CONDITIONS FOR CONVERGENCE OF MONTE-CARLO *EM* SEQUENCES WITH AN APPLICATION TO PRODUCT DIFFUSION MODELING

Robert P. Sherman

California Institute of Technology

Division of Humanities and Social Sciences 228–77

Pasadena, CA 91125, USA

sherman@amdg.caltech.edu

Yu-Yun K. Ho

Telcordia Technologies

Information Analysis & Services Research Department

445 South Street, Morristown, NJ 07960, USA

yyh@research.telcordia.com

Siddhartha R. Dalal

Telcordia Technologies

Information Analysis & Services Research Department

445 South Street, Morristown, NJ 07960, USA

sid@research.telcordia.com

Abstract

Intractable maximum likelihood problems can sometimes be finessed with a Monte-Carlo implementation of the *EM* algorithm. However, there appears to be little theory governing when Monte-Carlo *EM* (*MCEM*) sequences converge. Consequently, in some applications, convergence is assumed rather than proved. Motivated by this problem in the context of modeling market penetration of new products and services over time, we develop (i) high-level conditions for rates of almost-sure convergence and convergence in distribution of any *MCEM* sequence and (ii) primitive conditions for almost-sure monotonicity and almost-sure convergence of an *MCEM* sequence when Monte-Carlo integration is carried out using independent Gibbs runs. We verify the main primitive conditions for the Bass product diffusion model and apply the methodology to data on wireless telecommunication services.

KEY WORDS: Monte-Carlo *EM*, Convergence Conditions, Gibbs Sampling, Temporal Diffusion, New Products and Services, Bass Model.

1. Introduction

The *EM* algorithm (Dempster, Laird, and Rubin, 1977) is a popular sequential procedure for computing maximum likelihood estimates in incomplete data problems. The *E* step of the algorithm requires computation of a conditional expectation of the augmented, or complete, data likelihood function. When this computation is infeasible, recourse can be made to direct Monte-Carlo integration or, more commonly, to a blending of Monte-Carlo integration with Markov chain sampling techniques like the Gibbs sampler, the Metropolis algorithm, or the Hastings algorithm. In any case, the resulting scheme is called a Monte-Carlo *EM*, or *MCEM* algorithm. Wei and Tanner (1990) were the first to propose such a scheme. Guo and Thompson (1991), Shephard (1993), Chan and Ledolter (1995), and Meng and Schilling (1996) present substantial applications. See also related work by Hajivassiliou and Ruud (1994).

Biscarat (1994) states conditions for almost-sure convergence of a general class of stochastic algorithms and verifies these conditions for an *MCEM* algorithm applied to a mixture problem involving direct Monte-Carlo integration of the *E* step. While Chan and Ledolter (1995) focus on applying the $MCEM$ algorithm to a certain type of time series count data, in the course of doing so they implicitly develop some convergence theory for a general $MCEM$ sequence where one long Gibbs run is used in the E step of each iteration. They establish convergence in probability of such an $MCEM$ sequence to a local maximizer of the incomplete data likelihood provided the starting values for the sequence are in a small neighborhood of the local maximizer. They also give heuristic arguments for asymptotic normality and efficiency of the $MCEM$ sequence.

In this paper, we seek to clarify and complement the results of Biscarat (1994) and Chan and Ledolter (1995) by deriving (i) high-level conditions for almost-sure convergence and convergence in distribution of any $MCEM$ sequence and (ii) primitive conditions for almost-sure monotonicity and almost-sure convergence of an $MCEM$ sequence when the E step of each iteration is implemented using Monte-Carlo integration with independent Gibbs runs.

The high-level conditions in (i) are independent of the Monte-Carlo method used to implement the E step. For example, they cover direct Monte-Carlo integration, or Monte-Carlo integration using either the Metropolis algorithm, the Hastings algorithm, the Gibbs sampler with either one long run or many independent runs, or a mixture of these techniques. In fact, nothing intrinsic to maximum likelihood estimation is implicit in the high-level conditions. The conditions cover Monte-Carlo approximations of convergent sequences defined by optimization of any criterion function at each iteration. Our aim in (ii) is not to find the most general conditions, but rather a set of conditions that are understandable to practitioners and appear to have a reasonably wide range of applicability.

Throughout, we assume that an EM sequence, started at a parameter value θ^0, converges to a value $\hat{\theta}$. We then develop conditions under which an $MCEM$ sequence, started at θ^0, converges in an appropriate sense to $\hat{\theta}$. By taking this approach we separate convergence issues for EM from those for *MCEM*. (See Dempster et al. (1977), Wu (1983), and Biscarat (1994) for more on the former.) For notational simplicity, we develop convergence conditions in the context of maximum likelihood estimation rather than maximum posterior likelihood estimation. In the final section, we indicate how to translate the frequentist conditions to a Bayesian setting.

We were introduced to the *MCEM* methodology through work on modeling the temporal diffusion of telecommunication services into residential and business markets. This work, together with an apparent lack of clear guidance in the literature on requirements for convergence of *MCEM* sequences, provided the impetus for this paper. We present some details of the Bass model (Bass, 1969), a widely used temporal diffusion model for business applications, and verify the main primitive conditions for convergence for this model.

The paper is organized as follows: In Section 2, we discuss the *EM* and *MCEM* algorithms in more detail, introducing necessary notation. The convergence theorems are presented in Section 3. Section 4 introduces the Bass model. We verify the main primitive conditions implying almost-sure convergence of the *MCEM* sequence for this model and present simulation results illustrating the performance of the *MCEM* procedure in this context. In addition, we apply the methodology to data on wireless telecommunications services in the United States. In Section 5, we summarize and discuss related issues.

2. The *EM* and *MCEM* Algorithms

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ denote the observed data vector where n is the sample size. Let $\mathcal{Y}$ denote the support of $\boldsymbol{Y}$ and let $\boldsymbol{y}$ denote an element of $\mathcal{Y}$. Write $f(\boldsymbol{y} \mid \theta)$ for the observed data likelihood where θ is an element of the parameter space, Θ. Let θ_0 denote the value of the parameter that generated the observed data. Our objective is to estimate θ_0 by the method of maximum likelihood.

Sometimes direct maximization of $f(\boldsymbol{y} \mid \theta)$ is intractable due to the incomplete nature of the observed data set. Still, it may be possible to obtain a maximum likelihood estimator indirectly through the sequential EM procedure. Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ denote the missing data vector, $\mathcal{X}$ the support of $\boldsymbol{X}$, and $\boldsymbol{x}$ an element of $\mathcal{X}$. Write $f(\boldsymbol{x}, \boldsymbol{y} \mid \theta)$ for the complete data likelihood. Let ϕ denote an arbitrary element of Θ. Write $f(\cdot \mid \boldsymbol{y}, \phi)$ for the conditional density of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ and ϕ, and $l(\boldsymbol{x}, \boldsymbol{y} \mid \theta)$ for $\log f(\boldsymbol{x}, \boldsymbol{y} \mid \theta)$. For each θ and ϕ in Θ, let

$$Q(\theta \mid \phi) = \int_{\mathcal{X}} l(\boldsymbol{x}, \boldsymbol{y} \mid \theta)\, f(\boldsymbol{x} \mid \boldsymbol{y}, \phi)\, d\boldsymbol{x}$$

and define the set

$$\theta(\phi) = \operatorname*{argmax}_{\Theta} Q(\theta \mid \phi)\,.$$

Throughout we shall assume that $\theta(\phi)$ contains at least one element in Θ for each ϕ in Θ. The EM algorithm generates a sequence $\theta^1, \theta^2, \ldots$ where $\theta^i \in \theta(\theta^{i-1})$, $i = 1, 2, \ldots$ with θ^0 an arbitrary starting value. In this paper, we assume that $\theta^i \to \hat{\theta}$ as $i \to \infty$ for some $\hat{\theta}$ in Θ. For example, Wu (1983) shows that if $f(\boldsymbol{y} \mid \theta)$ is unimodal, then under mild regularity conditions, θ^i converges to this mode as i tends to infinity.

It may not be possible to evaluate $Q(\theta \mid \phi)$ either directly or through numerical integration. However, it may be possible to estimate $Q(\theta \mid \phi)$ through Monte-Carlo integration. If it is easy to generate samples from $f(\cdot \mid \boldsymbol{y}, \phi)$, then Monte-Carlo integration is straightforward. If not, a Markov Chain sampling technique like the Gibbs sampler may be used to facilitate Monte-Carlo integration. To fix ideas, we develop notation for a Gibbs implementation.

For each $k \geq 1$, write $f_k(\cdot \mid \boldsymbol{y}, \phi)$ for the conditional density of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$ and ϕ after k Gibbs iterations. Let $\boldsymbol{X}_i^k(\phi)$, $i = 1, \ldots, m$, denote a sample of independent observations from

$f_k(\cdot \mid \boldsymbol{y}, \phi)$ and write $P^m_{k,\phi}$ for the empirical measure that puts mass $\frac{1}{m}$ on each $\boldsymbol{X}^k_i(\phi)$. Notice that $\boldsymbol{X}^k_i(\phi)$ depends on θ_0 through the observed data $\boldsymbol{y}$. For each θ and ϕ in Θ, let

$$Q^m_k(\theta \mid \phi) = \int_{\mathcal{X}} l(\boldsymbol{x}, \boldsymbol{y} \mid \theta)\, P^m_{k,\phi}(d\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} l(\boldsymbol{X}^k_i(\phi), \boldsymbol{y} \mid \theta)$$

and define the set

$$\hat{\theta}(\phi) = \operatorname*{argmax}_{\Theta} Q^m_k(\theta \mid \phi)\,.$$

As with $\theta(\phi)$, we shall assume that $\hat{\theta}(\phi)$ contains at least one element for each ϕ in Θ. The *MCEM* algorithm generates a sequence of iterates $\hat{\theta}^1, \hat{\theta}^2 \ldots$ where $\hat{\theta}^i \in \hat{\theta}(\hat{\theta}^{i-1})$, $i = 1, 2, \ldots$ with θ^0 an arbitrary starting value.

