $\Delta(\mathbf{h})$, whose variance is equal $\sigma^2(\mathbf{h})$. This result, together with the stationary AR(1) approximation in Section 4.2, implies that under suitable regularity conditions, m^2 times the right side of (5), and hence $m^2 s^2$, is approximately equal to $E(\sigma^2(\mathbf{H}))$, where $\mathbf{H}$ is $N(\mathbf{0}, \mathbf{\Sigma}_*)$.

Recall that $\tilde{\Delta} l_{\mathbf{Y}}(\tau, \delta) = -\log(\sum_{j=1}^m f_\tau(\mathbf{X}_j)/(m f_\delta(\mathbf{X}_j)))$, where $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ is the Gibbs sampler generated with δ as the true parameter. Suppose that τ and δ are close to ω. Expanding $\tilde{\Delta} l_{\mathbf{Y}}(\tau, \delta)$ about (ω, ω), we get

$$\begin{aligned}\tilde{\Delta} l_{\mathbf{Y}}(\tau, \delta) = {} & \frac{1}{m}\sum \mathbf{D}^t l_{\mathbf{X}_j}(\omega)(\tau - \delta) \\ & + \frac{1}{2}\Bigg\{(\tau - \omega)^t \frac{1}{m}\sum \mathbf{D}^2 f_\omega(\mathbf{X}_j)/f_\omega(\mathbf{X}_j)(\tau - \omega) \\ & \quad - 2(\tau - \omega)^t \frac{1}{m}\sum \mathbf{D} l_{\mathbf{X}_j}(\omega)\mathbf{D}^t l_{\mathbf{X}_j}(\omega)(\delta - \omega) \\ & \quad + (\delta - \omega)^t \frac{1}{m}\Big(\sum -\mathbf{D}^2 l_{\mathbf{X}_j}(\omega) \\ & \quad + \mathbf{D} l_{\mathbf{X}_j}(\omega)\mathbf{D}^t l_{\mathbf{X}_j}(\omega)\Big)(\delta - \omega)\Bigg\} \\ & + \text{higher-order terms.} \end{aligned} \tag{A.1}$$

Multiplying both sides of the (A.1) by m and letting $\tau = \theta_{\mathbf{h},m}$, $\delta = \mathcal{M}_m(\theta_{\mathbf{h},m})$, $\omega = \theta^*$, and $f_* = f_{\theta^*}$, (A.1) becomes

$$\begin{aligned} m\tilde{\Delta} l_{\mathbf{Y}}(\theta_{\mathbf{h},m}, \mathcal{M}_m(\theta_{\mathbf{h},m})) = {} & o_p(1) + \frac{1}{\sqrt{m}}\sum \mathbf{D}^t l_{\mathbf{X}_j}(\theta^*)\sqrt{m}(\mathcal{M}_m(\theta_{\mathbf{h},m}) - \theta_{\mathbf{h},m}) \\ & + (1/2)\Bigg\{\sqrt{m}(\theta_{\mathbf{h},m} - \theta^*)^t \frac{1}{m}\sum \mathbf{D}^2 f_*(\mathbf{X}_j)/f_*(\mathbf{X}_j)\sqrt{m}(\theta_{\mathbf{h},m} - \theta^*) \\ & \quad - 2\sqrt{m}(\theta_{\mathbf{h},m} - \theta^*)^t \frac{1}{m}\sum \mathbf{D} l_{\mathbf{X}_j}(\theta^*)\mathbf{D}^t l_{\mathbf{X}_j}(\theta^*)\sqrt{m}(\mathcal{M}_m(\theta_{\mathbf{h},m}) - \theta^*) \\ & \quad + \sqrt{m}(\mathcal{M}_m(\theta_{\mathbf{h},m}) - \theta^*)^t \frac{1}{m}\Big(\sum -\mathbf{D}^2 l_{\mathbf{X}_j}(\theta^*) + \mathbf{D} l_{\mathbf{X}_j}(\theta^*)\mathbf{D}^t l_{\mathbf{X}_j}(\theta^*)\Big)\sqrt{m}(\mathcal{M}_m(\theta_{\mathbf{h},m}) - \theta^*)\Bigg\}. \end{aligned} \tag{A.2}$$

