Bernoulli **6**(3), 2000, 457–489

The stochastic EM algorithm: Estimation and asymptotic results

SØREN FEODOR NIELSEN

Department of Theoretical Statistics, University of Copenhagen, Universitetsparken 5, DK-2100 København Ø, Denmark. E-mail: feodor@stat.ku.dk

The EM algorithm is a much used tool for maximum likelihood estimation in missing or incomplete data problems. However, calculating the conditional expectation required in the E-step of the algorithm may be infeasible, especially when this expectation is a large sum or a high-dimensional integral. Instead the expectation can be estimated by simulation. This is the common idea in the stochastic EM algorithm and the Monte Carlo EM algorithm.

In this paper some asymptotic results for the Stochastic EM algorithm are given, and estimation based on this algorithm is discussed. In particular, asymptotic equivalence of certain simple estimators is shown, and a simulation experiment is carried out to investigate this equivalence in small and moderate samples. Furthermore, some implementation issues and the possibility of allowing unidentified parameters in the algorithm are discussed.

Keywords: EM algorithm; incomplete observations; simulation

1. Introduction

Missing data problems often lead to complicated likelihood functions involving high-dimensional integrals or large sums, which are difficult if not impossible to calculate. Also any differentiation needed to find the maximum likelihood estimator (MLE) may be infeasible. The EM algorithm is an appealing method for maximizing the likelihood, since derivatives are not needed. Instead, from a given value of the unknown parameter the complete data log-likelihood is predicted and then maximized. Often the prediction – a conditional expectation of the complete data log-likelihood given the observed data – is easy to calculate and the maximization can be done either explicitly or by standard methods such as Newton–Raphson or scoring.

However, in particular with high-dimensional data or incomplete observations such as censored data, the conditional expectation is a high-dimensional integral or an integral over an irregular region and cannot be calculated explicitly. Instead one could try to estimate it by simulation. This is done in the stochastic EM algorithm suggested by Celeux and Diebolt (1986) – see Diebolt and Ip (1996) for a recent review and further references – and in the Monte Carlo EM (MCEM) algorithm (Wei and Tanner 1990). The main difference between these two algorithms is that the MCEM algorithm uses (infinitely) many simulations to obtain a good estimate of the conditional expectation (at least for the last

1350–7265

iterations of the algorithm), whereas the stochastic EM algorithm uses only one in each iteration.

In this paper asymptotic results applicable to both algorithms are shown. However, since we focus on the situation where the number of simulations in each iteration is small compared to the sample size (for reasons to be discussed in Section 4), we will use the term stochastic EM algorithm. Furthermore, for brevity we shall use the acronym StEM to denote the stochastic EM algorithm. Often SEM is used as an acronym, but we prefer StEM to avoid confusion with the simulated EM algorithm (Ruud 1991) and the supplementary EM algorithm (Meng and Rubin 1991), both of which are called SEM as well.

1.1. The StEM algorithm

Let $X_1, X_2 \ldots, X_n$ be independent identically distributed (i.i.d.) random variables from a distribution indexed by an unknown parameter θ. Suppose that a (many-to-one) mapping Y_i of X_i is observed rather than X_i. Throughout this paper X denotes a generic random variable from the unknown distribution and Y the corresponding incomplete observation.

The StEM algorithm takes the following form. From an arbitrary starting value $\tilde{\theta}_n(0)$ a sequence $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ is formed by going through a stochastic E-step or StE-step (simulation) and an M-step (maximization):

StE-step. Given a value of $\tilde{\theta}_n(k)$, simulate values $\tilde{X}_i$ from the conditional distribution of X given $Y = y_i$ under $\tilde{\theta}_n(k)$, i.e. draw $\tilde{X}_i \sim \mathscr{L}_{\tilde{\theta}_n(k)}(X|Y = y_i)$.

M-step. Maximize the resulting complete data log-likelihood, $\sum_{i=1}^{n} \log f_\theta(\tilde{X}_i)$, and let the maximizer be the next value, $\tilde{\theta}_n(k+1)$.

The StE-step completes the data set, and the maximization in the M-step is thus a complete data maximum likelihood estimation. Hence the M-step is typically easy to solve either explicitly or iteratively using standard algorithms such as Newton–Raphson or scoring.

It is clear that by simulating new independent $\tilde{X}_i$s in each step, the sequence of maximizers, $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$, is a time-homogeneous Markov chain, when the observed data, $y_1, y_2, \ldots, y_n$, are fixed (i.e. conditioned upon). If ergodic, the algorithm will converge in the sense that as k, the number of iterations, tends to infinity, $\tilde{\theta}_n(k)$ converges in distribution to a random variable, $\tilde{\theta}_n$, where $\tilde{\theta}_n$ is distributed according to the stationary distribution of the Markov chain. We shall show that $\tilde{\theta}_n$ is asymptotically normal and use this result to find estimators of the unknown parameter θ.

1.2. Outline of paper

In the following section we introduce notation and regularity assumptions to be used in the proofs of the large-sample results. Also some necessary facts about the convergence of sequences of Markov chains are discussed.

In Section 3 we first show that the Markov chains are well behaved – aperiodic, irreducible and Feller – and give weak sufficient conditions for ergodicity (Theorem 1). In Section 3.2 we prove that the stationary distribution of the Markov chain converges to a

normal distribution if suitably normalized (Theorem 2) as the sample size tends to infinity. This result has been suggested previously, for instance by Chan and Ledolter (1995), who gave a heuristic argument under stronger regularity assumptions, and proved for a special case by Celeux and Diebolt (1993). The general proof given in Section 3.2 appears to be new, however.

In Section 4 we give large-sample results for some simple estimators of the unknown parameter θ derived from the StEM algorithm, and show asymptotic equivalence of some estimators obtained from simple extensions of the StEM algorithm. A simulation study is given to illustrate small to moderate sample size behaviour of the various estimators, and estimation of the asymptotic variance is discussed.

The paper concludes with a discussion of some implementation issues and an indication of what effect unidentified parameters may have on the algorithm.

2. Preliminaries

In this section necessary results about sequences of Markov chains are given (Section 2.1), some notation is introduced (Section 2.2), and assumptions to be used in showing the asymptotic results are given (Section 2.3). The results in Section 2.1 are 'discrete-time' versions of some 'continuous-time' results given by Ethier and Kurtz (1986, Chapter 4).

2.1. Sequences of Markov chains

Let P_n be the transition probabilities for a Markov chain on a finite-dimensional Euclidean state space S and let μ_n be the corresponding stationary initial distributions, which are assumed to exist. $\mathscr{C}_b(S)$ denotes the set of continuous, bounded functions $\kappa: S \to \mathbb{R}$.

The following assumptions (referred to collectively as Assumption C) will be used:

Assumption C1. *$P_n(x, \cdot) \xrightarrow{w} P(x, \cdot)$ uniformly over compacta, i.e. for all compact sets, $K \subseteq S$, $\sup_{x \in K} \left| \int \kappa(y) P_n(x, \mathrm{d}y) - \int \kappa(y) P(x, \mathrm{d}y) \right| \to 0$ as $n \to \infty$ for all $\kappa \in \mathscr{C}_b(S)$.*

Assumption C2. *$x \to \int \kappa(y) P(x, \mathrm{d}y)$ is continuous for all $\kappa \in \mathscr{C}_b(S)$.*

Assumption C implies that $\int \kappa(y) P_n(\cdot, \mathrm{d}y)$ converges to $\int \kappa(y) P(\cdot, \mathrm{d}y)$ continuously, i.e. $\int \kappa(y) P_n(x_n, \mathrm{d}y) \to \int \kappa(y) P(x, \mathrm{d}y)$ when $x_n \to x$. Conversely, continuous convergence of $\int \kappa(y) P_n(\cdot, \mathrm{d}y)$ to $\int \kappa(y) P(\cdot, \mathrm{d}y)$ implies Assumptions C1 and C2 (Roussas 1972, p. 132). Thus assumption C may be replaced by:

Assumption C*. *$\int \kappa(y) P(\cdot, \mathrm{d}y)$ converges continuously to $\int \kappa(y) \mathrm{P}(\cdot, \mathrm{d}y)$ for each $\kappa \in \mathscr{C}_b(S)$, where P is a transition probability.*

We are interested in the limiting distribution of the stationary distributions. The following results characterizes the limit in terms of the limit of the transition probabilities.

Proposition 1. *Suppose that Assumption* C *holds. If a subsequence of* $(\mu_n)_{n\in\mathbb{N}}$ *converges weakly to a probability* μ*, then* μ *is a stationary initial distribution for the Markov chain with transition kernel P.*

Proof. Suppose $(\mu_{n'})_{n'}$ is a convergent subsequence with limit μ. Then

$$\begin{aligned}\left|\iint\kappa(y)P(x,\,\mathrm{d}y)\,\mathrm{d}\mu(x)-\int\kappa(x)\,\mathrm{d}\mu(x)\right| &\leqslant \left|\iint\kappa(y)P(x,\,\mathrm{d}y)\,\mathrm{d}\mu(x)-\iint\kappa(y)P(x,\,\mathrm{d}y)\,\mathrm{d}\mu_{n'}(x)\right| \\ &\quad+\left|\int\left(\int\kappa(y)P(x,\,\mathrm{d}y)-\int\kappa(y)P_{n'}(x,\,\mathrm{d}y)\right)\mathrm{d}\mu_{n'}(x)\right| \\ &\quad+\left|\int\kappa(x)\,\mathrm{d}\mu_{n'}(x)-\int\kappa(x)\,\mathrm{d}\mu(x)\right|. \qquad (1)\end{aligned}$$

Here the first and the third terms vanish as $n'\to\infty$. Let $\varepsilon>0$ be arbitrary. According to Prohorov's theorem, a compact set K can be chosen such that $\inf_{n'}\mu_{n'}(K)>1-\varepsilon/2$. Thus

$$\begin{aligned}&\left|\int\left(\int\kappa(y)P(x,\,\mathrm{d}y)-\int\kappa(y)P_{n'}(x,\,\mathrm{d}y)\right)\mathrm{d}\mu_{n'}(x)\right| \\ &\leqslant\frac{\varepsilon}{2}+\sup_{x\in K}\left|\int\kappa(y)P(x,\,\mathrm{d}y)-\int\kappa(y)P_{n'}(x,\,\mathrm{d}y)\right|, \qquad (2)\end{aligned}$$

where the second term can be made arbitrarily small (smaller than $\varepsilon/2$, say) by Assumption C1. □

From Proposition 1 follows:

Corollary 1. *Suppose Assumption* C *holds. If* μ *is the unique stationary distribution corresponding to P and* $(\mu_n)_{n\in\mathbb{N}}$ *is tight, then* $\mu_n\xrightarrow{w}\mu$.