To recap, we assume that an *EM* sequence, started at θ^0, converges to some value $\hat{\theta}$ in Θ. We make no claims about the nature of $\hat{\theta}$. It may be the global maximizer of $f(\boldsymbol{y} \mid \theta)$, a local maximizer, a saddle point, or some other value. For our purposes, $\hat{\theta}$ is simply the limit point of an *EM* sequence started at θ^0. In the next section, we develop conditions under which an *MCEM* sequence, started at θ^0, converges almost surely and in distribution to $\hat{\theta}$. The issue of determining the nature of $\hat{\theta}$ is taken up briefly in Section 5.

3. Convergence Results

In this section, we present some convergence results for the *MCEM* procedure. Theorem $1'$ and Theorem $2'$ provide high-level conditions for rates of almost-sure convergence and convergence in distribution of any *MCEM* sequence. These theorems state how fast n, the sample size of the observed data, i, the number of *MCEM* iterations, and m, the Monte-Carlo sample size for each iteration, must simultaneously tend to infinity to ensure convergence. Theorem 1 and Theorem 2 are analogous, though weaker, results. However, their conditions are also weaker. Theorem 1 gives conditions for almost-sure convergence of any $MCEM$ sequence, conditional on the observed sample. Theorem 2 gives conditions for convergence in distribution. These theorems state how fast i and m must simultaneously tend to infinity to ensure convergence. Theorem 3 states primitive conditions for almost-sure pointwise convergence of the argmax functional defining the $MCEM$ sequence when independent Gibbs runs are used to facilitate Monte-Carlo integration. Almost-sure pointwise convergence and continuity of the incomplete data likelihood imply a monotonicity property analogous to that for the EM procedure. Theorem 4 and Corollary 5 provide primitive conditions implying the high-level conditions of Theorem 1 and Theorem 2 when independent Gibbs runs are used in the integration step. These results state how fast i, m, and k, the number of Gibbs steps in each Monte-Carlo sample, must simultaneously tend to infinity to ensure convergence.

We shall establish notions of convergence of $MCEM$ sequences as the indices n, i, m, and k tend to infinity. The following compact notation will be convenient to use in this regard. Define the limit operator

$$L(\gamma_1, \ldots, \gamma_\kappa) = \lim_{\gamma_1,\ldots,\gamma_\kappa \to \infty}$$

where each γ_i has domain the positive integers. For example,

$$L(m,k)|\hat{\theta}(\phi) - \theta(\phi)| = \lim_{m,k\to\infty} |\hat{\theta}(\phi) - \theta(\phi)| \,.$$

In the last expression, m and k tend to infinity simultaneously and unconditionally. Sometimes it will be useful to impose conditions on how indices tend to infinity. For example, the expression "$L(n,i,m,k)\sqrt{n}|\hat{\theta}^i - \hat{\theta}| = 0$ $as.$ provided $n = O(i^\alpha)$ and $i = O(m^\beta)$" means that $\sqrt{n}|\hat{\theta}^i - \hat{\theta}|$ converges almost surely to zero as n, i, k, and m tend to infinity simultaneously but subject to the conditions that n be bounded by a multiple of i^{α} and i be bounded by a multiple of m^{β}. As a final example, consider

$$L(n)L(i,m,k)|\sqrt{n}(\hat{\theta}^i - \theta^i)| = \lim_{n\to\infty} \left[L(i,m,k))|\sqrt{n}(\hat{\theta}^i - \hat{\theta})| \right] .$$

In this last expression, first i, k, and m tend to infinity simultaneously, and then n tends to infinity.

Recall the notation from Section 2. To guarantee rates of almost-sure convergence we require the following uniform convergence and smoothness conditions on the argmax functionals defining the $MCEM$ and EM sequences. Write $S(n,m,k)$ for $\sup_{\phi\in\Theta} |\hat{\theta}(\phi) - \theta(\phi)|$.

A1′. For some $\delta > 0$, $L(n,m,k)\mathbb{E}m^{\delta}S(n,m,k) < \infty$.

A2′. For some $M \geq 0$, $|\theta(\phi) - \theta(\phi')| \leq M|\phi - \phi'|$ for each ϕ, ϕ' in Θ.

A1′ and A2′ are the needle and thread needed to stitch together the EM and $MCEM$ sequences. The uniform convergence requirement in A1′ can be understood by comparing a consistency argument for a maximization estimator with one for an iterative estimator defined by a maximization at each iteration. The former typically requires the sample objective function to converge uniformly in some sense to a corresponding population objective function. The uniform convergence condition in A1′ is an analogue for an iterative estimator. It requires uniform convergence one level deeper, namely, on the level of argmax functionals rather than objective functions. This stronger condition is needed to string together a sequence of maximization estimators. The Lipschitz condition A2′ is used to develop a recursion formula used in the consistency proof.

Theorem 1′: *Suppose A1′ and A2′ hold, and* $L(n,i)a_n|\theta^i - \hat{\theta}| = 0$ *where* $\{a_n\}$ *is a sequence of nonnegative real numbers. Then* $L(n,i,m,k)a_n|\hat{\theta}^i - \hat{\theta}| = 0$ *as. provided*

(a) *If* $0 \leq M < 1$, *then for* $A > 0$ *and* $B > 0$ *satisfying* $B\delta - A > 1$, $a_n = O(i^A)$ *and* $i = O(m^{1/B})$.

(b) *If $M = 1$, then for $A > 0$ and $B > 0$ satisfying $B\delta - A > 2$, $a_n = O(i^A)$ and $i = O(m^{1/B})$.*

(c) *If $M > 1$, then for $A > 2$, $a_n = O(M^{i\delta}/i^A)$ and $i = O(\log_M m)$.*

Proof. We begin by proving the result under condition (a). Apply condition A2$'$ to get, for each $i \geq 1$,

$$\begin{aligned} |\hat{\theta}^i - \theta^i| &= |\hat{\theta}(\hat{\theta}^{i-1}) - \theta(\theta^{i-1})| \\ &\leq |\hat{\theta}(\hat{\theta}^{i-1}) - \theta(\hat{\theta}^{i-1})| + |\theta(\hat{\theta}^{i-1}) - \theta(\theta^{i-1})| \\ &\leq \sup_{\phi\in\Theta} |\hat{\theta}(\phi) - \theta(\phi)| + M|\hat{\theta}^{i-1} - \theta^{i-1}| \,. \end{aligned}$$

Write $S_j(n, m, k)$ for the random variable $\sup_{\phi\in\Theta} |\hat{\theta}(\phi) - \theta(\phi)|$ constructed from the Monte-Carlo samples generated for the the jth iteration, $1 \leq j \leq i$. Apply the last inequality recursively to see that

$$a_n|\hat{\theta}^i - \theta^i| \leq \sum_{j=1}^{i} a_n S_j(n, m, k) M^{i-j} \,.$$

Condition A1$'$ implies that for n, m, and k large enough, $I\!\!E m^\delta S(n, m, k) \leq \kappa < \infty$. Deduce from this and the identical distributions of the $S_j(n, m, k)$'s that

$$I\!\!E a_n|\hat{\theta}^i - \theta^i| \leq a_n \frac{\kappa}{m^\delta} \sum_{j=1}^{i} M^{i-j} \leq \frac{a_n}{m^\delta} \frac{\kappa}{1 - M} \,.$$

By condition (a), a_n/m^δ is bounded by a multiple of $i^{A/B\delta}$ and $B\delta - A > 1$. Thus, there exists a constant $C > 0$ such that

$$\sum_{i=1}^{\infty} \mathbb{E} a_n |\hat{\theta}^i - \theta^i| \quad \leq \quad \sum_{i=1}^{\infty} \frac{C}{i^{B\delta - A}} < \infty \, .$$

Convergence in mean sufficiently fast implies convergence almost surely (e.g. Serfling 1980, p.11). Deduce that $L(n, i, m, k) a_n |\hat{\theta}^i - \theta^i| = 0 \; as.$ The result follows from this and the condition $L(n, i) a_n |\theta^i - \hat{\theta}| = 0$.

The proof under condition (b) or (c) is similar, after noting that $\sum_{j=1}^{i} M^{i-j} = i$ when $M = 1$ and $\sum_{j=1}^{i} M^{i-j} \leq i M^i$ when $M > 1$. *QED.*

Remark 1. Note that the result is independent of the type of Monte-Carlo implementation of the E step. For example, it can cover Monte-Carlo integration using independent Gibbs runs or one long Gibbs run. In fact, nothing intrinsic to maximum likelihood estimation is contained in the conditions or the proof of Theorem $1'$. The result can be used to establish rates of almost-sure convergence of a Monte-Carlo approximation to a convergent sequence defined by optimization of any criterion function at each iteration.