The limiting distribution of $\tilde{\Delta} l_{\mathbf{Y}}(\theta_{\mathbf{h},m}, \mathcal{M}_m(\theta_{\mathbf{h},m}))$ is derived using the concepts of contiguity. Suppose that $\{\gamma_{\mathbf{h},m}\}$ is a sequence of parameters converging to γ and that $\sqrt{m}(\gamma_{\mathbf{h},m} - \gamma) \to \mathbf{h}$ as $m \to \infty$. Let $P_{\gamma_{\mathbf{h},m}}$ be the distribution of the Gibbs sampler $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ with $\gamma_{\mathbf{h},m}$ as the true parameter, and let $P_{\gamma,m}$ be the distribution of $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ when the true parameter is γ. $\{P_{\gamma_{\mathbf{h},m}}\}$ is said to be contiguous to $\{P_{\gamma,m}\}$ if for any random variable, $f_m = f_m(\mathbf{X}_1, \ldots, \mathbf{X}_m)$, that converges to zero in probability under $P_{\gamma,m}$, it holds that f_m converges to zero in probability under $P_{\gamma_{\mathbf{h},m}}$. Assuming contiguity, the limiting distribution of a sequence of random variables f_m under $P_{\gamma_{\mathbf{h},m}}$ can be derived from that under $P_{\gamma,m}$. For example, if the weak law of large numbers holds under $P_{\gamma,m}$, then it also holds under $P_{\gamma_{\mathbf{h},m}}$. To state a needed consequence of contiguity, we introduce some further notations. Let $l_{\gamma,m}$ be the relative log-likelihood of $P_{\gamma_{\mathbf{h},m}}$ with respect to $P_{\gamma,m}$. Then

$$\begin{aligned} l_{\gamma,m} \approx {} & \sqrt{m}(\gamma_{\mathbf{h},m} - \gamma)^t \frac{1}{\sqrt{m}}\mathbf{U}(\gamma) \\ & + \sqrt{m}(\gamma_{\mathbf{h},m} - \gamma)^t\{\mathbf{DU}(\gamma)/m\}\sqrt{m}(\gamma_{\mathbf{h},m} - \gamma)/2, \end{aligned}$$

where $\mathbf{U}(\gamma)$ is the score vector of $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ and is equal to $\mathbf{D} l_{\mathbf{X}_1}(\gamma) + \sum_{i=1}^{m-1} \mathbf{D} q_{\mathbf{X}_i,\mathbf{X}_{i+1}}(\gamma)$, with $q_{\mathbf{X}_i,\mathbf{X}_{i+1}}(\gamma)$ being the conditional log-likelihood of $\mathbf{X}_{i+1}$ given $\mathbf{X}_i$ evaluated at γ. Under suitable regularity conditions, $E_\gamma(\mathbf{D} q_{\mathbf{X}_i,\mathbf{X}_{i+1}}(\gamma)|\mathbf{X}_1, \ldots, \mathbf{X}_i) = \mathbf{0}$. Therefore, $\mathbf{U}(\gamma)$ is a sum of martingale differences and hence $\mathbf{U}(\gamma)/\sqrt{m}$ is asymptotically $N(\mathbf{0}, \mathbf{W}_\gamma)$, where $\mathbf{W}_\gamma = E_\gamma(\mathbf{D} q_{\mathbf{X}_1,\mathbf{X}_2}(\gamma)\mathbf{D}^t q_{\mathbf{X}_1,\mathbf{X}_2}(\gamma))$ (see Billingsley 1968 for the martingale central limit theorem). Here we assume that $\mathbf{W}_\gamma$ is finite. By the law of large numbers, $\mathbf{DU}(\gamma)/m \to E_\gamma(\mathbf{D}^2 q_{\mathbf{X}_1,\mathbf{X}_2}(\gamma))$ a.s. under $P_{\gamma,m}$ and, under the assumption that differentiation and expectation commute, the latter limit is equal to $-\mathbf{W}_\gamma$. Hence $l_{\gamma,m} \to N(-\sigma^2(\gamma)/2, \sigma^2(\gamma))$ in distribution, where $\sigma^2(\gamma) = \mathbf{h}^t\mathbf{W}_\gamma\mathbf{h}$. The convergence of $l_{\gamma,m} \to N(-\sigma^2/2, \sigma^2)$ for some $\sigma^2 < \infty$ is a sufficient condition for the contiguity of $\{P_{\gamma_{\mathbf{h},m}}\}$ to $\{P_{\gamma,m}\}$. With these notations, we are ready to introduce an extremely useful consequence of contiguity. Suppose that under $P_{\gamma,m}$, $(l_{\gamma,m}, f_m)'$ converges in distribution to a bivariate Normal with mean vector $(-\sigma_1^2/2, \mu_2)'$ and covariance matrix whose diagonal elements are σ_1^2 and σ_2^2 and off-diagonal element is $\sigma_{1,2}$. Then under $P_{\gamma,m}$, f_m converges in distribution to $N(\mu_2 + \sigma_{1,2}, \sigma_2^2)$; Hajek and Sidak (1967) called this result LeCam's third lemma. For Markov processes, contiguity is often satisfied under very general conditions (see Jeganathan 1988 and Roussas 1972). Henceforth, we assume that contiguity holds.