Proof. Tightness implies that any subsequence of $(\mu_n)_{n\in\mathbb{N}}$ has a convergent sub-subsequence, $(\mu_{n'})$. According to Proposition 1, $\mu_{n'}\xrightarrow{w}\mu$, and the result follows from Theorem 2.3 in Billingsley (1968). □

In order to apply Corollary 1, a criterion for tightness is necessary:

Proposition 2. *Suppose that, for each* $n\in\mathbb{N}$, $(Z_k^n)_{k\in\mathbb{N}_0}$ *is an ergodic Markov chain on a finite-dimensional Euclidean state space S with initial distibution* υ_n *and transition kernel* P_n*. Let* μ_n *be the corresponding stationary initial distribution. If there exist functions* $\varphi_n: S\to[0;\,\infty[$, $\psi_n: S\to\mathbb{R}$, *and* $\psi: S\to[-C;\,\infty[$ *for some* $C>0$, *such that*

(i) $\int\varphi_n\,\mathrm{d}\upsilon_n$ *is finite for all* $n\in\mathbb{N}$,
(ii) $\psi_n\geqslant\psi$ *for all* $n\in\mathbb{N}$,
(iii) ψ *is a norm-like function, i.e.* $\overline{\{z:\,\psi(z)\leqslant c\}}$ *is compact for each* $c>0$,
(iv) and $\mathrm{E}(\varphi_n(Z_l^n))\leqslant\mathrm{E}(\varphi_n(Z_{l-1}^n))-\mathrm{E}(\psi_n(Z_{l-1}^n))$ *for all* $n,\ l\in\mathbb{N}$,
then $(\mu_n)_{n\in\mathbb{N}}$ *is tight.*

Proof. Letting $K_c = \overline{\{z: \psi(z) \leqslant c\}}$, we get $\psi_n \geqslant c \cdot 1_{K_c^c} - C \cdot 1_{K_c} = c - (C + c) \cdot 1_{K_c}$ by (ii). Now

$$0 \leqslant \mathrm{E}(\varphi_n(Z_l^n)) \leqslant \mathrm{E}(\varphi_n(Z_0^n)) - \mathrm{E}\left(\sum_{k=0}^{l-1} \psi_n(Z_k^n)\right)$$

$$\geqslant \int \varphi_n d\upsilon_n + (C + c)\mathrm{E}\left(\sum_{k=0}^{l-1} 1_{K_c}(Z_k^n)\right) - cl, \tag{3}$$

so that

$$\frac{c}{C + c} - \frac{1}{l}\int \varphi_n \mathrm{d}\upsilon_n \frac{1}{C + c} \leqslant \mathrm{E}\left(\frac{1}{l}\sum_{k=0}^{l-1} 1_{K_c}(Z_k^n)\right). \tag{4}$$

When $l \to \infty$ the right-hand side converges to $\mu_n(K_c)$ and the second term on the left-hand side vanishes. As c may be chosen arbitrarily large and K_c is compact, $(\mu_n)_{n \in \mathbb{N}}$ is tight. □

Remark. Conditions (i)–(iv) of the proposition need only hold for n sufficiently large. Condition (i) holds if (as will typically be the case) υ_n is degenerate, i.e. if Z_0^n is fixed.

ψ is norm-like if it is continuous and $\psi(z) \to \infty$ for $\|z\| \to \infty$.

A sufficient condition for condition (iv) is that $\mathrm{E}(\varphi_n(Z_l^n)|Z_{l-1}^n) \leqslant \varphi_n(Z_{l-1}^n) - \psi_n(Z_{l-1}^n)$, and this will typically be easier to verify in practice. This is equivalent to showing that $\varphi_n(Z_l^n) + \sum_{k=0}^{l-1}\psi_n(Z_k^n)$ is a super-martingale for each $n \in \mathbb{N}$.

2.2. Notation

Let f_θ be the density of X with respect to some dominating probability measure, μ. The resulting probability measure is denoted P_θ, and expectation with respect to this probability is denoted E_θ. Let $s_x(\theta)$ be the corresponding score function and $V(\theta) = \mathrm{E}_\theta(s_X(\theta)^{\otimes 2})$ the complete data information.

The conditional density of X given $Y = y$ with respect to some probability measure υ_y is denoted $k_\theta(x|y)$, and the corresponding score function $s_{x|y}(\theta)$. Let $I_y(\theta) = \mathrm{E}_\theta(s_{X|y}(\theta))^{\otimes 2}|Y = y)$.

The density of Y is denoted h_θ and the corresponding score function is $s_y(\theta)$. Let $I(\theta) = \mathrm{E}_\theta(s_Y(\theta)^{\otimes 2})$ denote the observed data information.

Notice that $s_y(\theta) = s_x(\theta) - s_{x|y}(\theta)$ and that $I(\theta) = V(\theta) - \mathrm{E}_\theta I_Y(\theta)$, when $\mathrm{E}_\theta(s_{X|y}(\theta)|Y = y) = 0$, as we will assume below.

Let $F(\theta) = \mathrm{E}_\theta I_Y(\theta)V(\theta)^{-1}$ be the expected fraction of missing information, and let θ_0 denote the true unknown value of $\theta \in \Theta \subseteq \mathbb{R}^d$. It is easily shown that $F(\theta)$ has d eigenvalues in]0; 1[, when both $I(\theta)$ and $\mathrm{E}_\theta I_Y(\theta)$ are non-singular, as we will assume below.

As in Section 1, we shall use $\tilde{X}_i$ for simulated values. The distribution of the simulated values is denoted $\tilde{P}_\theta$. This is of course just the conditional distribution of the unobserved X_is given the observed values of $Y_i = y_i$ under P_θ. Generally the simulated values are not

from the distribution indexed by the same value of the parameter θ as the observed Y_is (i.e. the correct value, θ_0), and the notation introduced should help to keep clear the distinction between simulated and observed variables and their distributions. Notice that the $\tilde{P}_\theta$ notation is shorthand in the sense that the dependence upon the observed y_is is suppressed.

2.3. Assumptions

The assumptions can be divided roughly into three groups corresponding to which model – the observed, the complete, or the incomplete data model – they relate to.

On the model for the observed data, $Y_1, Y_2, \ldots, Y_n$, it is assumed that the unknown parameter, θ, is identifiable. Furthermore, we assume that there is an MLE, $\hat{\theta}_n$, which solves the likelihood equation, $\frac{1}{n}\sum_{i=1}^{n} s_{y_i}(\theta) = 0$, such that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\mathscr{D}} N(0, I(\theta_0)^{-1})$. Note that we here implicitly assume that $I(\theta_0)$ is non-singular.

It is also assumed that $\hat{\theta}_n$ converges to θ_0 almost surely. This assumption is not necessary but it will simplify the proofs of Section 3.2. The strong consistency may be replaced by taking almost surely convergent subsequences of arbitrary subsequences and applying Theorem 2.3 in Billingsley (1968).

The assumptions on the observed data model may be difficult to verify in practice, but will typically follow if the complete data model and the missing data model are sufficiently smooth. The assumptions do not appear to be unreasonable since the StEM algorithm attempts to mimic maximum likelihood estimation. Hence, we would not expect it to have better properties than maximum likelihood. We note that in the proofs to follow the assumption that θ is identifiable is never explicitly used. We conjecture, however, that some of the regularity assumptions – in particular, those involved in the discussion of tightness (cf Proposition 3) – will fail to hold if θ is unidentifiable. We return to this discussion in Section 5.2.

On the model for the complete data, $X_1, X_2, \ldots, X_n$, it is assumed that (for n sufficiently large) there is (with probability 1) an MLE.

We must also assume that if $\theta_n = \theta_0 + O(1/\sqrt{n})$ then, for all almost all y-sequences,

$$\sqrt{n}(\tilde{\theta}_n(1) - \theta_n) = V(\theta_0)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{\tilde{X}_i}(\theta_n) + o_{\tilde{P}_{\theta_n}}(1), \tag{5}$$

where $\tilde{X}_i \sim \mathscr{L}_{\theta_n}(X|Y = y_i)$ and $\tilde{\theta}_n(1)$ is the complete data MLE based on the simulated data. This condition may be interpreted as an assumption of the complete data MLE based on approximately correct simulations is approximately efficient. This is close to assuming local efficiency of the complete data MLE. If $\tilde{\theta}_n(1)$ is uniformly asymptotically linear (Bickel *et al.* 1993, Definition 2.2.6) with influence function $V(\theta_0)^{-1} s_{\tilde{X}_i}$ in the model where $Y_i \sim P_{\theta_0}$ and $\tilde{X}_i \sim \mathscr{L}_{\tilde{\theta}}(X|Y = y_i)$, with $\tilde{\theta}$ playing the role of the unknown parameter, then (5) holds.

It is difficult to give simple, yet general, sufficient conditions for (5), but it should not be difficult to verify in actual applications. For instance, it is easily seen to hold if the complete data model is a full exponential family. If the complete data model is smooth (in the sense of Lehmann 1983, say), (5) follows if we show that

$$-\frac{1}{n}\sum_{i=1}^{n} D_\theta s_{\tilde{X}_i}(\theta_n) \xrightarrow{\tilde{P}_{\theta_n}} V(\theta_0) \tag{6}$$

for almost every y-sequence when $\theta_n = \theta_0 + O(1/\sqrt{n})$ and either that the integrable majorizer of the third derivative can be chosen to depend on y only or that a law of large numbers (again given y) applies to this majorizer.

Finally, some assumptions are necessary on the model for the missing data, i.e. the conditional distribution of X given $Y = y$. This is assumed to be regular in the sense of Bickel *et al.* (1993, Section 2.1) In particular, this means that $\mathrm{E}_{\theta_0} I_Y(\theta_0)$ is non-singular. Let

$$\theta \to D_\theta k_\theta^{\frac{1}{2}}(\cdot k|y) = \dot{k}_\theta^{\frac{1}{2}}(\cdot|y) \tag{7}$$

denote the ($L^2(\upsilon_y)$-) derivative of the root density,

$$\theta \to (k_\theta(\cdot|y))^{\frac{1}{2}} = k_\theta^{\frac{1}{2}}(\cdot|y). \tag{8}$$

Assume that for almost every y-sequence and every compact $C \subseteq \Theta$ we have the following:

Assumption U. $\sup_{\theta\in C}|\frac{1}{n}\sum_{i=1}^{n} I_{y_i}(\theta) - \mathrm{E}_{\theta_0} I_Y(\theta)| \to 0$ as $n \to \infty$.

Assumption D. $\forall h \in \mathbb{R}^d$,

$$\sup_{\theta\in C} \sup_{u\in[0;1/\sqrt{n}]} \frac{1}{n}\sum_{i=1}^{n} \int \left(h^{\mathrm{T}} \dot{k}_{\theta+uh}^{1/2}(x|y_i) - h^{\mathrm{T}} \dot{k}_\theta^{1/2}(x|y_i) \right)^2 \mathrm{d}\upsilon_{y_i}(x) \to 0$$

as $n \to \infty$.

Assumption L. $\forall \varepsilon > 0$: $\sup_{\theta\in C} \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathrm{E}}_\theta[(h^{\mathrm{T}} s_{X_i|y_i}(\theta))^2 1_{\{|h^{\mathrm{T}} s_{X_i|y_i}(\theta)| \geqslant \varepsilon/\sqrt{n}\}}] \to 0$ as $n \to \infty$.

Then (by a straightforward extension of Theorem II.6.2 in Ibragimov and Has'minskii 1981).

$$\sum_{i=1}^{n} \log k_{\theta+\frac{1}{\sqrt{n}}h}(X_i/y_i) - \sum_{i=1}^{n} \log k_\theta(X_i|y_i)$$

$$= h^{\mathrm{T}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{X_i|y_i}(\theta) - \frac{1}{2} h^{\mathrm{T}} \mathrm{E}_{\theta_0} I_Y(\theta) h + R_n(\theta, h), \tag{9}$$

where

$$\sup_{|h|\leqslant M} \sup_{\theta\in C} \tilde{P}_\theta\{|R_n(\theta, h)| > \varepsilon\} \to 0 \tag{10}$$

for every ε, $M > 0$ and every compact set $C \subseteq \Theta$, and given y

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{X_i|y_i}(\theta_n) \xrightarrow{\mathscr{D}} N(0, \mathrm{E}_{\theta_0} I_Y(\theta_0)) \tag{11}$$

for every sequence $(\theta_n)_{n\in\mathbb{N}}$ converging to θ_0.