Remark 2. To see how one might apply Theorem $1'$, consider the interesting special case where (i) $0 < \delta \leq 1/2$ (see Theorem 4 below) (ii) $a_n = \sqrt{n}$, the rate at which regular maximum likelihood estimators converge, and (iii) $\hat{\theta}$ is a fixed point of $\theta(\phi)$ and the Lipschitz constant M in A$2'$ satisfies $0 \leq M < 1$ on Θ, a ball centered at $\hat{\theta}$. Under (iii), the argmax functional $\theta(\phi)$ that defines the EM sequence is a contraction mapping on Θ. It follows easily from the contraction property that $\sqrt{n}|\theta^i - \hat{\theta}| = O(\sqrt{n} M^i)$ and so $L(n, i)\sqrt{n}|\theta^i - \hat{\theta}| = 0$ when n and i satisfy the constraints of condition (a) in Theorem $1'$. These constraints require that $n = O(i^{2A}) = O(m^{2A/B})$. Since $0 < \delta \leq 1/2$, it follows that $2A < 2B\delta - 2 < B - 2$ and so n must grow at a slower rate than m to satisfy condition (a). If we choose B large and $2A \approx B$, then convergence will follow if n and m grow at nearly the same rate, and i grows much more slowly than n and m.

Theorem $1'$ provides the means of establishing the asymptotic distributional properties of general $MCEM$ sequences. Recall that θ_0 denotes the parameter value that generated the observed data. To fix ideas, suppose it is true that

$$L(n)\sqrt{n}(\hat{\theta} - \theta_0) \Longrightarrow N(0, V)$$

where the symbol $\Longrightarrow$ denotes convergence in distribution. We want to show that under appropriate conditions on how n, i, m, and k tend to infinity,

$$L(n, i, m, k)\sqrt{n}(\hat{\theta}^i - \theta_0) \Longrightarrow N(0, V).$$

We can then justify doing inference with $\hat{\theta}^i$ in place of $\hat{\theta}$. If, in addition, $\hat{\theta}$ maximizes the observed data likelihood, then the $MCEM$ sequence will inherit any efficiency properties the maximum likelihood sequence may possess. Since

$$\sqrt{n}(\hat{\theta}^i - \theta_0) = \sqrt{n}(\hat{\theta}^i - \hat{\theta}) + \sqrt{n}(\hat{\theta} - \theta_0)$$

it is enough to show that $L(n, i, k, n)\sqrt{n}|\hat{\theta}^i - \hat{\theta}| = 0$ *as*. Theorem $1'$ provides conditions under which this happens. We get the following result.

Theorem $2'$: *Suppose $L(n)\sqrt{n}(\hat{\theta} - \theta_0) \Longrightarrow N(0, V)$ and $L(n, i)a_n|\theta^i - \hat{\theta}| = 0$. If A1$'$ and A2$'$ hold for condition (a), (b), or (c) of Theorem 1$'$, then $L(n, i, m, k)\sqrt{n}(\hat{\theta}^i - \theta_0) \Longrightarrow N(0, V)$.*

The conditions of Theorem $1'$ are stronger than needed to prove that $\hat{\theta}^i$ converges almost surely to $\hat{\theta}$ given the observed data $\boldsymbol{Y} = \boldsymbol{y}$. Theorem 1 below gives weaker conditions sufficient to prove that $L(i,m,k)|\hat{\theta}^i - \hat{\theta}| = 0$ *as.* The proof of Theorem 1 is very similar to the proof of Theorem $1'$ and so is omitted. Let $\mathbb{E}_{\boldsymbol{y}}$ denote an expectation taken conditional on $\boldsymbol{Y} = \boldsymbol{y}$.

A1. For some $\delta > 0$, $L(m,k)\mathbb{E}_{\boldsymbol{y}}\left[m^{\delta}S(n,m,k)\right] < \infty$.

Theorem 1: *Suppose A1 and A2′ hold, and* $L(i)|\theta^i - \hat{\theta}| = 0$. *Then* $L(i,m,k)|\hat{\theta}^i - \hat{\theta}| = 0$ *as. provided*

(a) *If* $0 \leq M < 1$, *then for* $B > 1/\delta$, $i = O(m^{1/B})$.

(b) *If* $M = 1$, *then for* $B > 2/\delta$, $i = O(m^{1/B})$.

(c) *If* $M > 1$, *then for* $0 < B < \delta$, $i = O(\log_M m^B)$.

Refer to the discussion preceding the statement of Theorem $2'$. A simple dominated convergence argument can be used to establish the following result.

Theorem 2: *Suppose* $\sqrt{n}(\hat{\theta} - \theta_0) \Longrightarrow N(0,V)$ *and* $L(i)|\theta^i - \hat{\theta}| = 0$. *If A1 and A2′ hold for conditions (a), (b), or (c) of Theorem 1, then* $L(n)L(i,m,k)\sqrt{n}(\hat{\theta}^i - \theta_0) \Longrightarrow N(0,V)$.

Theorem 1 and Theorem 2 provide conditions for almost-sure convergence and convergence in distribution of any *MCEM* sequence. In order to apply these theorems, we must find reasonable primitive conditions under which assumptions A1 and A2′ hold. We now develop such conditions in the context of Monte-Carlo integration carried out with independent Gibbs runs. The following definitions, taken from Pakes and Pollard (1989), will be useful in developing these primitive conditions.

Let $\mathcal{F}$ denote a class of functions on a set $\mathcal{X}$. An envelope for $\mathcal{F}$ is any function F such that $|f| \leq F$ for each f in $\mathcal{F}$. Call $\mathcal{F}$ *Euclidean* for an envelope F if there exist positive constants A and V with the following property: if $0 < \epsilon \leq 1$ and μ is a measure for which $\int F d\mu < \infty$, then there exist functions $f_1, \ldots, f_N$ in $\mathcal{F}$ such that

(i) $N \leq A\epsilon^{-V}$

(ii) $\mathcal{F}$ is covered by a union of closed balls with radius $\epsilon \int F d\mu$ and centers $f_1, \ldots, f_N$. That is, for each f in $\mathcal{F}$, there is an f_i such that $\int |f - f_i| d\mu \leq \epsilon \int F d\mu$.

(iii) The constants A and V do not depend on μ.

Informally, a class of functions $\mathcal{F}$ is Euclidean if only a small number of functions in the class is needed to approximate any function in the class to any degree of accuracy. Many classes of functions satisfy this property, including any class indexed by a bounded, finite-dimensional parameter with uniformly bounded derivatives. The class of functions encountered in the diffusion modeling application in Section 4 is of this type. However, the Euclidean property is satisfied by many other classes including classes of functions that are neither bounded nor smooth, and even for these classes, the property is often easy to verify. See, for example, Pakes and Pollard (1989) or Nolan and Pollard (1987) for examples and many simple ways to check this condition.

Recall, once again, the notation in Section 2. Assume that an EM sequence $\theta^0, \theta^1, \ldots \theta^i$ where $\theta^i = \theta(\theta^{i-1})$ converges to a value $\hat{\theta}$ in Θ. Let ϕ denote a current estimate of $\hat{\theta}$. The EM algorithm satisfies the monotonicity property

$$f(\boldsymbol{y} \mid \theta(\phi)) \geq f(\boldsymbol{y} \mid \phi)\,.$$

That is, at each step of the EM algorithm, the incomplete data likelihood is nondecreasing. This property is lost for the $MCEM$ procedure. In other words, $f(\boldsymbol{y} \mid \hat{\theta}(\phi))$ need not be greater than or equal to $f(\boldsymbol{y} \mid \phi)$. However, if $f(\boldsymbol{y} \mid \theta)$ is continuous in θ and $\hat{\theta}(\phi)$ converges almost surely to $\theta(\phi)$ as m and k tend to infinity, then with probability one, as m and k tend to infinity, monotonicity in the incomplete data likelihood is preserved. Our next result states conditions under which $\hat{\theta}(\phi)$ converges almost surely to $\theta(\phi)$ for each ϕ in Θ. This result will be useful, not only for establishing the monotonicity property mentioned above, but also for developing primitive conditions sufficient to imply conditions A1 and A2$'$.

Write $\mathcal{F}$ for the class of functions $\{l(\boldsymbol{x}, \boldsymbol{y} \mid \theta) : \boldsymbol{x} \in \mathcal{X}, \theta \in \Theta\}$. We say that a function F on $\mathcal{X}$ is $\mathcal{L}^2$-integrable uniformly in k if $\sup_{k \geq 1} \int_{\mathcal{X}} F^2(\boldsymbol{x}) \left[f_k^{1/2}(\boldsymbol{x} \mid \boldsymbol{y}, \phi) + f^{1/2}(\boldsymbol{x} \mid \boldsymbol{y}, \phi) \right]^2 \, d\boldsymbol{x} < \infty$. Note that all constant functions satisfy this condition. We make the following assumptions:

B1. For each $\delta > 0$, $\sup_{|\theta - \theta(\phi)| > \delta} Q(\theta \mid \phi) < Q(\theta(\phi) \mid \theta)$.

B2. $\mathcal{F}$ is Euclidean for an envelope F that is $\mathcal{L}^2$-integrable uniformly in k.