Using the concepts discussed in the previous paragraph, the following result can be obtained, from which the assertions on the limiting behavior of $m\tilde{\Delta}l_{\mathbf{Y}}(\theta_{\mathbf{h},m}, \mathcal{M}_m(\theta_{\mathbf{h},m}))$ follow readily.

Lemma A. Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ be the Gibbs sampler generated with the true parameter being $\mathcal{M}_m(\theta_{\mathbf{h},m})$. Let $\mathbf{B} = E_*(\mathbf{D}q_{\mathbf{X}_1,\mathbf{X}_2}(\theta^*)\mathbf{D}^t q_{\mathbf{X}_1,\mathbf{X}_2}(\theta^*))$ and $\mathbf{V}_* = E_*(-\mathbf{D}^2 l_{\mathbf{X}}(\theta^*)|\mathbf{y})$, where $E_*(\cdot)$ denotes the expectation under θ^*. Also, let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ be iid $N(\mathbf{0}, \mathbf{\Gamma}_*)$. All of the following summations are from 1 to m. Then we have the following results, with all convergence being in distribution, that as $m \to \infty$,

a. $\sqrt{m}(\mathcal{M}_m(\theta_{\mathbf{h},m}) - \theta^*) \to \mathbf{V}_*^{-1}(\mathbf{Z}_1 + \mathbf{Bh})$
b. $1/m \sum \mathbf{D}^2 f_*(\mathbf{X}_t)/f_*(\mathbf{X}_t) \to E_*(\mathbf{D}^2 f_*(\mathbf{X})/f_*(\mathbf{X})|\mathbf{y})$
c. $1/m \sum \mathbf{D}l_{\mathbf{X}_t}(\theta^*)\mathbf{D}^t l_{\mathbf{X}_t}(\theta^*) \to \mathbf{V}_*$
d. $1/m \sum -\mathbf{D}^2 l_{\mathbf{X}_t}(\theta^*) \to \mathbf{V}_*$
e. $1/\sqrt{m} \sum \mathbf{D}l_{\mathbf{X}_t}(\theta^*) \to \mathbf{B}\mathbf{V}_*^{-1}(\mathbf{Z}_1 + \mathbf{Bh}) + \mathbf{Z}_2$.