Assumption L may be verified in practice by showing that a Lyapunov-type condition

holds. Assumptions U and D may be shown using empirical process techniques or from further smoothness.

Finally, the distributions of the conditional model, $(k_\theta(\cdot|y)\cdot\nu_y)_{\theta\in\Theta}$, must be mutually equivalent for each given y.

3. Asymptotic results for the stochastic EM algorithm

In Section 3.1 the convergence of the StEM algorithm (for a fixed sample size n) is discussed. The properties of the Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ are discussed, and sufficient criteria for ergodicity are given.

In Section 3.2 large-sample results for the sequence $(\tilde{\theta}_n)_{n\in\mathbb{N}}$ are given. The main result is Theorem 2, where the limiting distribution of $\sqrt{n}(\tilde{\theta}_n-\theta_0)$ is identified as $n\to\infty$. A criterion for tightness of the sequence $\sqrt{n}(\tilde{\theta}_n-\theta_0)$ is given.

We apply the obtained result to a simple example in Section 3.3.

3.1. Convergence

It is clear that by simulating new independent $\tilde{X}_i$s in each step, the sequence of maximizers, $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$, is a time-homogeneous Markov chain. We wish to show that the Markov chain is ergodic, i.e. that the distribution of $\tilde{\theta}_n(k)$ converges in total variation to the stationary initial distribution as $k\to\infty$. Ergodicity is expected to hold quite generally, since there clearly is a drift in the Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$. At each iteration, we maximize an unbiased estimate of the conditional expectation $Q(\theta|\tilde{\theta}_n(k-1))=\sum_{i=1}^n \mathrm{E}_{\tilde{\theta}_n(k-1)}(\log f_\theta(X)|Y=y_i)$ from the E-step of the EM algorithm. Consequently, we expect that on average the new θ value, $\tilde{\theta}_n(k)$, will increase the observed data log-likelihood just as the resulting θ value from one iteration of the EM algorithm would. This idea is used to give sufficient conditions for ergodicity in Theorem 1.

Also the simulated x values make up a Markov chain; this is denoted $(\tilde{X}(k))_{k\in\mathbb{N}_0}$ (suppressing the dependence on n). We start by showing a few properties of the Markov chains $(\tilde{X}(k))_{k\in\mathbb{N}_0}$ and $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$.

Lemma 1. *The Markov chain $(\tilde{X}(k))_{k\in\mathbb{N}_0}$ is irreducible and aperiodic.*

Remark. In the proofs of this subsection the dependence on n will be suppressed in the notation. We may – and will – assume that $\operatorname{supp}\nu_y=\operatorname{supp}(k_\theta(\cdot|y)\cdot\nu_y)$ for all $\theta\in\Theta$ since the conditional distributions are mutually equivalent.

Proof. As $\tilde{P}\{\tilde{X}(1)\in B|\tilde{X}(0)=x\}=\tilde{P}\{\tilde{X}(1)\in B|\tilde{\theta}(0)=\theta\}=\int_B k_\theta(x'|y)\,\mathrm{d}\nu_y(x')>0$ for any x – here θ is the complete data MLE corresponding to the observation x – and any measurable set B such that $\nu_y(B)>0$, the chain is ν_y-irreducible.

To show aperiodicity, suppose that the chain is periodic with period at least two. Then there exist two disjoint sets, D_1 and D_2, such that $\tilde{P}(\tilde{X}(1)\in D_2|\tilde{X}(0)=x\}=1$ for all

$x \in D_1$ (see Meyn and Tweedie 1993, Theorem 5.4.4). Consequently, $\int_{D_2} k_\theta(x|y)\,\mathrm{d}\nu_y(x) = 1$ for all values of θ, and D_1 must be a null set for all θ. In other words, $\tilde{P}\{\tilde{X}(1) \in D_1 | \tilde{X}(0) = x\} = 0$ for all x, contradicting Theorem 5.4.4 in Meyn and Tweedie (1993). Thus the chain is aperiodic. □

The Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ inherits some of the properties of the chain $(\tilde{X}(k))_{k\in\mathbb{N}_0}$.

Corollary 2. *The Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ is irreducible and aperiodic. If the Markov chain $(\tilde{X}(k))_{k\in\mathbb{N}_0}$ is ergodic, then so is $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$. In particular, if the simulations have a finite sample space, $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ is ergodic.*

The assumed regularity of the conditional distributions has the following consequence:

Lemma 2. *The Markov chain $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ has the weak Feller property, and compact sets are small.*

Proof. The assumed regularity of the model implies that for all measurable sets B we obtain $\int_B k_{\theta'}(x|y)\,\mathrm{d}\nu_y(x) \to \int_B k_\theta(x|y)\nu_y(x)$ when $\theta' \to \theta$. In particular, for an open set $O \subseteq \Theta$, $\tilde{P}\{\tilde{\theta}(1) \in O | \tilde{\theta}(0) = \theta'\} \to \tilde{P}\{\tilde{\theta}(1) \in O | \tilde{\theta}(0) = \theta\}$ as $\theta' \to \theta$. Thus the Markov chain has the weak Feller property.

If $K \subseteq \Theta$ is compact, then there is for each measurable B a $\theta' \in K$ such that $\inf_{\theta\in K} \int_B k_\theta(x|y)\,\mathrm{d}\nu_y(x) = \int_B k_{\theta'}(x|y)\,\mathrm{d}\nu_y(x)$ since the map $\theta \to \int_B k_\theta(x|y)\,\mathrm{d}\nu_y(x)$ is continuous. Thus K is small. □

It is worth noticing that $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ typically has a smaller state space than Θ. For instance, if the sample space of the simulations is finite, then so is the actual state space of the Markov chain. Also the observed y values may restrict the sample space of the simulations. The actual state space is the set of possible complete data maximum likelihood estimates based on simulated data; or, more precisely, the image of the support of the conditional distribution of $(X_i)_{i=1,\ldots,n}$ given $(Y_i)_{i=1,\ldots,n} = (y_i)_{i=1,\ldots,n}$ under the mapping that transforms complete data to the corresponding complete data MLE. We let $\tilde{\Theta}_n$ denote the actual state space; the dependence on $y_1, y_2, \ldots, y_n$ is suppressed.

Since compact sets are small, we get:

Corollary 3. *If $\tilde{\Theta}_n$ is compact, then $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$, is ergodic.*

More generally, we would try to verify a drift criterion to show ergodicity. The drift in the EM algorithm towards high-density areas suggests looking at the observed data log-likelihood:

Theorem 1. *Put $v(\theta) = \log h_{\hat{\theta}}(y) - \log h_\theta(y)$. Let $c = \log \int k_{\tilde{\theta}}(\tilde{x}|y)\,\mathrm{d}\nu_y(\tilde{x})$, where $\tilde{\theta}$ is the complete data MLE based on $\tilde{x}$.*

Suppose that $c < \infty$ and that the function

$$\theta \to \Delta(\theta) = \tilde{\mathrm{E}}(\log f_{\tilde{\theta}(k+1)}(\tilde{X}) - \log f_{\tilde{\theta}(k)}(\tilde{X})|\tilde{\theta}(k) = \theta), \tag{12}$$

where $\tilde{X} \sim \mathscr{L}_{\tilde{\theta}(k)}(X|Y = y)$, *is greater than* $(1 + \delta)c$ *outside a compact set* $K \subseteq \tilde{\theta}_n$ *for some* $\delta > 0$. *Then:*

(i) there is a stationary initial distribution, $\tilde{P}$, *of the Markov chain* $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$. *In particular, the distribution of* $\tilde{\theta}_n(k)$ *converges in total variation to* $\tilde{P}$ *as* $k \to \infty$ *for* $\tilde{P}$*-almost all values of* $\tilde{\theta}_n(0)$.
(ii) if $\theta \to v(\theta)$ *is norm-like, then the Markov chain* $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ *is ergodic.*
(iii) if $\theta \to \Delta(\theta)$ *is norm-like, then the Markov chain* $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ *is ergodic.*

Proof. First note that $\Delta(\theta)$ is positive, since $\tilde{\theta}(k+1)$ maximizes $\theta \to \log f_\theta(\tilde{X})$. Now

$$\begin{aligned} v(\tilde{\theta}(k+1)) &= v(\tilde{\theta}(k)) + \log h_{\tilde{\theta}(k)}(y) - \log h_{\tilde{\theta}(k+1)}(y) \\ &= v(\tilde{\theta}(k)) + (\log f_{\tilde{\theta}(k)}(\tilde{X}) - \log f_{\tilde{\theta}(k+1)}(\tilde{X})) \\ &\quad - (\log k_{\tilde{\theta}(k)}(\tilde{X}|y) - \log k_{\tilde{\theta}(k+1)}(\tilde{X}|y)), \end{aligned} \tag{13}$$

where $\tilde{X} \sim \mathscr{L}_{\tilde{\theta}(k)}(X|Y = y)$. Now by Jensen's inequality

$$\begin{aligned} &-\tilde{\mathrm{E}}(\log k_{\tilde{\theta}(k)}(\tilde{X}|y) - \log k_{\tilde{\theta}(k+1)}(\tilde{X}|y)|\tilde{\theta}(k)) \\ &\leqslant \log \tilde{\mathrm{E}}\left(\frac{k_{\tilde{\theta}(k+1)}(\tilde{X}|y)}{k_{\tilde{\theta}(k)}(\tilde{X}|y)}\middle|\tilde{\theta}(k)\right) = \log \int k_{\tilde{\theta}(k+1)}(\tilde{x}|y)\mathrm{d}\nu_y(\tilde{x}) = c. \end{aligned} \tag{14}$$

Hence

$$\tilde{\mathrm{E}}(v(\tilde{\theta}(k+1))|\tilde{\theta}(k) = \theta) \leqslant v(\theta) - \Delta(\theta) + c. \tag{15}$$

Now:

(i) since the chain is Feller and $\Delta(\theta) - c \geqslant \delta c - (1 + \delta)c1_K(\theta)$, (15) ensures the existence of a stationary initial distribution, $\tilde{P}$ (Meyn and Tweedie 1993, Theorem 12.3.4). This implies convergence in total variation for $\tilde{P}$-almost all starting values.
(ii) if $\theta \to v(\theta)$ is norm-like, then the chain is Harris recurrent according to Theorems 9.1.8 and 9.4.1 in Meyn and Tweedie (1993). In combination with the positivity shown in (i), this gives ergodicity (Meyn and Tweedie 1993, Theorem 13.3.3).
(iii) if $\theta \to \Delta(\theta)$ is norm-like, then $\Delta(\theta) - c \geqslant (1 + \Delta(\theta))/2 - (c + 1/2)1_{K_1}(\theta)$ for the compact set $K_1 = \overline{\{\theta \in \tilde{\theta}_n : \Delta(\theta) \leqslant 1 + 2c\}}$, which by Theorem 14.0.1 in Meyn and Tweedie (1993) ensures ergodicity. □

Remark. Since $\int k_{\tilde{\theta}}(\tilde{x}|y)\,\mathrm{d}\nu_y(\tilde{x}) \geqslant \int k_{\hat{\theta}}(\tilde{x}|y)\,\mathrm{d}\nu_y(\tilde{x}) = 1$, c is positive but may be infinite. It is finite if $(\theta, x) \to k_\theta(x|y)$ is bounded; here it may be useful to remember that we need only consider $\theta \in \tilde{\theta}_n$. The assumption that $\log \int k_{\tilde{\theta}}(\tilde{x}|y)\,\mathrm{d}\nu_y(\tilde{x})$ is finite may be relaxed to assuming that $\tilde{\mathrm{E}}(\log k_{\tilde{\theta}(k+1)}(\tilde{X}|y) - \log k_{\tilde{\theta}(k)}(\tilde{X}|y)|\tilde{\theta}(k) = \theta)$ is bounded by some $c < \infty$ (as a function of θ). We expect this to hold quite generally since it is bounded by the mean of half the

(conditional) likelihood ratio test statistic (' $-2\log Q$') for testing the null hypothesis $\theta = \tilde{\theta}(k)$, which (in large samples) is approximately χ^2-distributed with d degrees of freedom.