B3. $L(k) \int_{\mathcal{X}} |f_k(\boldsymbol{x} \mid \boldsymbol{y}, \phi) - f(\boldsymbol{x} \mid \boldsymbol{y}, \phi)| \, d\boldsymbol{x} = 0$.

B1 is a strong maximization assumption and serves as an identifying condition in the consistency proof. B1 can be established in various ways. For example, B1 holds if $Q(\theta \mid \phi)$ is strictly concave in θ. B1 also holds if Θ is compact, $Q(\theta \mid \phi)$ is continuous in θ, and $Q(\theta \mid \phi)$ is uniquely maximized at $\theta(\phi)$. Write P for the distribution of $\boldsymbol{X}$ given $\boldsymbol{y}$ and ϕ. A dominated convergence argument shows that $Q(\theta \mid \phi)$ is continuous in θ if $\mathcal{F}$ has an integrable envelope and $l(\boldsymbol{x}, \boldsymbol{y} \mid \theta)$ is continuous in θ for P almost all $\boldsymbol{x}$. Assumptions B2 and B3 are used to establish a uniform strong law of large numbers. B2 governs a stochastic term while B3 controls the bias in the Gibbs sampler. B3 requires the weakest form of convergence for the Gibbs sampler. See, for example, Gelfand and Smith (1990), Schervish and Carlin (1992), Chan (1993), Tierney (1994), and Sethuramen et al. (1996) for ways to check B3.

Theorem 3: *If B1 through B3 hold, then for each ϕ in Θ, $L(m,k)|\hat{\theta}(\phi) - \theta(\phi)| = 0$ as.*

Proof. We show that on a set of probability one,

$$L(m,k) \sup_{\theta \in \Theta} |Q_k^m(\theta \mid \phi) - Q(\theta \mid \phi)| = 0\,. \tag{1}$$

This and B1 will imply the result.

Write $Q_k(\theta \mid \phi)$ for $\int_{\mathcal{X}} l(\boldsymbol{x}, \boldsymbol{y} \mid \theta) f_k(\boldsymbol{x} \mid \boldsymbol{y}, \phi)\, d\boldsymbol{x}$. We prove (1) by showing that with probability one,

$$L(m) \sup_{k \geq 1, \theta \in \Theta} |Q_k^m(\theta \mid \phi) - Q_k(\theta \mid \phi)| = 0 \tag{2}$$

and

$$L(k) \sup_{\theta \in \Theta} |Q_k(\theta \mid \phi) - Q(\theta \mid \phi)| = 0\,. \tag{3}$$

Property (2) follows from B2, the fact that $\mathcal{F}$ does not depend on k, and Corollary 4(i) in Sherman (1994). Property (2) also follows from Theorem 2.8.1 in van der Vaart and Wellner (1996) taking $\mathcal{P}$ to be the set of probability measures corresponding to the densities $f_k(\cdot \mid \boldsymbol{y}, \phi)$ for $k \geq 1$. By the Cauchy-Schwarz inequality, the left-hand side of (3) is bounded by

$$\left(\int_{\mathcal{X}} F^2(\boldsymbol{x}) \left[f_k^{1/2}(\boldsymbol{x} \mid \boldsymbol{y}, \phi) + f^{1/2}(\boldsymbol{x} \mid \boldsymbol{y}, \phi) \right]^2 d\boldsymbol{x} \right)^{1/2} \left(\int_{\mathcal{X}} \left[f_k^{1/2}(\boldsymbol{x} \mid \boldsymbol{y}, \phi) - f^{1/2}(\boldsymbol{x} \mid \boldsymbol{y}, \phi) \right]^2 d\boldsymbol{x} \right)^{1/2}.$$

Use the fact that $|a^{1/2} - b^{1/2}| \leq |a - b|^{1/2}$ for all nonnegative numbers a and b and invoke assumptions B2 and B3 to get (3). *QED.*

Remark 3. There is a subtle point in the proof of Theorem 1 that is worth mentioning. Condition (2) states that the strong law of large numbers holds uniformly, not only over θ, but also over k. Uniformity in k is crucial for establishing almost-sure convergence of $\hat{\theta}(\phi)$ to $\theta(\phi)$ as m and k tend to infinity simultaneously and unconditionally. That condition (2) holds uniformly over k follows from (i) the fact that the class of functions $\mathcal{F} = \{l(\boldsymbol{x}, \boldsymbol{y} \mid \theta) : \boldsymbol{x} \in \mathcal{X}, \theta \in \Theta\}$ is Euclidean and does not depend on k and (ii) the envelope condition in B2. To see this, note that an empirical process maximal inequality (See the Maximal Inequality in Section 5, as well as the Main Corollary and Corollary 4(i) in Section 6 of Sherman, 1994) is used to bound the supremum over θ in (2). The upper bound of this maximal inequality depends on the Euclidean constants A and V associated with $\mathcal{F}$ and the second moment $\int_{\mathcal{X}} F^2(\boldsymbol{x}) f_k(\boldsymbol{x} \mid \boldsymbol{y}, \phi) d\boldsymbol{x}$ where F is the envelope for $\mathcal{F}$. The envelope condition in B2 implies that this second moment can be bounded by a constant that does not depend on k. Consider the Euclidean constants A and V. Since $\mathcal{F}$ does not depend on k, $Q_k^m(\theta \mid \phi) = \frac{1}{m} \sum_{i=1}^m l(\boldsymbol{X}_i^k(\phi), \boldsymbol{y} \mid \theta) = P_{k,\phi}^m l(\cdot \mid \boldsymbol{y}, \phi)$ depends on k only through the simulated data $\{\boldsymbol{X}_i^k(\phi)\}$. By definition, the constants A and V do not depend on any measure μ for which $\mu F^2 < \infty$. The envelope condition in B2 and a law of large numbers guarantees that for m large enough, $P_{k,\phi}^m F^2(\cdot) < \infty$. Thus, A and V do not depend on $P_{k,\phi}^m$, and therefore, they do not depend on k. Thus, convergence holds uniformly over k as well as θ.

Our next results provide primitive conditions that imply conditions A1 and A2$'$. Define $\mathcal{G} = \{\nabla_\theta l(\boldsymbol{x}, \boldsymbol{y} \mid \theta) : \boldsymbol{x} \in \mathcal{X}, \theta \in \Theta\}$ and $\mathcal{H} = \{\nabla_{\theta\theta} l(\boldsymbol{x}, \boldsymbol{y} \mid \theta) : \boldsymbol{x} \in \mathcal{X}, \theta \in \Theta\}$. Write $H(\theta \mid \phi)$ for $\nabla_{\theta\theta} Q(\theta \mid \phi)$ and $\|m\|$ for the matrix norm $(\sum_{i,j} m_{ij}^2)^{1/2}$. Call a function F on $\mathcal{X}$ $\mathcal{L}^2$-integrable uniformly over k and ϕ if $\sup_{k \geq 1, \phi \in \Theta} \int_{\mathcal{X}} F^2(\boldsymbol{x}) \left[f_k^{1/2}(\boldsymbol{x} \mid \boldsymbol{y}, \phi) + f^{1/2}(\boldsymbol{x} \mid \boldsymbol{y}, \phi) \right]^2 d\boldsymbol{x} < \infty$.

C1. Θ is compact and convex.

C2. For each ϕ in Θ, $\theta(\phi)$ uniquely maximizes $Q(\theta \mid \phi)$.

C3. $H(\theta \mid \phi)$ is continuous in θ and ϕ on $\Theta \otimes \Theta$.

C4. For each ϕ in Θ, $L(m,k)|\hat{\theta}(\phi) - \theta(\phi)| = 0 \ as.$

C5. $\mathcal{G}$ and $\mathcal{H}$ are Euclidean for envelopes that are $\mathcal{L}^2$-integrable uniformly over k and ϕ.

C6. $\sup_{\phi \in \Theta} \int_{\mathcal{X}} |f_k(\boldsymbol{x} \mid \boldsymbol{y}, \phi) - f(\boldsymbol{x} \mid \boldsymbol{y}, \phi)| \, d\boldsymbol{x} \leq Dr^k$ for some $D > 0$ and $r \in [0, 1)$.

Conditions C1 through C6 are used to establish the uniform convergence condition in A1′. Conditions C1, C2, and C3 are used in tandem with an implicit function theorem to establish A2′. Conditions C1 through C4 help ensure the uniform boundedness of $[H(\theta(\phi) \mid \phi)]^{-1}$. Conditions C5 and C6 are used to establish uniform strong laws for gradient and hessian processes. As before, C5 controls stochastic terms and C6 controls the bias in the Gibbs sampler.

Various sufficient conditions can be developed for establishing C1 through C6. For example, assumptions C2 and C4 are implied by assumptions B1 through B3. Condition C3 would follow from continuity of $\nabla_{\theta\theta} l(\boldsymbol{x}, \boldsymbol{y} \mid \theta)$ in θ and continuity of $f(\boldsymbol{x} \mid \boldsymbol{y}, \phi)$ in ϕ. Easy ways to establish C5 were discussed earlier in this section. Consider C6. Suppose that for each ϕ in Θ, the integral in C6 is bounded by $D(\phi)[r(\phi)]^k$ where $0 \leq r(\phi) < 1$, $D(\phi) > 0$, and $D(\phi)$ and $r(\phi)$ are continuous functions of ϕ on Θ. This together with C1 would imply C6. The pointwise bound on the integral is called a geometric ergodicity condition and implies a strong form of convergence for the Gibbs sampler. Schervish and Carlin (1992) and Chan (1993) state checkable conditions for geometric ergodicity.