Proof of Lemma A

Expanding $\mathbf{D}_1 Q^{(m)}(\cdot|\theta)$ about θ^*, we have

$$\sqrt{m}(\mathcal{M}_m(\theta_{\mathbf{h},m}) - \theta^*) = (-\mathbf{D}_{11}Q^{(m)}(\theta^*|\theta_{\mathbf{h},m}))^{-1}\sqrt{m}\mathbf{D}_1 Q^{(m)}(\theta^*|\theta_{\mathbf{h},m}) + o_p(1).$$

Note that

$$-\mathbf{D}_{11}Q^{(m)}(\theta^*|\theta_{\mathbf{h},m})) = -\frac{1}{m}\sum \mathbf{D}^2 l_{\mathbf{X}_t}(\theta^*),$$

where $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ is the Gibbs sampler generated with $\theta_{\mathbf{h},m}$ as the true parameter. It follows from ergodicity and contiguity that $-\mathbf{D}_{11}Q^{(m)}(\theta^*|\theta_{\mathbf{h},m})) \to \mathbf{V}_*$. Applying LeCam's third lemma, it is not difficult to check that $\sqrt{m}\mathbf{D}_1 Q^{(m)}(\theta^*|\theta_{\mathbf{h},m}) \to N(\mathbf{Bh}, \mathbf{\Gamma}_*)$ in distribution; hence $\sqrt{m}(\mathcal{M}_m(\theta_{\mathbf{h},m}) - \theta^*) \to \mathbf{V}_*^{-1}(\mathbf{Z}_1 + \mathbf{Bh})$ in distribution. It follows from Skohorod's theorem (see Pollard 1984) that there exists a sequence of random vectors (ζ_m) such that (a) the distribution of ζ_m is that of $\sqrt{m}(\mathcal{M}_m(\theta_{\mathbf{h},m}) - \theta^*)$, and (b) $\zeta_m \to \mathbf{V}_*^{-1}(\mathbf{Z}_1 + \mathbf{Bh})$ almost surely. Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ be the Gibbs sampler generated with the true parameter being $\gamma_{\mathbf{h},m} = \zeta_m/\sqrt{m} + \theta^*$. Its unconditional distribution is equal to that of $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ generated with the true parameter being $\mathcal{M}_m(\theta_{\mathbf{h},m})$. Hence in verifying b–e, it suffices to assume that $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ is generated with the true parameter being $\gamma_{\mathbf{h},m}$. Note that $\sqrt{m}(\gamma_{\mathbf{h},m} - \theta^*) \to \mathbf{V}_*^{-1}(\mathbf{Z}_1 + \mathbf{Bh})$ a.s. Formulas b–e follow from contiguity by first conditioning on $\mathbf{Z}_1$. For example, given $\mathbf{Z}_1 = \mathbf{z}_1$, the $\mathbf{X}$'s in the expression of $1/\sqrt{m} \sum \mathbf{D}l_{\mathbf{X}_t}(\theta^*)$ are generated with the true parameter being $\gamma_{\mathbf{h},m} = \theta^* + \mathbf{V}_*^{-1}(\mathbf{z}_1 + \mathbf{Bh})/\sqrt{m} + o(1/\sqrt{m})$. Then applying LeCam's third lemma, we obtain that $1/\sqrt{m} \sum \mathbf{D}l_{\mathbf{X}_t}(\theta^*) \to \mathbf{B}\mathbf{V}_*^{-1}(\mathbf{z}_1 + \mathbf{Bh}) + \mathbf{Z}_2$ in distribution; hence the result as stated in e. This completes the proof of Lemma A.