Assuming that $\{\theta \in \tilde{\Theta}_n : \Delta(\theta) < (1+\delta)c\}$ is (contained in) a compact set holds if $\Delta(\theta)$ is norm-like.

The assumption of $\upsilon(\theta)$ being norm-like seems to be weak but difficult to check in practice. It holds if the observed data likelihood $\theta \to h_\theta(y)$ is continuous and goes to 0 as θ goes away from the observed data MLE.

The assumption in (iii) may be relaxed to K_1 defined in the proof above being compact. The implications of the inequality (15) are stronger than just ergodicity under the assumption made in (iii) (see Meyn and Tweedie 1993, Chapter 14) but they do not appear to be generally useful in this context.

3.2. Asymptotic normality

In this subsection asymptotic normality of the sequence $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is shown. The aim is to apply Corollary 1 to the sequence $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n)$ conditional on the observed y-sequence.

In order to do this, we have to look at how the transition probabilities from a fixed point of the sample space behave as n tends to infinity. Here the 'fixed points' of the Markov chains, $(\sqrt{n}(\tilde{\theta}_n(k) - \hat{\theta}_n))_{k\in\mathbb{N}_0}$, have the form $h = \sqrt{n}(\theta_n - \hat{\theta}_n)$. We will verify Assumption C* the following lemma. Hence we need to look at a convergent sequence, $h_n = \sqrt{n}(\theta_n - \hat{\theta}_n)$, of points in the sample space, i.e. θ values of the type $\theta_n = \hat{\theta}_n + (1/\sqrt{n})h + o(1/\sqrt{n})$, and show continuous convergence of the transition probabilities.

Lemma 3. *Let $\tilde{X}_n \sim \mathscr{L}_{\theta_n}(X|Y = y_i)$, where $\theta_n = \hat{\theta}_n + (1/\sqrt{n})h + o(1/\sqrt{n}) = \tilde{\theta}_n(0)$.*

For almost all y-sequences and conditional on y,

$$\sqrt{n}(\tilde{\theta}_n(1) - \hat{\theta}_n) \xrightarrow{\mathscr{D}} N(F(\theta_0)^{\mathrm{T}}h,\ V(\theta_0)^{-1}\mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}), \tag{16}$$

where $\tilde{\theta}_n(1)$ is the complete data MLE based on the simulated $\tilde{X}_i$s, and $F(\theta_0) = \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}$ is the expected fraction of missing information.

Proof. Observe that from (9) the difference of the log-likelihood, based on the simulated $\tilde{X}_i$, evaluated at θ_n and $\hat{\theta}_n$, respectively, is

$$l_n(\theta_n; \hat{\theta}_n) = h^{\mathrm{T}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{\tilde{X}_i|y_i}(\hat{\theta}_n) - \frac{1}{2} h^{\mathrm{T}} \mathrm{E}_{\theta_0} I_Y(\theta_0) h + o_{\tilde{P}_{\hat{\theta}_n}}(1). \tag{17}$$

From (5) follows that when $\tilde{X}_n \sim \mathscr{L}_{\hat{\theta}_n}(X|Y = y_i)$

$$\sqrt{n}(\tilde{\theta}_n(1) - \hat{\theta}_n) = V(\theta_0)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{\tilde{X}_i}(\hat{\theta}_n) + o_{\tilde{P}_{\hat{\theta}_n}}(1)$$

$$= V(\theta_0)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_{\tilde{X}_i|y_i}(\hat{\theta}_n) + o_{\tilde{P}_{\hat{\theta}_n}}(1) \tag{18}$$

since $\sum_{i=1}^{n} s_{y_i}(\hat{\theta}_n) = 0$. Therefore, under $\hat{\theta}_n$, i.e. when $\tilde{X}_i \sim \mathscr{L}_{\hat{\theta}_n}(X|Y = y_i)$,

$$\begin{pmatrix} l_n(\theta_n; \hat{\theta}_n) \\ \sqrt{n}(\tilde{\theta}_n(1) - \hat{\theta}_n) \end{pmatrix} \xrightarrow{\mathscr{D}} N\left(\begin{pmatrix} -\frac{1}{2} h^{\mathrm{T}} \mathrm{E}_{\theta_0} I_Y(\theta_0) h \\ 0 \end{pmatrix}, \begin{pmatrix} h^{\mathrm{T}} \mathrm{E}_{\theta_0} I_Y(\theta_0) h & h^{\mathrm{T}} F(\theta_0) \\ F(\theta_0)^{\mathrm{T}} h & V(\theta_0)^{-1} \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1} \end{pmatrix} \right) \tag{19}$$

according to (9). LeCam's third lemma implies that under θ_n

$$\sqrt{n}(\tilde{\theta}_n(1) - \hat{\theta}_n) \xrightarrow{\mathscr{D}} N(F(\theta_0)^{\mathrm{T}} h,\ V(\theta_0)^{-1} \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}), \tag{20}$$

as we wished to show. □

Lemma 3 shows that the transition probabilities of the Markov chain $(\sqrt{n}(\tilde{\theta}_n(k) - \hat{\theta}_n))_{k \in \mathbb{N}_0}$ converge continuously to the transition probabilities of a multivariate Gaussian AR(1) process.

Theorem 2. *Suppose $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n)$ is tight for almost every y-sequence. Then:*

(i) for almost every y-sequence and conditionally on y,

$$\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) \xrightarrow{\mathscr{D}} N(0,\ I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]);$$

(ii) unconditionally $\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{\mathscr{D}} N(0,\ I(\theta_0)^{-1}[2I - \{I + F(\theta_0)\}^{-1}])$.

Here $F(\theta_0) = \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}$ is the expected fraction of missing information.

Proof. From Lemma 3 and Corollary 1 it follows that the limiting stationary distibution of $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n)$ given almost every y-sequence is the one corresponding to the Gaussian AR(1) process with parameter $F(\theta_0)^{\mathrm{T}} = V(\theta_0)^{-1} \mathrm{E}_{\theta_0} I_Y(\theta_0)$ and innovation variance $V(\theta_0)^{-1} \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}$. This distribution is normal with expectation 0 and variance given by

$$\sum_{k=0}^{\infty} F(\theta_0)^{\mathrm{T}\,k} V(\theta_0)^{-1} \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1} F(\theta_0)^k = F(\theta_0)^{\mathrm{T}} \sum_{k=0}^{\infty} F(\theta_0)^{\mathrm{T}\,k} V(\theta_0)^{-1} F(\theta_0)^k$$

$$= F(\theta_0)^{\mathrm{T}} V(\theta_0)^{-1} \sum_{k=0}^{\infty} F(\theta_0)^{2k} = V(\theta_0)^{-1} F(\theta_0)\{I - F(\theta_0)^2\}^{-1}$$

$$= V(\theta_0)^{-1}\{I - F(\theta_0)\}^{-1} F(\theta_0)\{I + F(\theta_0)\}^{-1} = I(\theta_0)^{-1} F(\theta_0)\{I + F(\theta_0)\}^{-1}$$

$$= I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]. \tag{21}$$

Part (ii) follows from Lemma 1 in Schenker and Welsh (1987). □

Tightness still remains to be shown. We shall give a sufficient condition. Let M denote the EM update, i.e. the mapping which maps θ to $\arg\max_{\theta'} Q(\theta'|\theta)$, the result after one iteration of the EM algorithm. Recall from the general theory of the EM algorithm (see, for example, Dempster *et al.* 1977 that $\hat{\theta}_n$ is a fixed point of M. Suppose that $\theta \to M(\theta)$ is differentiable and let $\lambda_n(\theta)$ be the largest eigenvalue of $D_\theta M(\theta) = V(M(\theta))^{-1}$. $(1/n)\sum_{i=1}^n I_{y_i}(\theta)$. Since the model is regular $V(\theta) - (1/n)\sum_{i=1}^n I_{y_i}(\theta)$ is positive definite and $0 \leqslant \lambda_n(\hat{\theta}_n) < 1$. Continuity (also a consequence of regularity) ensures that $\lambda_n(\theta) < 1$ in a neighbourhood of $\hat{\theta}_n$.

Proposition 3. *Suppose that there is a $\lambda^* < 1$ such that $\lambda_n(\theta) \leqslant \lambda^*$ for all $\theta \in \Theta_n$ for n sufficiently large and almost every y-sequence. If there exists a $c < 1 - \lambda^*$ such that, for some $C < \infty$,*

$$\tilde{\mathrm{E}}(\sqrt{n}\|\tilde{\theta}_n(1) - M(\tilde{\theta}_n(0))\|_2 | \sqrt{n}(\tilde{\theta}_n(0) - \hat{\theta}_n) = h) \leqslant C + c\|h\|_2 \tag{22}$$

$\tilde{P}$-almost surely for n sufficiently (and almost every y-sequence) large, then the sequence $(\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n))_{n\in\mathbb{N}_0}$ is tight.

Proof. Since $M(\hat{\theta}_n) = \hat{\theta}_n$ we get

$$\sqrt{n}\|\tilde{\theta}(1) - \hat{\theta}_n\|_2 \leqslant \sqrt{n}\|\tilde{\theta}(1) - M(\tilde{\theta}(0))\|_2 + \sqrt{n}\|M(\tilde{\theta}(0)) - \hat{\theta}_n\|_2$$

$$\leqslant \sqrt{n}\|\tilde{\theta}(1) - M(\tilde{\theta}(0))\|_2 + \lambda^* \sqrt{n}\|\tilde{\theta}(0) - \hat{\theta}_n\|_2, \tag{23}$$

so that, for n sufficiently large,

$$\tilde{\mathrm{E}}(\sqrt{n}\|\tilde{\theta}_n(1) - \hat{\theta}_n\|_2) \leqslant \tilde{\mathrm{E}}(\sqrt{n}\|\tilde{\theta}_n(0) - \hat{\theta}_n\|_2) - (1 - \lambda^* - c)\tilde{\mathrm{E}}(\sqrt{n}\|\tilde{\theta}_n(0) - \hat{\theta}_n\|_2) + C. \tag{24}$$

Thus the assumptions of Propositions 2 hold with $\varphi_n(\cdot) = \|\cdot\|_2$ and $\psi_n(\cdot) = \psi(\cdot) = (1 - \lambda^* - c)\|\cdot\|_2 - C$. □

Remark. Note that Proposition 3 gives sufficient conditions for tightness given the observed y-sequence. This is what we need in Theorem 2. Of course, unconditional tightness of $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ follows easily from conditional tightness of $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n)$.