Theorem 4: *If C1 through C6 hold, then A1 and A2′ hold for* $0 < \delta \leq 1/2$ *if* $\log m \ll k$.

Proof. Start with A2′. Condition C2 implies that for each ϕ in Θ, $|\det H(\theta(\phi) \mid \phi)| > 0$. This, C3, and the implicit function theorem (see, for example, Apostol, 1957, p.147) imply that $\theta(\phi)$ is continuously differentiable in a neighborhood of ϕ for each ϕ in Θ. Condition C2 also implies that $\theta(\phi)$ is a single-valued function for each ϕ in Θ. Deduce that $\theta(\phi)$ is continuously differentiable on Θ. This and C1 imply A2$'$.

Turn to A1$'$. Write $G_k^m(\theta \mid \phi)$ for $\nabla_\theta Q_k^m(\theta \mid \phi)$. By definition of $\hat{\theta}(\phi)$,

$$0 = G_k^m(\hat{\theta}(\phi) \mid \phi) .$$

Expand the right-hand side of the last equation about $\theta(\phi)$ to get

$$0 = G_k^m(\theta(\phi) \mid \phi) + [\hat{\theta}(\phi) - \theta(\phi)]' H_k^m(\theta^*(\phi) \mid \phi)$$

for $\theta^*(\phi)$ between $\hat{\theta}(\phi)$ and $\theta(\phi)$. Rearrange the last equation to get

$$\hat{\theta}(\phi) - \theta(\phi) = -[H_k^m(\theta^*(\phi) \mid \phi)]^{-1} G_k^m(\theta(\phi) \mid \phi) .$$

The following two conditions are sufficient to prove the result:

$$L(m,k) I\!\!E_{\boldsymbol{y}} \left[m^\delta \sup_{\phi \in \Theta} |G_k^m(\theta(\phi) \mid \phi)| \right] < \infty \tag{4}$$

$$L(m,k) \sup_{\phi \in \Theta} [H_k^m(\theta^*(\phi) \mid \phi)]^{-1} = O(1) \ as. \tag{5}$$

Start with (4). Write $G(\theta \mid \phi)$ for $\nabla_\theta Q(\theta \mid \phi)$. By definition of $\theta(\phi)$, for each ϕ in Θ,

$$G(\theta(\phi) \mid \phi) = 0 . \tag{6}$$

Write $G_k(\theta \mid \phi)$ for $\int_{\mathcal{X}} \nabla_\theta l(\boldsymbol{x}, \boldsymbol{y} \mid \theta) f_k(\boldsymbol{x} \mid \boldsymbol{y}, \phi)\, d\boldsymbol{x}$. Invoke C5, the fact that $\mathcal{G}$ does not depend on

k or ϕ (see Remark 3 after the proof of Theorem 3), and Corollary 4(i) in Sherman (1994) to get

$$L(m)\mathbb{E}_{\boldsymbol{y}}\left[m^{\delta}\sup_{k\geq 1,(\theta,\phi)\in\Theta\otimes\Theta}|G_k^m(\theta\mid\phi)-G_k(\theta\mid\phi)|\right]<\infty\,. \tag{7}$$

A dominated convergence argument using C5 allows us to pass the derivative operator through the integral to get $G(\theta\mid\phi)=\int_{\mathcal{X}}\nabla_{\theta}l(\boldsymbol{x},\boldsymbol{y}\mid\theta)f(\boldsymbol{x}\mid\boldsymbol{y},\phi)\,d\boldsymbol{x}$. Then, by C5 and C6 and an argument similar to the one used to prove condition (3) in Theorem 1,

$$\begin{aligned} m^{\delta}|G_k(\theta\mid\phi)-G(\theta\mid\phi)| &\leq m^{\delta}C\sup_{\phi\in\Theta}\int_{\mathcal{X}}|f_k(\boldsymbol{x}\mid\boldsymbol{y},\phi)-f(\boldsymbol{x}\mid\boldsymbol{y},\phi)|\,d\boldsymbol{x} \\ &= O(m^{\delta}r^{k})\,. \end{aligned}$$

This last bound is finite provided $\log m\ll k$. This and (7) imply that if $\log m\ll k$, then

$$L(m,k)\mathbb{E}_{\boldsymbol{y}}\left[m^{\delta}\sup_{(\theta,\phi)\in\Theta\otimes\Theta}|G_k^m(\theta\mid\phi)-G(\theta\mid\phi)|\right]<\infty\,. \tag{8}$$

Conditions (6) and (8) imply condition (4).

To establish (5), first note that $\|H_k^m(\theta^*(\phi)\mid\phi)-H(\theta(\phi)\mid\phi)\|$ is bounded by

$$\sup_{(\theta,\phi)\in\Theta\otimes\Theta}\|H_k^m(\theta\mid\phi)-H(\theta\mid\phi)\|+\sup_{\phi\in\Theta}\|H(\theta^*(\phi)\mid\phi)-H(\theta(\phi)\mid\phi)\|\,. \tag{9}$$

Apply C5 and C6 and argue as above for establishing (4) to see that with probability one,

$$L(m,k)\sup_{(\theta,\phi)\in\Theta\otimes\Theta}\|H_k^m(\theta\mid\phi)-H(\theta\mid\phi)\|=0\,.$$

Next, write $\gamma(\theta,\phi)$ for $\|H(\theta\mid\phi)-H(\theta(\phi)\mid\phi)\|$. By C3 and the continuity of $\theta(\phi)$, $\gamma(\theta,\phi)$ is continuous in both arguments. Deduce that $\Gamma(\theta) \equiv \sup_{\phi\in\Theta} \gamma(\theta, \phi)$ is continuous in θ. By C4 and the definition of $\theta^*(\phi)$, $L(m, k)|\theta^*(\phi) - \theta(\phi)| = 0$ *as.* This, the continuity of $\Gamma(\theta)$, and the fact that $\Gamma(\theta(\phi)) = 0$ imply that $L(m, k)\Gamma(\theta^*(\phi)) = 0$ *as.* Thus, on a set of probability one,

$$L(m, k) \sup_{\phi\in\Theta} \|H_k^m(\theta^*(\phi) \mid \phi) - H(\theta(\phi) \mid \phi)\| = 0\,. \tag{10}$$

Condition C3 and the continuity of $\theta(\phi)$ imply that $\det H(\theta(\phi) \mid \phi)$ is continuous in ϕ on Θ. Condition C2 implies that $|\det H(\theta(\phi) \mid \phi)| > 0$ for each ϕ in Θ. These last two facts and C1 imply that $\inf_{\phi\in\Theta} |\det H(\theta(\phi) \mid \phi)| > 0$. It follows that the components of $[H(\theta(\phi) \mid \phi)]^{-1}$ are uniformly bounded in ϕ on Θ. This and (10) imply (5), proving Theorem 4. *QED.*

Corollary 5: *Suppose C1 through C6 hold and* $L(i)|\theta^i - \hat{\theta}| = 0$. *Then* $L(i, m, k)|\hat{\theta}^i - \hat{\theta}| = 0$ *as. provided*

(a) *If* $0 \le M < 1$, *then for* $0 < \delta \le 1/2$ *and* $B > 1/\delta$, $i = O(m^{1/B})$ *and* $\log m \ll k$.

(b) *If* $M = 1$, *then for* $0 < \delta \le 1/2$ *and* $B > 2/\delta$, $i = O(m^{1/B})$ *and* $\log m \ll k$.

(c) *If* $M > 1$, *then for* $0 < \delta \le 1/2$ *and* $0 < B < \delta$, $i = O(\log_M m^B)$ *and* $\log m \ll k$.

If, in addition, $\sqrt{n}(\hat{\theta} - \theta_0) \Longrightarrow N(0, V)$, *then* $L(n)L(i, m, k)\sqrt{n}(\hat{\theta}^i - \theta_0) \Longrightarrow N(0, V)$.

Remark 4. Suppose the conditions of Corollary 5 hold and $0 \le M < 1$. Choose $\delta = 1/2$ and $B > 2$. Then the conclusions of the Corollary 5 hold provided i grows much slower than $\sqrt{m}$ and k grows much faster than $\log m$.

Remark 5. Note that if direct Monte-Carlo integration is possible, then conditions B3 and C6 can be dropped and the dependence on k ignored in all the results presented in the section.

4. An Application to Diffusion of New Products and Services

A problem of great practical importance to business decision makers is predicting market penetration of new products and services, that is, predicting their growth in net sales over time. In the simplest setting, the data consist of the points $(t_0, n_0), (t_1, n_1), \ldots, (t_q, n_q)$ where the t_j's are successive time points and the n_j's are the corresponding cumulative numbers of adopters of a new product, say. The t_j's are selected independently of the adoption process and are typically spaced equal time units (months, quarters, years) apart. There may also be information on the price charged for the product and the value of promotions during the time periods delimited by the t_j's. The objective is to describe some of the dynamics of the temporal diffusion and predict net sales into the future.

Let M denote the the total number of people in the population of interest and assume that M is known and constant over time. Let π denote the proportion of potential adopters of the new product. Thus, the quantity $M\pi$ represents the number of potential adopters of the product.