Proof of Lemma 1

We need to verify the two conditions in the definition of asymptotic stability. By definition, for any neighborhood of θ^*, there exists a subneighborhood V, of θ^*, over which $l_{\mathbf{Y}}(\cdot)$ attains its maximum at θ^* and that does not include another stationary point of $l_{\mathbf{Y}}(\cdot)$. Without loss of generality, assume that V is compact. For some $\varepsilon > 0$, the level set $V_0 = \{\theta : l_{\mathbf{Y}}(\theta) \geq l_{\mathbf{Y}}(\theta^*) - \varepsilon\}$ has a connected component that contains θ^* and lies inside the interior of V. Call this connected component V_1. Then for all $\theta \in V_1$, $l_{\mathbf{Y}}(\mathbf{M}(\theta)) > l_{\mathbf{Y}}(\theta)$ unless $\theta = \theta^*$; hence $\mathbf{M}(V_1) \subseteq V_0$. Because $\mathbf{M}(V_1)$ is connected and $\mathbf{M}(\theta^*) = \theta^* \in V_1$, it follows that $\mathbf{M}(V_1) \subseteq V_1$. This shows that Condition 1 is satisfied. Let $\theta \in V_1$. Because $\{l_{\mathbf{Y}}(\mathbf{M}^k(\theta))\}$ is nondecreasing and bounded above by $l_{\mathbf{Y}}(\theta^*)$, it converges to some finite number, say c. By the compactness of V_1, any subsequence of $\{\mathbf{M}^k(\theta)\}$ has a convergent subsequence. To show that $\mathbf{M}^k(\theta) \to \theta^*$ and $\{l_{\mathbf{Y}}(\mathbf{M}^k(\theta))\}$ converges to $l_{\mathbf{Y}}(\theta^*)$, it suffices to show that any convergent subsequence of $\{\mathbf{M}^k(\theta)\}$ converges to θ^*. For simplicity, assume that $\{\mathbf{M}^k(\theta)\}$ is convergent and converges to $\theta^\dagger$. We want to show that $\theta^\dagger = \theta^*$. Suppose not; then $l_{\mathbf{Y}}(\mathbf{M}^{k+1}(\theta)) \to l_{\mathbf{Y}}(\mathbf{M}(\theta^\dagger)) > c$, contradicting the fact that $l_{\mathbf{Y}}(\theta^\dagger) = c$. This implies that $\mathbf{M}^k(\theta) \to \theta^*$ as $k \to \infty$. Thus Condition 2 is satisfied. This completes the proof of Lemma 1.

Proof of Theorem 1

Let V_1 be as constructed in the proof of Lemma 1. Note that V_1 is compact. $\mathbf{M}(V_1)$ lies inside the interior of V_1. Let $V_2 = \{\theta \in V_1, |\theta - \theta^*| \geq \varepsilon\}$, for given $\varepsilon > 0$. By shrinking ε if necessary, it can be assumed that V_2 is nonempty. Now, for all $\theta \in V_2$, $\mathbf{M}(\theta) \neq \theta$, and hence there exists δ and δ_1, all larger than zero, such that for all $\theta \in V_2$ and for all θ' with $|\theta' - \mathbf{M}(\theta)| < \delta$, $l_{\mathbf{Y}}(\theta') - l_{\mathbf{Y}}(\theta) > \delta_1$. By continuity and compactness, there exists $\delta_2 > 0$ such that for all $\theta \in V_1$ and for all θ' with $|\theta' - \mathbf{M}(\theta)| < \delta_2$, $\theta' \in V_1$. Without loss of generality, assume that $\delta_2 < \delta$. Let $R = \max_{\theta', \theta \in V_1}\{l_{\mathbf{Y}}(\theta) - l_{\mathbf{Y}}(\theta')\} > 0$. Let $k_0 = [R/\delta_1] + 1$, where $[\cdot]$ denotes the integral part of a real number. Given $\theta_k^{(m)} = \theta \in V_1$, the event that $\theta_{k+1}^{(m)} \in V_1$ contains the event that $\{|\mathcal{M}(\bar{\mathbf{Z}}_m(\theta_k^{(m)})) - \mathbf{M}(\theta)| < \delta_2\}$ which has a probability $p = p(\delta_2) > 0$, uniformly for $\theta \in V_1$. This is because $\mathbf{M}(\theta) = \mathcal{M}(E_\theta(\mathbf{Z}))$ and $\bar{\mathbf{Z}}_m \to E_\theta(\mathbf{Z})$ in probability uniformly for θ lying over a compact set. Therefore, the probability of the event $F = \{\theta_k^{(m)} \in V_1$, for all $k = 0, 1, \ldots, k_0\}$ is not less than p^{k_0}. On the event F, $|\theta_k^{(m)} - \theta^*| < \varepsilon$ for some $k \leq k_0$; otherwise for all $k \leq k_0$, $\theta_k^{(m)} \in V_2$ and hence $l_{\mathbf{Y}}(\theta_{k_0}^{(m)}) - l_{\mathbf{Y}}(\theta_0) > \delta_1 k_0 > R$, leading to a contradiction. Because $p \to 1$ as $m \to \infty$, this completes the proof.