For exponential family models – with θ the expectation of the canonical statistic – $M(\tilde{\theta}_n(0)) = \tilde{\mathrm{E}}(\tilde{\theta}_n(1)|\tilde{\theta}_n(0))$ and

$$\tilde{\mathrm{E}}(\sqrt{n}\|\tilde{\theta}_n(1) - M(\tilde{\theta}_n(0))\|_2 | \sqrt{n}(\tilde{\theta}_n(0) - \hat{\theta}_n) = h) \leqslant \sqrt{\mathrm{tr}\{\widetilde{\mathrm{var}}(\tilde{\theta}_n(1)|\tilde{\theta}_n(0))\}}. \tag{25}$$

In this case (22) holds if the conditional variance of $\tilde{\theta}_n(1)$ given $\tilde{\theta}_n(0)$ is small compared to the distance of $\tilde{\theta}_n(0)$ to the observed data MLE, when this is large.

The assumptions used to obtain Proposition 3 are stronger than the assumptions used to prove the previous results: here we have assumed differentiability of M and existence of moments of the transition probabilities. We can ensure that the moments exist by reparametrizing so that the unknown parameter is restricted to lie in a bounded set. If $\sqrt{n}(\varphi(\tilde{\theta}_n) - \varphi(\hat{\theta}_n))$ is tight then we can apply Theorem 2 to this sequence and afterwards transform to obtain the asymptotic distribution of $\sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n)$.

Differentiability of M was used to ensure that the EM update is a contraction. Of course, this may hold without M being differentiable. If M is a contraction, then the EM algorithm converges to the observed data MLE. Examples exist where the EM algorithm has more than one fixed point, and in these examples M is only locally a contraction. However, this is typically (when θ is identified from the observed data) a small-sample problem. Thus we expect that in these examples the assumptions of Proposition 3 will hold (for n sufficiently large).

3.3. An example

In order to illustrate the theoretical results in the two previous subsections we look at a simple example.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. exponentially distributed random variables with mean θ. Suppose X_i is only observed if $X_i < c$ for some fixed positive c. The observed data MLE is $\hat{\theta}_n = (1/N)\sum_{i=1}^{n} X_i \wedge c$, where N is the number of uncensored X_is. Obviously, if $N = 0$, i.e. all the observations are censored, the MLE does not exist. Moreover, it is easy to see that the StEM algorithm will not converge in this case. However $P\{N = 0\}$ goes to zero exponentially fast, so in large samples this will not be a problem. To avoid notational difficulties, put $\hat{\theta}_n = 0$ if $N = 0$. Suppose for simplicity that it is the first N X_is that are observed. The regularity conditions are easily checked in this example.

In the $(k+1)$th iteration of the StEM algorithm we first simulate $\tilde{X}_n = c + \tilde{\theta}_n(k)\varepsilon_i$, where ε_i are i.i.d. standard exponentially distributed random variables, if X_i is censored and put $\tilde{X}_n = \mathrm{X}_i$ otherwise. In the M-step we get $\tilde{\theta}_n(k+1) = (1/n)\sum_{i=1}^{n} \tilde{X}_n = (1/n)\sum_{i=1}^{n} X_i \wedge c + (1/n)\sum_{i=N+1}^{n} \tilde{\theta}_n(k)\varepsilon_i$. Here the actual state space, Θ_n, is $[\frac{1}{n}\sum_{i=1}^{n} X_i \wedge c;\ \infty[$.

Simple but tedious manipulations show that (when $0 < N < n$)

$$\tilde{\mathrm{E}}(\log k_{\tilde{\theta}(k+1)}(\tilde{X}|y) - \log k_{\tilde{\theta}(k)}(\tilde{X}|y)|\tilde{\theta}(k)) \leqslant -n\left(\frac{n-N}{n}\log\left(\frac{n-N}{n}\right) + \frac{1}{n-N-1}\right) - N, \tag{26}$$

while

$$\Delta(\theta) = n\tilde{\mathrm{E}}\left(-\log\left(\frac{1}{n}\sum_{i=N+1}^{n}\varepsilon_i + \frac{1}{\theta n}\sum_{i=N+1}^{n} X_i \wedge c\right)\right) + \frac{1}{\theta}\sum_{i=N+1}^{n} X_i \wedge c - N$$

$$\to n\left(\log n + \gamma - \sum_{k=1}^{n-N-1}\frac{1}{k}\right) - N \qquad \text{as } \theta \to \infty. \tag{27}$$

Here $\gamma = 0.5772\ldots$ is Euler's constant. Subtracting the left-hand side of (26) fromt he left-hand side of (27) and dividing by n gives

$$-\frac{N}{n}\log\left(\frac{n-N}{n}\right) + \left(\log(n-N) - \sum_{k=1}^{n-N-2}\frac{1}{k}\right) + \gamma. \tag{28}$$

which is strictly positive. Since $\theta \to \log h_{\hat{\theta}}(y) - \log h_{\theta}(y)$ is norm-like, the Markov chain is ergodic by Theorem 1(ii). Of course, if $N = n$ there are no missing data and the Markov chain is trivially ergodic.

We see that

$$M(\tilde{\theta}_n(k)) = \tilde{\mathrm{E}}(\tilde{\theta}_n(k+1)|\tilde{\theta}_n(k)) = \frac{1}{n}\sum_{i=1}^{N} X_i + \frac{n-N}{n}(c + \tilde{\theta}_n(k)), \tag{29}$$

so that $\lambda_n(\theta) = D_\theta M(\theta) = (n-N)/n$. Since $(n-N)/n \to \exp(-c/\theta_0)$ for almost every y-sequence, $\lambda_n(\theta) \leqslant \lambda^*$ for any $\lambda^* > \exp(-c/\theta_0)$ for n sufficiently large. To show tightness, note that

$$\tilde{\mathrm{E}}(\sqrt{n}|\tilde{\theta}_n(1) - M(\tilde{\theta}_n(0))||\sqrt{n}(\tilde{\theta}_n(0) - \hat{\theta}_n)) \leqslant \frac{\sqrt{n-N}}{n}\tilde{\theta}_n(0)$$

$$\leqslant \frac{1}{\sqrt{n}}\tilde{\theta}_n(0) \leqslant \frac{1}{n}\sqrt{n}|\tilde{\theta}_n(0) - \hat{\theta}_n| + \frac{1}{\sqrt{n}}\hat{\theta}_n, \tag{30}$$

for n sufficiently large. Since $\hat{\theta}_n/\sqrt{n} \to 0$ for almost every y-sequence and $1/n$ can be made arbitrarily small, tightness follows from Proposition 3.

To find the asymptotic distribution of $\tilde{\theta}_n$, note that the complete data information, $V(\theta_0)$, is $1/\theta_0^2$ and that $F(\theta_0)^{\mathrm{T}} = \mathrm{E}_{\theta_0}(D_\theta M(\theta_0)) = \exp(-c/\theta_0)$. Theorem 2 and straightforward calculations show that the asymptotic variance of $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ is

$$\frac{\theta_0^2}{1 - \exp(-c/\theta_0)} \cdot \left(2 - \frac{1}{1 + \exp(-c/\theta_0)}\right), \tag{31}$$

where the first factor is the asymptotic variance of the observed data MLE.

Remark. Ergodicity can be shown quite easily in this example for all values of n with $N > 1$ by noting that

$$\tilde{\mathrm{E}}(\tilde{\theta}_n(1)|\tilde{\theta}_n(0)) = \frac{1}{n}\sum_{i=1}^{n} X_i \wedge c + \frac{n-N}{n}\tilde{\theta}_n(0) \tag{32}$$

and invoking Theorem 14.0.1 in Meyn and Tweedie (1993).

4. Estimation

We will now discuss various ways of using the StEM algorithm for estimating the unknown parameter θ. Theorem 2 immediately yields a consistent asymptotically normal estimator of θ, namely $\tilde{\theta}_n$, with asymptotic variance $I(\theta_0)^{-1}[2I - \{I + F(\theta_0)\}^{-1}]$. The asymptotic variance can be split into a 'model part', $I(\theta_0)^{-1}$, and a 'simulation part', $I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]$. The latter is the asymptotic variance of $\tilde{\theta}_n$ given the observed data, i.e. the additional variance due to the simulations. The following result bounds this additional variance.

Proposition 4. *The asymptotic relative efficiency of $u^{\mathrm{T}}\tilde{\theta}_n$ to $u^{\mathrm{T}}\hat{\theta}_n$, for any $u \in \mathbb{R}^d$, is bounded by $2 - 1/(1+\lambda)$, where λ is the largest eigenvalue of the fraction of missing information, $F(\theta_0)$. In particular, the simulation increase the variance by less than 50% compared to the observed data MLE.*

Proof. Let $I(\theta_0)^{1/2}$ the positive definite square root of the positive definite matrix $I(\theta_0)$ and let $I(\theta_0)^{-1/2}$ denote its inverse.

The asymptotic relative efficiency of $u^{\mathrm{T}}\tilde{\theta}_n$ to $u^{\mathrm{T}}\hat{\theta}_n$ is

$$\frac{u^{\mathrm{T}} I(\theta_0)^{-1}[2I - \{I + F(\theta_0)\}^{-1}]u}{u^{\mathrm{T}} I(\theta_0)^{-1} u} = \frac{v^{\mathrm{T}} I(\theta_0)^{-1/2}[2I - \{I + F(\theta_0)\}^{-1}] I(\theta_0)^{1/2} v}{v^{\mathrm{T}} v} \leqslant \lambda_1, \quad (33)$$

where $v = I(\theta_0)^{-1/2}u$ and λ_1 is the largest eigenvalue of $I(\theta_0)^{-1/2}[2I - \{I + F(\theta_0)\}^{-1}]I(\theta_0)^{1/2}$. This matrix has the same eigenvalues as $[2I - \{I + F(\theta_0)\}^{-1}]$ and by straightforward manipulations we see that $\lambda_1 = 2 - 1/(1+\lambda)$. Since $\lambda < 1$ we get that $\lambda_1 < 3/2$. □

We see that the additional variance due to the simulations is an increasing function of the fraction of missing information. Thus the increase in variance can to some extent be controlled by choosing the complete data model so that the fraction of missing information is small.

Obviously, it will be of interest to reduce the simulation part of the variance, i.e. to reduce the additional variance due to simulations. The first thing that comes to mind is to average (the last part of) the Markov chain, $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$, i.e. to use $\tilde{\mathrm{E}}(\tilde{\theta}_n)$ as an estimator of θ. However, more than ergodicity of the Markov chain is needed to ensure that this mean exists and hence that the average of the chain converges to $\tilde{\mathrm{E}}(\tilde{\theta}_n)$. Indeed, in most examples it is not difficult to construct parameters where the mean of $\tilde{\theta}_n$ does not exist. For instance, in the example discussed in Section 3.3, if we reparametrize to (the hardly relevant parameter) $\theta^* = 1/(\theta - c)$, then the Markov chain derived from the StEM algorithm will still be ergodic and tight, but the mean of the stationary distribution does not exist. Furthermore, more than ergodicity is needed to ensure that $\tilde{\mathrm{E}}(\tilde{\theta}_n)$ is a consistent estimator

of θ, if Θ is not bounded. To get $\sqrt{n}$-consistency, even more is needed. Finally, in order to estimate the mean of $\tilde{\theta}_n$ we need to run the Markov chain for more iterations than are needed to (approximately) obtain a realization of $\tilde{\theta}_n$. As we consider large-sample properties of the StEM algorithm and as the simulation burden increases with the sample size, we shall here only discuss some finite simulation estimators of θ.

4.1. Averaging the Markov chain

Even with only finite simulation, averaging the last m iterations of the Markov chain, $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$, improves the estimation.