We view n_j as a realization of a stochastic process $N(t)$ at time t_j. We assume that $N(t)$ is a pure birth Markov process with stationary transition probabilities and adoption rate $[M\pi - N(t)]\lambda(N(t))$ (cf. Karlin and Taylor, 1984). The quantity $M\pi - N(t)$ represents the number of potential adopters and $\lambda(N(t))$ the individual adoption rate at time t.

Bass (1969, 1994) popularized the specification $\lambda(N(t)) = \alpha + \beta N(t)$. The parameter $\alpha \geq 0$ is called the innovator coefficient and can be interpreted as the strength of the tendency within each potential adopter to innovate, or adopt the new product irrespective of the behavior of other adopters. The parameter $\beta \geq 0$ is called the imitator coefficient and quantifies the tendency within each potential adopter to adopt due to the influence of other adopters. The quantity $\beta N(t)$ can be interpreted as the strength of this tendency within each potential adopter at time t.

The Bass model has been the workhorse in diffusion modeling over the course of the last 25 years. It has been used to describe the empirical adoption curve for a wide variety of new products and services, from consumer durables like refrigerators and television sets to services like telephone banking and financial investment instruments (see Sultan, Farley, and Lehmann (1990) for more details). The model is simple and has intuitive appeal in explaining how a homogeneous group of potential adopters comes to adopt a new product. Further, it can be modified in a straightforward manner to accommodate the effect of price changes and promotions as well as nonlinearity of $\lambda(N(t))$ in $N(t)$.

A number of different procedures have been proposed to estimate the parameters of the Bass model, from the ordinary least squares procedure proposed by Bass (1969) to the beta-binomial approximation of Dalal and Weerahandi (1995). See Mahajan and Wind (1986) for a detailed account of other procedures. Notable for its absence in this regard is the method of maximum likelihood. Direct maximum likelihood estimation of the parameters of the Bass model has been avoided in the past due to the intractability of the likelihood function. However, it is possible to view this estimation problem as an incomplete data problem and apply $MCEM$ technology to obtain parameter estimates that converge in an appropriate sense to maximum likelihood estimates. We develop these ideas in the next subsections.

4.1 $MCEM$ Estimation

Recall that we observe the points $(t_0, n_0), (t_1, n_1), \ldots, (t_q, n_q)$ where n_j is a realization of a process $N(t)$ at time t_j. For simplicity, take $(t_0, n_0) = (0, 0)$. We assume that $N(t)$ is a pure birth Markov process with stationary transition probabilities and population adoption rate $[M\pi - N(t)][\alpha + \beta N(t)]$. Let S_i denote the ith "sojourn time", or time between the ith and $(i+1)$st adoptions, $i = 0, 1, \ldots, n_q - 1$. The Markovian and stationarity assumptions guarantee that the

S_i's are independent and exponentially distributed (cf. Ross, 1983, p.142). The rate assumption implies that S_i has rate $(M\pi - i)(\alpha + \beta i)$. These facts form the basis for writing down the likelihood function for the observed data.

Define the parameter vector $\theta = (\alpha, \beta, \pi)$ and $\Lambda_i(\theta) = (M\pi - i)(\alpha + \beta i)$. Note that the population adoption rate $\Lambda_i(\theta)$ must be positive for all θ and for all i. Write $\boldsymbol{D}$ for the random vector $(0,0), (t_1, N(t_1)), \ldots, (t_q, N(t_q))$ and $\boldsymbol{d}$ for the realization $(0,0), (t_1, n_1), \ldots, (t_q, n_q)$. Write $f(\boldsymbol{d} \mid \theta)$ for the corresponding likelihood function.

For $s, t \geq 0$, write $P_{n,m}(t, \theta)$ for $I\!\!P(N(s+t) = n \mid N(s) = m)$. Stationarity implies that $P_{n,m}(t, \theta)$ does not depend on s. In particular, $P_{n,m}(t, \theta) = I\!\!P(N(t) = n \mid N(0) = m)$. Apply this fact and the Markovian property to write

$$f(\boldsymbol{d} \mid \theta) = \left[\prod_{j=1}^{k} P_{n_j, n_{j-1}}(t_j - t_{j-1}, \theta) \right] .$$

Standard results on birth processes (cf. Bartlett, 1978, p.59) show that

$$P_{n,m}(t, \theta) = \sum_{i=m}^{n} \omega_i(\theta) \exp(-\Lambda_i(\theta)\, t)$$

where

$$\omega_i(\theta) = \frac{\prod_{k=m}^{n-1} \Lambda_k(\theta)}{\prod_{k=m, k \neq i}^{n} [\Lambda_k(\theta) - \Lambda_i(\theta)]} .$$

Recall that θ_0 denotes the true value of the parameter vector. We would like to estimate θ_0 by the method of maximum likelihood. However, as reported in Dalal and Weerahandi (1995), except for very small values of $M\pi$, the observed data likelihood function $f(\boldsymbol{d} \mid \theta)$ cannot be reliably evaluated, much less maximized. This explains why direct maximum likelihood estimation has been avoided in the past. Notice, though, that we can view this estimation problem as a missing data problem where the individual adoption times are the missing data.

Write T_i for $\sum_{j=0}^{i-1} S_j$, the ith adoption time, and $\boldsymbol{T}$ for $(T_1, \ldots, T_{n_q})$. Let $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{n_q})$ denote a realization of $\boldsymbol{T}$ and write $f(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)$ for the complete-data likelihood function. Let Θ denote the parameter space for θ_0. We shall assume that Θ is such that $\Lambda_i(\theta) > 0$ for all θ in Θ and for all i. For the Bass model, $\Lambda_i(\theta) > 0$ provided $\alpha > 0$, $\beta \geq 0$, and $\pi \geq n_q/M$. Deduce that for all θ in Θ,

$$f(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta) = \left[\prod_{i=0}^{n_q-1} \Lambda_i(\theta) \exp(-\Lambda_i(\theta)[\tau_{i+1} - \tau_i]) \right] \exp(-\Lambda_{n_q}(\theta)[t_q - \tau_{n_q}])$$

where $\tau_0 = 0$, $\tau_i \leq \tau_j$ for $i < j$, and $\tau_{n_j} \leq t_j < \tau_{n_j+1}$, $j = 1, 2, \ldots, q$. The final factor is the probability of not observing the $(n_q + 1)$st adoption in the interval $[\tau_{n_q}, t_q]$. This function has a relatively simple form that is easy to maximize.

Because $f(\cdot \mid \boldsymbol{d}, \phi)$ involves the intractable incomplete data likelihood and the dimension of $\boldsymbol{T}$ is typically large, it is infeasible to evaluate $Q(\theta \mid \phi)$ with any direct method. Therefore, we turn to Monte-Carlo integration to implement the E step of the EM algorithm. Direct sampling from $f(\cdot \mid \boldsymbol{d}, \phi)$ is difficult, again due to the involvement of the incomplete data likelihood. As a result, we turn to the Gibbs sampling technique to facilitate Monte-Carlo integration.

The Gibbs sampler is a cyclic, iterative technique for generating observations from the joint distribution of a random vector when it is possible to sample from all full univariate conditional distributions. See Tanner (1996) for a good introduction. We now explain how to implement the Gibbs sampler in our problem.

We wish to generate m observations from $f_k(\cdot \mid \boldsymbol{d}, \phi)$ to construct the objective function $Q_k^m(\theta \mid$

ϕ). To generate one such observation we must be able to sample from the conditional distribution of each T_i given all the other adoption times, $\boldsymbol{D} = \boldsymbol{d}$, and ϕ.

Rearrange terms in the complete-data likelihood to get, for each ϕ in Θ,

$$f(\boldsymbol{\tau}, \boldsymbol{d} \mid \phi) = \left[\prod_{i=1}^{n_q} \Lambda_{i-1}(\phi) \exp(-[\Lambda_{i-1}(\phi) - \Lambda_i(\phi)]\tau_i) \right] \exp(-\Lambda_{n_q}(\phi) t_q) \tag{11}$$

where $\tau_i \leq \tau_j$ for $i < j$ and $\tau_{n_j} \leq t_j < \tau_{n_j+1}$, $j = 0, 1, \ldots, q$. Let $\boldsymbol{T}_{-i}$ denote the vector of adoption times excluding T_i. Write $\boldsymbol{\tau}_{-i}$ for a realization of $\boldsymbol{T}_{-i}$ and $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ for the density of T_i given $\boldsymbol{T}_{-i} = \boldsymbol{\tau}_{-i}$, $\boldsymbol{D} = \boldsymbol{d}$ and ϕ. Write $f(\boldsymbol{\tau}_{-i}, \boldsymbol{d} \mid \phi)$ for the joint density of $\boldsymbol{T}_{-i}$ and $\boldsymbol{D}$ given ϕ. Note that $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ is the ratio of $f(\boldsymbol{\tau}, \boldsymbol{d} \mid \phi)$ to $f(\boldsymbol{\tau}_{-i}, \boldsymbol{d} \mid \phi)$ on the interval $[l_i, u_i]$ where

$$\begin{aligned} l_i &= \tau_{i-1} + [t_j - \tau_{i-1}] \{i - 1 = n_j\}, \\ u_i &= \tau_{i+1} + [t_j - \tau_{i+1}] \{i = n_j\} \end{aligned}$$

with $\{A\}$ denoting the indicator function of the set A. Moreover, $f(\boldsymbol{\tau}_{-i}, \boldsymbol{d} \mid \phi)$ does not involve τ_i and only the ith factor of $f(\boldsymbol{\tau}, \boldsymbol{d} \mid \phi)$ involves τ_i. Deduce that

$$f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi) = c_i(\phi) \exp(-\delta_i(\phi)\tau_i)\{l_i \leq \tau_i \leq u_i\} \tag{12}$$

where $\delta_i(\phi) = \Lambda_{i-1}(\phi) - \Lambda_i(\phi)$. Since $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ must integrate to unity we see that

$$c_i(\phi) = \delta_i(\phi)/[\exp(-\delta_i(\phi) l_i) - \exp(-\delta_i(\phi) u_i)] .$$

The distribution function corresponding to $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ is a linear function of $\exp(-\delta_i(\phi)\tau_i)$ and so is easily inverted. Therefore, it is trivial to sample from $f(\tau_i \mid \boldsymbol{\tau}_{-i}, \boldsymbol{d}, \phi)$ using the probability integral transformation technique.