[Received September 1992. Revised April 1994.]

REFERENCES

Anderson, B. D. O., and Moore, J. (1979), *Optimal Filtering,* Englewood Cliffs, NJ: Prentice-Hall.

Besag, J., and Green, P. J. (1993), "Spatial Statistics and Bayesian Computation" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 55, 25–38.

Billingsley, P. (1968), *Convergence of Probability Measures,* New York: John Wiley.

Chan, K. S. (1993), "Asymptotic Behavior of the Gibbs Sampler," *Journal of the American Statistical Association,* 88, 320–326.

Chan, K. S., and Geyer, C. J. (1994), Comment on "Markov Chains for Exploring Posterior Distributions," by L. Tierney, *The Annals of Statistics,* 22, 4.

Cox, D. R. (1981), "Statistical Analysis of Time Series: Some Recent Developments," *Scandinavian Journal of Statistics,* 8, 93–115.

Dempster, A. P., Laird, N. M., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 39, 1–38.

Devroye, L. (1984), "A Simple Algorithm for Generating Random Variates With a Log-Concave Density," *Computing,* 33, 247–57.

Geyer, C. J. (1991), "Monte Carlo Maximum Likelihood for Dependent Data," in *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface,* pp. 156–163.

Guo, S. W., and Thompson, E. A. (1991), "Monte Carlo Estimation of Variance Component Models for Large Complex Pedigrees," *IMA Journal of Mathematics Applied in Medicine and Biology,* 8, 171–189.

Hajek, J., and Sidak, Z. (1967), *Theory of Rank Tests,* New York: Academic Press.

Harvey, A. C., and Koopman, S. J. (1992), "Diagnostic Checking of Unobserved-Components Time Series Models," *Journal of Business & Economic Statistics,* 10, 377–389.

Jeganathan, P. (1988), "Some Aspects of Asymptotic Theory With Applications to Time Series Models," technical report, University of Michigan, Dept. of Statistics.

Lasalle, J. P. (1976), *The Stability of Dynamical System,* Philadelphia: Society of Industrial and Applied Mathematics.

Liu, J., Wong, W. H., and Kong, A. (1995), "Correlation Structure and Convergence Rate of the Gibbs Sampler for Various Scans," *Journal of the Royal Statistical Society,* Ser. B, 57, 157–169.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society,* Ser. B, 44, 190–200.

Pollard, D. (1984), *Convergence of Stochastic Processes,* New York: Springer-Verlag.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. (1986), *Numerical Recipes,* Cambridge, U.K.: Cambridge University Press.

Roussas, G. G. (1972), *Contiguity of Probability Measures: Some Applications in Statistics,* Cambridge, U.K.: Cambridge University Press.

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods" (with discussion), *Journal of the Royal Statistical Society,* Ser. B, 55, 3–24.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics,* Vol. 22.

Wald, A. (1949), "Note on the Consistency of Maximum Likelihood Estimate," *The Annals of Statistics,* 20, 595–601.

Wei, C. G., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association,* 85, 699–704.

Wu, J. C. F. (1983), "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics,* 11, 95–103.

Yue, H., and Chan, K. S. (1992), "Complete Monotonicity of the Gibbs Samplers and Asymptotic Efficiency of Sample Means," Technical Report 212, University of Iowa, Dept. of Statistics and Actuarial Science.

Zeger, S. L. (1988), "A Regression Model for Time Series of Counts," *Biometrika,* 75, 621–629.