Choose a sequence of integers, $(k_n)_{n\in\mathbb{N}}$, such that the total variation distance between $\tilde{\theta}_n(k_n)$ and $\tilde{\theta}_n$ is smaller than, say, $1/n$. This is possible due to the ergodicity of the Markov chain. Then, conditionally on y,

$$\sqrt{n}(\tilde{\theta}_n(k_n) - \hat{\theta}_n) \xrightarrow{\mathscr{D}} N(0,\, I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]) \tag{34}$$

We can now show the following result:

Proposition 5. *Let m be a fixed integer and assume ergodicity and tightness as in Theorem 2.*

(i) $(\sqrt{n}(\tilde{\theta}_n(k_n + j) - \hat{\theta}_n))_{j=0,\ldots,m-1}$ *converges in distribution given y for almost every y-sequence as $n \to \infty$ to a sequence of the same length of the stationary Gaussian AR(1) process with autoregression parameter $F(\theta_0)^{\mathrm{T}}$ and innovation variance $V(\theta_0)^{-1}\mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}$.*

(ii) $\sqrt{n}(\frac{1}{m}\sum_{j=0}^{m-1}\tilde{\theta}_n(k_n + j) - \theta_0)$ *converges in distribution as $n \to \infty$ to the normal distribution with mean 0 and variance given by*

$$\begin{aligned} &I(\theta_0)^{-1} + \frac{1}{m} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}] \\ &\quad + \frac{2}{m} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]F(\theta_0)(I - F(\theta_0))^{-1} \\ &\quad - \frac{2}{m^2} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]F(\theta_0)(I - F(\theta_0)^m)(I - F(\theta_0))^{-2}. \end{aligned} \tag{35}$$

Proof. The first statement of the proposition is easily shown by induction on m. The induction start is given by (34), and the induction step uses the same arguments as the proof of Proposition 1 with μ_n replaced by the distribution of $(\sqrt{n}(\tilde{\theta}_n(k_n + j) - \hat{\theta}_n))_{j=0,\ldots,k-1}$ and μ by the corresonding weak limit, which exists by the induction hypothesis.

From part (i) it follows that, conditionally on y, $\sqrt{n}((1/m)\sum_{j=0}^{m-1}\tilde{\theta}_n(k_n + j) - \hat{\theta}_n)$ converges in distribution as $n \to \infty$ to the normal distribution with mean 0 and variance given by

$$\frac{1}{m} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}] + \frac{1}{m^2} \sum_{1 \leqslant i < j \leqslant m} [(F(\theta_0)^{\mathrm{T}})^{j-i} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]$$

$$+ I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]F(\theta_0)^{j-i}]. \tag{36}$$

Since $(F(\theta_0)^{\mathrm{T}})^{j-i} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}] = I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]F(\theta_0)^{j-i}$, the variance simplifies to

$$\frac{1}{m} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]$$

$$+ \frac{2}{m} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]F(\theta_0)(I - F(\theta_0))^{-1}$$

$$- \frac{2}{m^2} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]F(\theta_0)(I - F(\theta_0)^m)(I - F(\theta_0))^{-2}. \tag{37}$$

The result now follows from Lemma 1 in Schenker and Welsh (1987). □

Notice that the mean of $\tilde{\theta}_n$ does not have to exist for this result to be valid, since it is a *fixed* m result.

4.2. More simulations per iteration

In this subsection we shall discuss three different estimators derived from simple extensions of the StEM algorithm. All these estimators are obtained by increasing the amount of simulation in each iteration by a factor m, and, though different, they turn out to be asymptotically equivalent. Just like the simple estimator, $\tilde{\theta}_n$, they can be further improved by averaging as discussed in the previous subsection, but for simplicity of exposition we shall avoid discussion of this in the following paragraphs.

4.2.1. Averaging log-likelihoods

The StE-step estimates the expectation from the E-step in the EM algorithm by (an average of) one simulated value. Thus the StE-step can be thought of as a very poor Monte Carlo integration, and the StEM algorithm may be seen as a simple version of the MCEM algorithm (Wei and Tanner 1990).

The resemblance can be strengthened by improving the Monte Carlo approximation, i.e. by simulating, for each observation y_i, m independent (given y_i) values, $\tilde{X}_{i,1} \ldots, \tilde{X}_{i,m}$, say, of the missing data, X_i, and maximizing the average, $(1/m)\sum_{j=1}^{m}\sum_{i=1}^{n} \log f_\theta(\tilde{X}_{i,j})$, of the m complete data log-likelihoods. This is again a complete data log-likelihood, namely the log-likelihood obtained when the $\tilde{X}_{i,j}$s are treated as if they were i.i.d. Obviously, $\tilde{X}_{i,j}$, $j = 1, \ldots, m$, are not (unconditionally) independent. The complete data log-likelihood allows us to use standard methods, i.e. the methods we would have used had we had complete data, in order to maximize the likelihood. The dependence in the simulated $\tilde{X}_{i,j}$s

may in some cases lead to multimodality of the likelihood function and thus complicate the M-step.

This algorithm again leads to a Markov chain of θ values, which is typically ergodic. Let $\tilde{\theta}_n^{\mathrm{MC}}$ denote a random variable drawn from the stationary distribution of this chain. We refer to this random variable as the Monte Carlo estimator.

To find the asymptotic distribution of $\tilde{\theta}_n^{\mathrm{MC}}$ we proceed as in Section 3.2. The multiple simulations affect Lemma 3 in Section 3.2; in the asymptotic distribution of the transition probabilities (16) the mean is unchanged but the variance is replaced by $\frac{1}{m} V(\theta_0)^{-1} \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}$. A detailed proof of this will not be given here; it closely mimics the proof of Lemma 3. The consequence of the changes to Lemma 3 is that the limiting AR(1) process will have an innovation variance, which is a factor $1/m$ smaller than the innovation variance in Theorem 2(i), but the same autoregression parameter $F(\theta_0)^{\mathrm{T}}$.

Hence, unconditionally,

$$\sqrt{n}(\tilde{\theta}_n^{\mathrm{MC}} - \theta_0) \xrightarrow{\mathscr{D}} N\left(0,\ I(\theta_0)^{-1} + \frac{1}{m} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]\right). \tag{38}$$

4.2.2. Multiple maximizations

In Section 4.2.1 we simulated m $\tilde{X}_n$s for each X_i and used these to improve the estimate of the conditional expectation required in the E-step of the EM algorithm. We could instead use the multiple simulations to construct m (pseudo-)complete data sets and then maximize these separately. The m maximizers found by maximizing each of the m complete data log-likelihoods separately could then be averaged to give the next value of $\tilde{\theta}_n(k)$. This is again a Markov chain. Obviously, if the complete data MLE is linear in the data, then this approach leads to the exact same result as the simultaneous maximization approach discussed in the previous paragraph. It is therefore not surprising that this algorithm leads to the same asymptotic result as above. In order to prove this claim, we generalize Lemma 3.

Lemma 4. *For* $j = 1, \ldots, m$, *let* $\tilde{X}_{i,j} \sim \mathscr{L}_{\theta_n}(X|Y = y_i)$, *where* $\theta_n = \hat{\theta}_n + (1/\sqrt{n})h + o(1/\sqrt{n}) = \tilde{\theta}_n^{\mathrm{MM}}(0)$. *Let* $\tilde{\theta}_{n1}, \ldots, \tilde{\theta}_{nm}$ *be the* m *complete data estimators calculated from the* m *simulated data sets. For almost all y-sequences and conditional on y,*

$$\sqrt{n}(\tilde{\theta}_n^{\mathrm{MM}}(1) - \hat{\theta}_n) \xrightarrow{\mathscr{D}} \left(F(\theta_0)^{\mathrm{T}} h,\ \frac{1}{m} V(\theta_0)^{-1} \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}\right), \tag{39}$$

where $\tilde{\theta}_n(1)^{\mathrm{MM}}$ *is the average of* $\tilde{\theta}_{n1}, \ldots, \tilde{\theta}_{nm}$.

Proof. Applying Lemma 3 to each of the m complete data MLEs $\tilde{\theta}_{n1}, \ldots, \tilde{\theta}_{nm}$ leads to $\sqrt{n}(\tilde{\theta}_{nj} - \hat{\theta}_n) \xrightarrow{\mathscr{D}} N\big(F(\theta_0)^{\mathrm{T}} h,\ V(\theta_0)^{-1} \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}\big)$, for $j = 1, \ldots, m$. Since the simulations are independent (given the observed data) the average has the asymptotic distribution (39). □

Having found the asymptotic distribution of the transition probabilities of this Markov chain, we can show (assuming convergence and tightness as in Theorem 2 and mimicking

the proof) that the stationary distribution of the Markov chain is asymptotically normal. Thus the estimator derived from this algorithm has the unconditional asymptotic distribution

$$\sqrt{n}(\tilde{\theta}_n^{\mathrm{MM}} - \theta_0) \xrightarrow{\mathscr{D}} \mathrm{N}\left(0,\, I(\theta_0)^{-1} + \frac{1}{m} I(\theta_0)^{-1}[I - \{I + F(\theta_0)\}^{-1}]\right), \tag{40}$$

just like the Monte Carlo estimator (38). We refer to $\tilde{\theta}_n^{\mathrm{MM}}$ as the multiple maximization estimator.

The advantage of the m-maximization approach compared to the simultaneous maximization approach is that the averaging of the log-likelihoods which may introduce multi-modality is avoided. This may in some cases make the m M-steps faster than the more complicated M-step needed in the Monte Carlo version.

4.2.3. Comparison

The two estimators discussed in the previous paragraphs are obviously asymptotically equivalent in the sense that they have the same asymptotic distribution. They are also asymptotically equivalent to the estimator obtained if we run the StEM algorithm m times in parallel and then average the m simple estimators obtained by just taking the last iteration of each of the m algorithm. This estimator can be seen as a natural step further in the direction of doing things in parallel. In Section 4.2.1 m simulations are done in parallel, averaged and maximized. In Section 4.2.2 both the simulations and the maximizations are done in parallel, and the average then taken. With multiple chains we do 'everything' including convergence in parallel before we average. The term 'parallel' is meant to be taken conceptually; real parallel computing is obviously not needed.

We mention the multiple chains estimator here for two reasons. First, running Markov chain algorithms from several (over-dispersed) starting points has been suggested as a way of checking convergence (cf. Gelman and Rubin 1992); we shall return briefly to this in Section 5.1. Thus it is a 'real' estimator in the sense that it occurs in practice; we refer to it as the multiple chains estimator. Second, one might expect the asymptotics to work slightly better for this estimator than for the the other two estimators since we average independent estimators; with multiple maximizations we average dependent estimators and with simultaneous maximizations we 'average the data' and then maximize. Hence in the simulation results we will discuss in the next subsection the multiple chains estimator will serve as a 'gold standard'.

We note that the time it takes to complete one iteration of one of these three algorithms differs from algorithm to algorithm. In every one of them we have m StE-steps. When using multiple chains there is only one averaging rather than one per iteration, but we would not expect this to make a significant difference in run-times. However, in the multiple maximization and the multiple chains algorithms we also have m M-steps rather than just one. Thus one might think that the Monte Carlo version is faster than the other two. However, as indicated previously, the maximization of the averaged log-likelihood may be more complicated and thus more time-consuming than m simpler maximizations. Thus it is far from obvious which algorithm has the fastest iterations. Nor is it obvious which algorithm requires less iterations to converge. Both questions are of interest since it is the

time to convergence measured in CPU time, rather than the number of iterations, that is of interest when deciding which algorithm to use.