Here are the details of the Gibbs scheme. Let $\boldsymbol{T}^{(j)} = (T_1^{(j)}, \ldots, T_{n_q}^{(j)})$ denote the jth Gibbs iterate, $j = 0, 1, \ldots, k$. The case $j = 0$ specifies a set of starting values for the adoption times that satisfy $T_i^{(0)} \leq T_j^{(0)}$ for $i < j$ and $T_{n_j}^{(0)} \leq t_j < T_{n_j+1}^{(0)}$, $j = 0, 1, \ldots, k$, but are otherwise arbitrary. Define $T_0^{(j)} = 0$, $j = 1, 2, \ldots, k$. Starting with $j = 1$, draw $T_i^{(j)}$, $i = 1, 2, \ldots, n_q$, from the distribution with density

$$\frac{\delta_i(\phi)\exp(-\delta_i(\phi)\tau_i)}{\exp(-\delta_i(\phi)u_i^{(j)}) - \exp(-\delta_i(\phi)l_i^{(j)})}\{l_i^{(j)} \leq \tau_i \leq u_i^{(j)}\}$$

where

$$\begin{aligned} l_i^{(j)} &= T_{i-1}^{(j)} + \left[t - T_{i-1}^{(j)}\right]\{i - 1 = n\}, \\ u_i^{(j)} &= T_{i+1}^{(j-1)} + \left[t - T_{i+1}^{(j-1)}\right]\{i = n\}, \end{aligned}$$

and (t, n) is a component of $\boldsymbol{d}$. Repeat this last step for $j = 2, 3, \ldots, k$. The final Gibbs iterate, $\boldsymbol{T}^{(k)}$, is an observation from $f_k(\cdot \mid \boldsymbol{d}, \phi)$. Independently generate m such observations, then construct $Q_k^m(\theta \mid \phi)$.

4.2 Condition Verification

The purpose of this subsection is to show how to verify the main conditions of Theorem 3 and Theorem 4 for the Bass model. We do not verify the Gibbs convergence conditions (B3 and C6), but rather refer the interested reader to the sufficient conditions and the appropriate references given after the statements of the assumptions in Section 3. Also, to make short work of verifying conditions B1 and C2, we shall assume that π, the proportion of eventual adopters, is known. In some applications, this may be a natural assumption to make. For example, it may be reasonable to assume that an entire population of interest will eventually adopt certain durable goods (e.g., telephones, television sets, refrigerators) that may become viewed as necessities. In this case, $\pi = 1$. A straightforward, but lengthier argument is needed to verify these conditions when π must be estimated.

Start with B1. For the sake of simplicity, assume that Θ is compact. This is a reasonable assumption since if either α or β were to exceed unity, then the expected number of adopters in either the first or second time periods would exceed $M\pi$, the number of potential adopters, making market saturation in the first few time periods almost a certainty. Therefore, assuming that $\theta = (\alpha, \beta)$ lies in a big compact set is not a restriction in practice. For concreteness, we shall assume that $\Theta \subseteq [\epsilon, K] \otimes [0, K]$ where $\epsilon > 0$ and $\epsilon < K < \infty$. This ensures that $\Lambda_i(\theta) = (M\pi - i)(\alpha + \beta i)$ is bounded and bounded away from zero on Θ for each $i = 0, 1, \ldots, n_q$.

We establish condition B1 by showing that $Q(\theta \mid \phi)$ is strictly concave in θ. To do this, we require that $n_q > 1$. Recall that $Q(\theta \mid \phi) = \int_{\mathcal{T}} l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta) f(\boldsymbol{\tau} \mid \boldsymbol{d}, \phi)\, d\boldsymbol{\tau}$ where $\mathcal{T}$ denotes the support of $\boldsymbol{T}$ and

$$l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta) = \left[\sum_{i=0}^{n_q-1} \log \Lambda_i(\theta) - \Lambda_i(\theta)[\tau_{i+1} - \tau_i]\right] - \Lambda_{n_q}(\theta)[t_q - \tau_{n_q}]\,.$$

The matrix of second partial derivatives for $l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)$ equals

$$-\begin{pmatrix} \sum_i \frac{1}{(\alpha+\beta i)^2} & -\sum_i \frac{i}{(\alpha+\beta i)^2} \\ -\sum_i \frac{i}{(\alpha+\beta i)^2} & \sum_i \frac{i^2}{(\alpha+\beta i)^2} \end{pmatrix}$$

where the index i runs from 0 to $n_q - 1$. Notice that this matrix does not depend on $\boldsymbol{\tau}$. Recall that $H(\theta \mid \phi) = \nabla_{\theta\theta} Q(\theta \mid \phi)$. Since $\Lambda_i(\theta)$ is bounded and bounded away from zero on Θ and $\mathcal{T}$ is bounded, dominated convergence arguments allow us to pass the second partial derivative operator through the integral sign to get

$$H(\theta \mid \phi) = \int_{\mathcal{T}} \nabla_{\theta\theta} l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta) f(\boldsymbol{\tau} \mid \boldsymbol{d}, \phi)\, d\boldsymbol{\tau}\,.$$

It follows from these last two facts that $H(\theta \mid \phi)$ is equal to the matrix defined above. We now show that $-H(\theta \mid \phi)$ is positive definite, implying strict concavity of $Q(\theta \mid \phi)$. We do so by showing that all the leading principal minors of $-H(\theta \mid \phi)$ are positive. By inspection, all the diagonal elements are positive. The determinant equals

$$\sum_i \sum_j \frac{i^2 - ij}{(\alpha + \beta i)^2 (\alpha + \beta j)^2}\,.$$

Note that the (i, j) and (j, i) terms have the same denominator. The contribution from these terms is

$$(i - j)^2 / (\alpha + \beta i)^2 (\alpha + \beta j)^2$$

which is strictly positive for $i \neq j$. This establishes B1.

Write $\mathcal{F}$ for the class of functions $\{l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta) : \boldsymbol{\tau} \in \mathcal{T}, \theta \in \Theta\}$. Since $\mathcal{T}$ is bounded and $l(\boldsymbol{\tau}, \boldsymbol{d} \mid \theta)$ is continuous on the compact set Θ, it follows that $\mathcal{F}$ is a bounded class of functions and so has a constant envelope. Further, it is easy to verify that the class of partial derivatives of $\mathcal{F}$ is bounded on Θ. It then follows from Lemma 2.13 in Pakes and Pollard (1989) that $\mathcal{F}$ is Euclidean for a constant envelope. This establishes B2.

Now consider conditions C2 through C5. Conditions C2 and C3 follow from the proof of B1. C4 would follow from B1, B2, and B3, and Theorem 1. Consider C5. Note that partial derivatives with respect to θ of all orders of functions in $\mathcal{F}$ are uniformly bounded in all arguments. C5 follows from this and Lemma 2.13 in Pakes and Pollard (1989).

4.3 Simulations and an Application

In this subsection, we present simulation results illustrating the performance of the *MCEM* procedure for estimating the parameters of the Bass model described in Section 3. We also present results of applying the methodology to data on the number of subscribers of wireless telecommunications services in the United States.

In our simulation, we take $\theta_0 = (\pi_0, \alpha_0, \beta_0) = (.5, .03, .0004)$. We take $M = 2000$ for the population size so that the number of potential adopters is $M\pi_0 = 1000$. We take $(t_0, n_0) = (0, 0)$ and $q = 60$ equally spaced time points, namely, $t_1, t_2, \ldots, t_{60} = .125, .25, \ldots, 7.5$. We then generate the corresponding cumulative numbers $n_1, n_2, \ldots, n_{60}$. A typical sample path generated from this model appears in the upper left-hand plot in Figure 1. We then use the *MCEM* procedure described in Section 3 to estimate θ_0.

Specifically, we take $([n_{60} + 1]/2000, .0005, .00005)$ as starting values for the parameters and a set of starting values for the Gibbs algorithm satisfying the constraints described in Section 3. We then run the procedure with $i = 10$ iterations, $m = 30$ samples per iteration, and $k = 50$ Gibbs iterations per sample. We replicate this procedure 1000 times to obtain 1000 estimates of θ_0. Histograms of the parameter estimates appear in Figure 1.

All histograms have modes near the true parameter values. The histograms for α and β appear roughly normally distributed. The histogram for π is slightly skewed to the right. It is interesting to note that the sample variances for these estimates are very close to the corresponding variance estimates obtained from a Monte-Carlo implementation of the method of Louis (1982) for computing variance estimates for EM estimators. We also note that the $MCEM$ iterates appear to have converged after only a small number of iterations.