When comparing the estimators given in this subsection and the averaged estimators discussed in Section 4.1, we note that the reduction in the variance for the same choice of m is always larger for the estimators discussed in this subsection. Furthermore, the difference between the variances of the estimators discussed here and the averaged estimator discussed in Section 4.1 is of the same magnitude, $1/m$, as the additional variance due to the simulations. This means that the relative efficiency of a multiple simulation estimator compared to an averaged estimator conditional on the observed data does not go to 1 even if m increases. In fact, at least for a one-dimensional parameter, this relative efficiency moves further away from 1 as m increases.

The gain of averaging the Markov chain depends greatly on the fraction of missing information. If the fraction of missing information is large then the gain is small, but for small fractions of missing information the gain can be quite large. Figure 1 shows the asymptotic relative efficiency of the estimators compared to the MLE for different values of m as a function of the fraction of missing information in the one-dimensional case.

Note, however, that averaging of the last iterations of the Markov chain is possible also

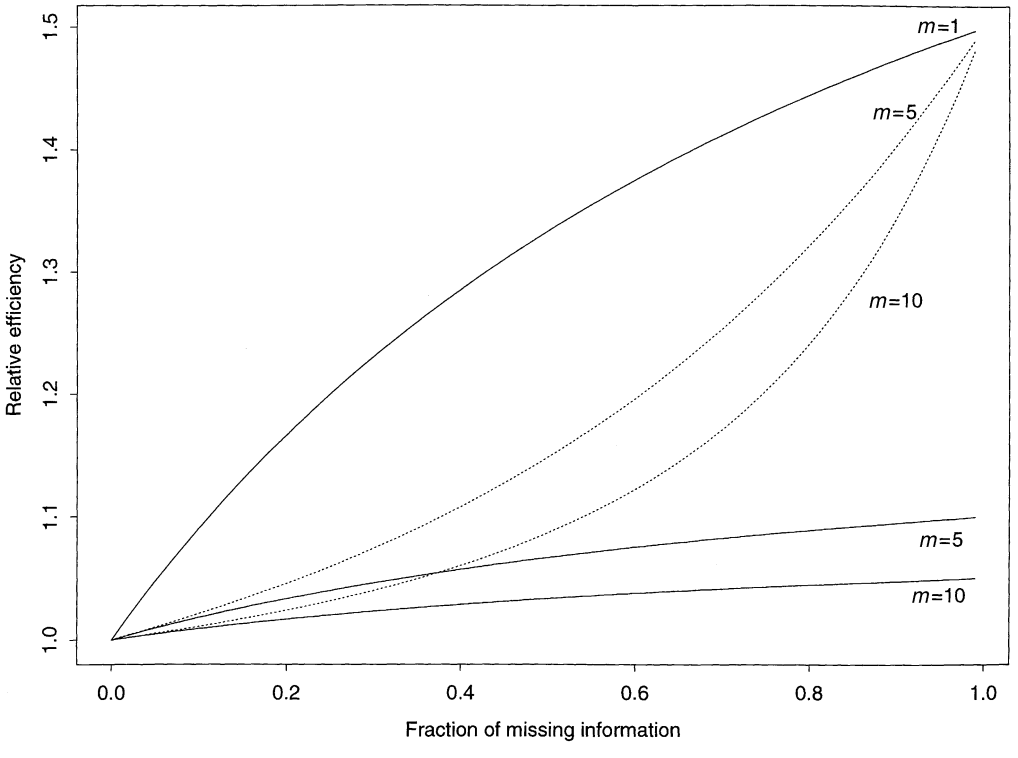

Figure 1. Relative efficiency of StEM estimators: full lines, multiple simulation estimators; broken lines, averaged single simulation estimators.

for the multiple simulations versions, so that there is not really a choice of either/or; we can always average the last iterations of the Markov chain. Also the other methods may be mixed; for instance, we could run m_1 chains of the multiple maximizations algorithm with m_2 simulations per X_i leading to an estimator with variance as in (38) or (40) with $m = m_1 m_2$. This variance can then be brought further down by averaging.

4.3. Simulation experiment

In order to illustrate the behaviour of the various estimators in small and moderate samples, we report some findings from a simulation experiment. A simple model has been chosen so that the results found should be an effect of the methods rather than an effect of a complex model.

The simulated data are a random sample from a standard exponential distribution. The incomplete data are obtained by censoring the simulated data at a fixed point. Two different sample sizes, ($n = 50$ and $n = 500$) and two different points of censoring (one corresponding to $F(\theta_0) = \frac{1}{4}$ and one to $F(\theta_0) = \frac{1}{2}$) have been used. Thus we have a small and a moderate sample size and a moderate and a large fraction of missing information. Three different choices of m (1, 5 and 10) have been used. We estimate the intensity rather than the scale parameter in order to obtain a complete data MLE that is nonlinear in the data; in this way the asymptotically equivalent estimators are only asymptotically equivalent, not actually identical. The algorithms have been run for a initial burn-in of 5000 iterations and then an additional 1000 iterations have been used to estimate the distribution of the estimators. Various convergence diagnostics (cf. Section 5.1) suggest that this burn-in is sufficient.

We summarize the distributions of the estimators in terms of the biases, i.e. the empirical means of the simulated estimators minus the observed data MLE, the standard deviations and the relative efficiencies of the estimators compared to the asymptotic distribution of the observed data MLE. The unconditional distribution of the StEM estimator is made up of the conditional distribution plus the distribution of the MLE. The latter is not of interest when evaluating the effects of the simulations since it is purely a function of the observed data. Since we are interested in comparing the various simulation estimators rather than comparing the simulation estimators to the MLE, it is the conditional distribution that is of interest. Therefore we report conditional biases and standard deviations rather than unconditional ones. Furthermore, the differences between unconditional biases and standard deviations would vanish compared to the biases and standard deviations of the MLE, in particular for $m = 10$. The relative efficiencies are calculated unconditionally using the asymptotic variance of the observed data MLE as comparison. We include these in order to show how little the simulation noise matters in practice.

4.3.1. Averaging

We start by giving a few results for the average of the last m iterations of the Markov chain from the StEM algorithm (Tables 1 and 2). We see, not surprisingly, that the variance goes

Table 1. Simulation results for averaged StEM estimator minus the MLE, $F(\theta_0) = 0.25$

$\bar{\theta}_n$	$n = 50$			$n = 500$			Asymp
$F(\theta_0) = 0.25$	Bias	StDev	RelEff	Bias	StDev	RelEff	RelEff
$m = 1$	0.006 70	0.075 295	1.2126	0.001 35	0.021 947	1.1806	1.2000
$m = 5$	0.004 13	0.040 840	1.0625	0.001 24	0.012 200	1.0558	1.0596
$m = 10$	0.004 83	0.029 534	0.0327	0.001 30	0.008 529	1.0273	1.0316

Notes: StDev is the standard deviation, RelEff the efficiency relative to the MLE, Asymp RelEff the asymptotic relative efficiency compared to the MLE.

Table 2. Simulation results for averaged StEM estimator minus the MLE, $F(\theta_0) = 0.50$

$\bar{\theta}_n$	$n = 50$			$n = 500$			Asymp
$F(\theta_0) = 0.25$	Bias	StDev	RelEff	Bias	StDev	RelEff	RelEff
$m = 1$	0.016 18	0.131 162	1.4301	0.017 23	0.038 064	1.3622	1.3333
$m = 5$	0.010 77	0.079 513	1.1581	0.016 05	0.022 609	1.1278	1.1483
$m = 10$	0.014 74	0.058 616	1.0859	0.016 01	0.016 356	1.0669	1.0867

Notes: see Table 1.

down as m increases. The relative efficiencies are not too far away from the corresponding asymptotic expressions given in the tables. We note that the estimators are biased. This is due to the chosen parametrization: the complete data MLE of the intensity has an expected bias of $1/(\mathrm{n}-1)$, and we would not expect the StEM estimator to do better. Had we chosen to estimate the scale parameter (as in Section 3.3) instead, the conditional mean of the StEM estimator would equal the observed data MLE. Using the readily available expressions for the conditional mean and variance of the StEM estimator of the scale parameter, a Taylor expansion suggests that the bias in the StEM estimator for the intensity is $\hat{\theta}_n/(1+2N)$ plus terms of lower order, where $\hat{\theta}_n$ is the observed data MLE of the intensity parameter, and N is the number of uncensored observations. In the small-sample cases this gives 0.0151 and 0.0215 for the moderate and large fraction of missing information respectively; in the large-sample cases we obtain 0.001 33 and 0.001 98 respectively. We see that the bias in the simulations is smaller than that obtained from the Taylor expansion, apart from the case of large-sample, moderate fraction of missing information, where the bias in simulations is close to what we would expect. This suggests that the discarded terms in the Taylor expansion are still fairly large in the small-sample cases and – to some degree – also in the large fraction of missing information case, where N is small compared to the sample size. Since the expression for the bias obtained from the Taylor expansion is asymptotic, it is not surprising that the biases in the large-sample cases are closer to the values given by the Taylor expansion.

4.3.2. Multiple simulations

Tables 3–6 give simulation results for the multiple simulations estimators. Each table gives results for all three estimators for a fixed sample size, a fixed fraction of missing information, and for both $m = 5$ and $m = 10$. The case $m = 1$ can be seen in Tables 1 and 2; all estimators are the same when $m = 1$, so we do not repeat these results.

Looking first at the results for the low fraction of missing information (Tables 3 and 4), we see that the standard deviations are virtually identical. The relative efficiencies are close to the expected 1.04 ($m = 5$) and 1.02 ($m = 10$). Again the estimators are biased. The bias of the Monte Carlo version is smallest, and this is to be expected. If we Taylor expand as in the previous subsection, we see that the bias is $\hat{\theta}_n/(1 + 2mN)$ (discarding terms of higher order) for the Monte Carlo estimator, whereas the biases of the other two estimators are unaffected by m. Thus the bias in these two cases is expected to be 0.0151 and 0.001 33 (in the small and moderate sample size cases respectively) as in the previous subsection, whereas for the Monte Carlo estimator we would expect 0.003 06 and 0.001 53 in the small samples for $m = 5$ and $m = 10$ respectively, and 0.000 23 and 0.000 13 when $n = 500$. The simulated bias again fits poorly to the asymptotic expression except in the Monte Carlo case, but we do find that the bias is unaffected by m in the multiple chains and multiple maximization estimators, but considerably lower and decreasing with m in the Monte Carlo estimators, though in the $n = 500$ case the biases are so small that they seem to disappear in the simulation noise. Since the expression we have derived for the bias is a large-sample expression – we discard higher-order terms in a Taylor expansion – it is not worrying that

Table 3. Simulation results for StEM estimators minus the MLE, $n = 50$, $F(\theta_0) = 0.25$

$n = 50$	$m = 5$			$m = 10$		
$F(\theta_0) = 0.25$	Bias	StDev	RelEff	Bias	StDev	RelEff
$\bar{\theta}$	0.004 13	0.033 582	1.0423	0.004 83	0.023 633	1.0209
$\tilde{\theta}^{\mathrm{MC}}$	0.003 02	0.032 767	1.0403	0.001 52	0.024 383	1.0223
$\tilde{\theta}^{\mathrm{MM}}$	0.005 21	0.033 508	1.0421	0.005 52	0.024 579	10.227

Notes: $\bar{\theta}$ is the average of m chains, $\tilde{\theta}^{\mathrm{MC}}$ based on m simulations, $\tilde{\theta}^{\mathrm{MM}}$ based on m maximizations. StDev is the standard deviation, RelEff is the efficiency relative to the MLE.