Next, we illustrate the use of the $MCEM$ methodology to predict the number of subscribers of wireless telecommunication services in the United States. These services are comprised of cellular telephone service, PCS (Personal Communication Services), and ESMR (Enhanced Specialized Mobile Radio) services. While count data of this sort need not follow a pure-birth model since subscribers may drop services, our results suggest that the pure-birth model may still be a reasonable approximation for this data.

The counts come from a semi-annual survey conducted since January, 1985 by CTIA, the Cellular Telecommunications Industry Association (see the web site *www.wow-com.com* for more information). After January of 1985, CTIA performed the semi-annual surveys every June and December. We have 25 observations from January of 1985 to December of 1996. These observations are represented as solid points in Figure 2. We shall use the first 23 points to estimate $\theta_0 = (\pi_0, \alpha_0, \beta_0)$ and then use the estimates to predict the number of wireless subscribers through December 2000. This will enable us to cross-validate the method for time points 24 and 25, the counts from the June 1996 and December 1996 surveys.

For this data, it is not clear how to determine the population size, M, since the same person may subscribe to more than one wireless service. For convenience, we take $M = 10^9$ so that π_0 is equal to $\frac{M^*}{M}\pi_0^*$ where M^* is the true population size for wireless services and π_0^* is the true proportion of potential adopters of wireless services. Note that $M\pi_0 = M^*\pi_0^*$ and so $M\pi_0$ still has the interpretation as the number of potential adopters of wireless services. By this means, we avoid having to specify M^*. Also note that that the choice of parametrization for the variable π does not affect estimation of α_0 and β_0 since π enters the complete data likelihood function only through $M\pi$.

Our estimate of θ_0, based on the first 23 observations, is $(.19, 4.3 \times 10^{-4}, 1.06 \times 10^{-9})$. As in our simulation, we obtain these estimates using $i = 10$ iterations, $m = 30$ samples per iteration, and $k = 50$ Gibbs steps per sample. The t-ratios for the parameter estimates are $(2.3, 2.5, 2.1)$. We obtain the standard error estimates used to construct the t-ratios from a Monte-Carlo implementation of the method of Louis (1982) mentioned previously. Our estimate of $M\pi_0$ is 1.9×10^8 with an estimated standard error of 8.3×10^7. Using these parameter estimates, we solve a system of 2 approximate partial differential equations to obtain estimates of the mean and variance of $N(t)$. Dalal and Weerahandi (1992) derive these approximations. The estimates of $\mathbb{E}N(t)$ for $t = 1, 2, \ldots, 23$ appear as circles in Figure 2. These estimates very accurately track the observed counts. The estimates of $\mathbb{E}N(t)$ for $t = 24, 25, \ldots, 33$ are represented as plus signs in Figure 2. The dashed lines interpolate points corresponding to ± 2 estimated standard errors from the estimates of the $\mathbb{E}N(t)$ values. Note that the estimates of $\mathbb{E}N(24)$ and $\mathbb{E}N(25)$ are close to the observed points, though they fall outside the dashed lines.

5. Summary

This paper establishes convergence conditions for an $MCEM$ sequence when the underlying EM sequence converges. High-level conditions are established for rates of almost-sure convergence and convergence in distribution. These conditions clarify fundamental requirements for convergence of any $MCEM$ sequence. Primitive conditions are developed for almost-sure monotonicity and almost-sure convergence when independent Gibbs samples are used to facilitate Monte-Carlo integration. These primitive conditions are not abstruse but accessible to practitioners. We verify the main primitive conditions for an important application requiring estimation of the parameters of the Bass model for the temporal diffusion of new products and services into business markets.

We present simulation results illustrating the performance of the $MCEM$ procedure in this context and apply the methodology to data on wireless telecommunications services.

We do not directly address the issue of when an EM sequence converges. Nor do we explore the nature of limit points of EM sequences. See Wu (1983) and Biscarat (1994) for more on these issues. We note that a Monte-Carlo version of the method of Louis (1982) for computing the observed information matrix using the complete data likelihood can be developed to help ascertain the nature of $MCEM$ limit points. Increased confidence in an $MCEM$ procedure can be achieved by running the procedure with different starting values representative of the parameter space and checking the nature of the limit points using Louis' method.

For simplicity, we have adopted a frequentist perspective in this paper. If we adopt a Bayesian stance, then statements about a maximum likelihood estimator can be reinterpreted as statements about a maximum posterior likelihood estimator when the prior is flat, or noninformative. If we have informative prior information on θ, then conditions for convergence to a mode of the posterior distribution of θ are identical to those developed in Section 3 after redefining $l(\boldsymbol{x}, \boldsymbol{y} \mid \theta)$ in Section 2 to be equal to $\log[\pi(\theta) f(\boldsymbol{x}, \boldsymbol{y} \mid \theta)]$ where $\pi(\theta)$ is the prior density for θ.

ACKNOWLEDGEMENTS. We thank the editor and two referees for useful comments and suggestions. One referee's particularly perceptive observations led to significant improvements in the manuscript.

REFERENCES

APOSTOL, T. M. (1957): *Mathematical Analysis: A Modern Approach to Advanced Calculus*, Reading, Massachusetts, Addison Wesley.

ALIPRANTIS, C. D., AND K. C. BORDER (1994): *Infinite Dimensional Analysis, A Hitchhiker's*

Guide, New York, Springer-Verlag.

Bartlett, M. S. (1978): *An Introduction to Stochastic Processes*, Cambridge, Cambridge University Press, 3rd Edition.

Bass, F. (1969): "A new product growth model for consumer durables." *Management Science*, 15, 215–227.

Bass, F. (1994): "Why the Bass model fits without decision variables." *Marketing Science*, 13, 203–223.

Biscarat, J. C. (1994):"Almost sure convergence of a class of stochastic algorithms," *Stochastic Processes and their Applications*, 50, 83–99.

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

Chan, K. S., and J. Ledolter (1995): "Monte-carlo EM estimation for time series models involving counts." *Journal of the American Statistical Association*, 90, 242–252.

Dalal, S., and S. Weerahandi (1992): "Some approximations for the moments of a process used in diffusion of new products." *Statistics and Probability Letters*, 15, 181–189.

Dalal, S., and S. Weerahandi (1995): "Estimation of innovation diffusion models with application to a consumer durable." *Marketing Letters*, 6, 123–136.

Dempster, A., Laird, N., and D. Rubin (1977): "Maximum likelihood from incomplete data via the EM algorithm (with discussion)." *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Gelfand, A., and A. Smith (1990): "Sampling-based approaches to calculating marginal densities." *Journal of the American Statistical Association*, 85, 398–409.

Guo, S. W., and E. A. Thompson (1991): "Monte-carlo estimation of variance component models for large complex pedigrees." *IMA Journal of Mathematics Applied in Medicine and Biology*, 8, 171–189.

Hajivassiliou, V. A., and P. A. Ruud (1994): "Classical estimation methods for LDV models using simulation", in *Handbook of Econometrics, Volume 4*, Editors R. F. Engle and D. L. McFadden, 2383–2441, North Holland, Amsterdam.

Karlin, S., and H. Taylor (1984): *An Introduction to Stochastic Modeling*, New York, Academic Press Inc.

Louis, T. (1982): "Finding observed information using the EM algorithm." *Journal of the Royal Statistical Society, Series B*, 44, 98–130.

Mahajan, V. and Y. Wind (1986): *Innovation Diffusion Models of New Product Acceptance*, Cambridge, MA, Ballinger.

Meng, X., and S. Schilling (1996): "Fitting full-information item factor models and an empirical investigation of bridge sampling." *Journal of the American Statistical Association*, 91, 1254–1267.

Nolan, D. and D. Pollard (1987): "U-processes: rates of convergence." *Annals of Statistics*, 15, 780–799.

Pakes, A., and D. Pollard (1989): "Simulation and the asymptotics of optimization estimators." *Econometrica*, 57, 1027–1057.

Ross, S. (1983): *Stochastic Processes*, New York, John Wiley and Sons.

Schervish, M. J., and B. P. Carlin (1992): "On the convergence of successive substitution sampling." *Journal of Computational and Graphical Statistics*, 1, 111–127.

Sethuramen, J., Athreya, K. B., and H. Doss (1996): "On the convergence of the markov chain simulation method." *Annals of Statistics*, 24, 69–100.

SHEPHARD, N. (1993): "Fitting non-linear time series models, with applications to stochastic variance models", *Journal of Applied Econometrics*", 8, 135–152.

SHERMAN, R. P. (1994): "Maximal inequalities for degenerate U-processes with applications to optimization estimators." *Annals of Statistics*, 22, 439–459.

SULTAN, F., FARLEY, J., AND D. LEHMANN (1990): "A meta-analysis of applications of diffusion models." *Journal of Marketing Research*, 27, 70–77.

TANNER, M. (1996): *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, New York, Springer-Verlag.

TIERNEY, L. (1994): "Markov chains for exploring posterior distributions." *Annals of Statistics*, 22, 1701–1762.

VAN DER VAART, A. W. AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*, New York, Springer-Verlag.

WEI, C. J. AND M. A. TANNER (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. (1983): "On the convergence properties of the EM algorithm." *Annals of Statistics*, 11, 95–103.