Table 4. Simulation results for StEM estimators minus the MLE, $n = 500$, $F(\theta_0) = 0.25$

$n = 500$	$m = 5$			$m = 10$		
$F(\theta_0) = 0.25$	Bias	StDev	RelEff	Bias	StDev	RelEff
$\bar{\theta}$	0.000 74	0.010 095	1.0382	0.000 49	0.006 440	1.0156
$\tilde{\theta}^{\mathrm{MC}}$	0.000 12	0.009 885	1.0366	0.000 26	0.006 941	1.0181
$\tilde{\theta}^{\mathrm{MM}}$	0.001 02	0.010 381	1.0404	0.001 08	0.007 215	1.0195

Notes: see Table 3.

the simulated biases differs from the 'expected'. The better agreement in the Monte Carlo is probably due to the discarded terms decreasing in m. It is interesting that the biases behave as we would expect: the Monte Carlo implementation of the StEM algorithm is closer to the EM algorithm, which returns the MLE. Therefore we should expect the bias to be smaller for the Monte Carlo estimator.

The largest differences in the simulation results are found when $m = 10$. The case of small sample size and large m reflects the lower expected bias in the Monte Carlo estimator. In all cases the multiple maximizations estimator has a larger bias than the other two estimators; in the $n = 500$, $m = 10$ case the difference is quite large. Incidentally, in this case there is a better agreement with the asymptotic bias and variance for the multiple maximizations estimator than for the two other estimators.

Tables 5 and 6 give results for the large fraction of missing information case. Here differences are more pronounced. The expected relative efficiencies are 1.0667 when $m = 5$, and 1.0333 when $m = 10$. The multiple chains estimator is fairly close but the relative efficiencies of the other estimators are a lot smaller. Obviously, this is also seen in the standard deviations of the Monte Carlo and the multiple maximization estimators, which are smaller than those of the multiple chains estimators. Again the tendencies in the biases are as before; roughly unaffected by m in the multiple chains and multiple maximizations estimators deviations, and generally smaller and decreasing with m for the Monte Carlo estimator. As in the low fraction of missing information cases, the simulated biases fit poorly to the approximation; the expected biases for the Monte Carlo estimators are 0.004 36 and 0.002 18 ($m = 5$, 10 respectively) in the small-sample cases, and 0.003 97 and 0.000 20 in the $n = 500$ case. For the other two estimators we obtain 0.0215 and 0.001 98

Table 5. Simulation results for StEM estimators minus the MLE, $n = 50$, $F(\theta_0) = 0.50$

$n = 50$	$m = 5$			$m = 10$		
$F(\theta_0) = 0.50$	Bias	StDev	RelEff	Bias	StDev	RelEff
$\bar{\theta}$	−0.021 09	0.053 748	1.0722	−0.017 11	0.039 386	1.0388
$\tilde{\theta}^{\mathrm{MC}}$	0.000 66	0.034 797	1.0303	0.001 72	0.023 999	1.0144
$\tilde{\theta}^{\mathrm{MM}}$	0.004 69	0.033 212	1.0276	0.004 67	0.022 719	1.0129

Notes: see Table 3.

Table 6. Simulation results for StEM estimators minus the MLE, $n = 500$, $F(\theta_0) = 0.50$

$n = 500$	$m = 5$			$m = 10$		
$F(\theta_0) = 0.50$	Bias	StDev	RelEff	Bias	StDev	RelEff
$\bar{\theta}$	0.018 11	0.016 860	1.0711	0.017 95	0.011 824	1.0524
$\tilde{\theta}^{\mathrm{MC}}$	0.000 73	0.010 099	1.0255	0.000 05	0.007 133	1.0191
$\tilde{\theta}^{\mathrm{MM}}$	−0.000 35	0.010 448	1.0273	0.000 66	0.007 613	1.0217

Notes: see Table 3.

($n = 50$, 500 respectively). We note that the larger the fraction of missing information, the larger the sample size needed to obtain the asymptotic results.

Inspection of various quantile–quantile plots as well as Kolmogorov–Smirnov test statistics (not shown here) suggests that in the low fraction of missing information case almost all the triples of estimators with the same values of n and m are similarly distributed. The only exceptions are the Monte Carlo estimator in the $n = 50$, $m = 10$ case which has a smaller bias than the other two, and the multiple maximizations estimator in the $n = 500$, $m = 10$ case, which has a considerably larger bias. In both cases the differences appear to be mainly a question of bias; if we subtract the bias, there appear to be no significant differences.

For the large fraction of missing information none of the estimators have similar distributions, the sole exceptions being the multiple maximization estimators and the Monte Carlo estimators when $m = 5$. The explanation of these similarities appears to be the roughly equal variances and the neglible biases.

4.4. Estimation of the asymptotic variance

The asymptotic variance of the various estimators can be estimated consistently from consistent estimates of any two of the four quantities $V(\theta_0)$, $I(\theta_0)$, $\mathrm{E}_{\theta_0} I_Y(\theta_0)$, and $F(\theta_0)$. The estimators discussed in this subsection can of course be applied to any of the algorithms, but for notational simplicity we only give formulae for the simple StEM algorithm.

If the complete data information is continuous, then $V(\tilde{\theta}_n)$ is consistent for $V(\theta_0)$, and by Assumption U $(1/n)\sum_{i=1}^{n} I_{Y_i}(\tilde{\theta}_n)$ is consistent for $\mathrm{E}_{\theta_0} I_Y(\theta_0)$.

With further assumptions, $(1/n)\sum_{i=1}^{n} s_{\tilde{X}_i|Y_i}(\tilde{\theta}_n)^{\otimes 2}$ may be a consistent estimator of $\mathrm{E}_{\theta_0} I_Y(\theta)$, when $\tilde{X}_n \sim \mathscr{L}_{\tilde{\theta}_n}(X|Y = y_i)$. This will be the case if

$$\left| \frac{1}{n}\sum_{i=1}^{n} s_{\tilde{X}_i|y_i}(\theta)^{\otimes 2} - \frac{1}{n}\sum_{i=1}^{n} I_{y_i}(\theta) \right| \xrightarrow{\tilde{P}_\theta} 0 \tag{41}$$

uniformly in a neighbourhood of θ_0 for almost every y-sequence.

These two estimators of $\mathrm{E}_{\theta_0} I_Y(\theta)$ may in practice be difficult to obtain due to the need for expressions for either $I_{y_i}(\theta)$ or $s_{x|y_i}(\theta)$. Following Louis (1982), we note that

$$\frac{1}{n}\sum_{i=1}^{n} D_\theta s_{\tilde{X}_i}(\tilde{\theta}_n) + \frac{1}{n}\sum_{i=1}^{n} s_{\tilde{X}_i}(\tilde{\theta}_n)^{\otimes 2} - \left(\frac{1}{n}\sum_{i=1}^{n} s_{\tilde{X}_i}(\tilde{\theta}_n)\right)^{\otimes 2} \tag{42}$$

is a consistent estimator of $I(\theta_0)$ assuming sufficient smoothness.

Finally, we note that $V(\theta_0)^{-1}\mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}$ and $F(\theta_0)$ can be estimated from the Markov chain, $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$. Let k_n be chosen as in Section 4.1 and put $\bar{\theta}_n = (1/m)\sum_{i=1}^{m} \tilde{\theta}_n(k_n + j)$. Then (using Proposition 5(i))

$$\tilde{F}_m = \left(\sum_{j=1}^{m-1} (\tilde{\theta}_n(k_n + j) - \bar{\theta}_n)^{\otimes 2} \right)^{-1} \cdot \sum_{j=1}^{m-1} (\tilde{\theta}_n(k_n + j + 1) - \bar{\theta}_n)(\tilde{\theta}_n(k_n + j) - \bar{\theta}_n)^{\mathrm{T}}$$

$$\xrightarrow{\mathscr{D}} \left(\sum_{j=1}^{m-1} (Z_j - \bar{Z})^{\otimes 2} \right)^{-1} \cdot \sum_{j=1}^{m-1} (Z_{j+1} - \bar{Z})(Z_j - \bar{Z})^{\mathrm{T}}, \tag{43}$$

where $Z_1, Z_2, \ldots, Z_m$ are distributed according to the limiting Gaussian AR(1) process, as $n \to \infty$. Note that since m is kept fixed, this estimator of $F(\theta_0)$ is not consistent.

Similarly, $V(\theta_0)^{-1} \mathrm{E}_{\theta_0} I_Y(\theta_0) V(\theta_0)^{-1}$ may be estimated (inconsistently for fixed m) from

$$\frac{1}{m} \sum_{j=1}^{m} (\tilde{\theta}_n(k_n + j) - \bar{\theta}_n)^{\otimes 2} \xrightarrow{\mathscr{D}} \frac{1}{m} \sum_{j=1}^{m} (Z_j - \bar{Z})^{\otimes 2}. \tag{44}$$

Due to the (asymptotically) positive autocorrelation this estimator will tend to underestimate the innovation variance.

Without further assumptions, these two estimators may even be inconsistent as $m \to \infty$. However, their large-sample distribution may be simulated in practice. The convergence in (43) may be strengthened to convergence in pth mean if we sum to $j = m$ (rather than $j = m - 1$) in the denominator of $\tilde{F}_m$ as in (44).

5. Concluding remarks

5.1. Implementation issues

As we have seen in Section 4.2, the number of simulations, m, determines the variance of the resulting estimator. By a straightforward extension of Proposition 4 the loss in efficiency is bounded by $(1 - 1/(1 + \lambda))/m \leqslant 1/(2m)$, where λ is the largest eigenvalue of the fraction of missing information. By specifying how small an efficiency loss, δ, due to simulations, we will accept, an appropriate value of m can be chosen; $m \geqslant 1/(2\delta)$ will do. A closer bound can be obtained if an upper bound, λ^*, on the largest eigenvalue of the fraction of missing information is available. Then choosing $m \geqslant 1/(\delta(1 - 1/(1 + \lambda^*))$ will ensure that the loss in efficiency is bounded by δ. In many cases the expected fraction of incomplete observations will be an upper bound on λ. Typically, λ^* must be estimated.

The value of m also affects the behaviour of the Markov chains. In nice exponential families where $\tilde{\mathrm{E}}(\tilde{\theta}_n(k) | \tilde{\theta}_n(\mathrm{k} - 1)) = M(\tilde{\theta}_n(\mathrm{k} - 1))$ we can write the multiple simulations versions of the StEM algorithm (discussed in Sections 4.2.1 and 4.2.2) as

$$\tilde{\theta}_n(k) = M(\tilde{\theta}_n(k - 1)) + \varepsilon_k, \tag{45}$$

where ε_k is approximately Gaussian for large values of n with mean zero and variance inversely proportional to m. So we can think of the StEM algorithm as made up from a drift part, which is just the EM update, M, and a noise part, ε_k. Suppose that θ'_n is a fixed point of M, i.e. that $M(\theta'_n) = \theta'_n$, and that for any $\theta \neq \theta'_n$ in an open ball B around θ'_n,

$\|M(\theta) - \theta'_n\| < \|\theta - \theta'_n\|$, so that the EM algorithm started inside this ball will converge to θ'_n. Then the StEM algorithm will tend to stay for some time in B if m is large. This is the case since the noise term will be so small that the probability of $\tilde{\theta}_n(k)$ not being in B, given that $\tilde{\theta}_n(k-1) \in B$, will be small. Thus large values of m will make the StEM algorithm move more slowly and possibly stay in neighbourhoods of fixed points for long periods of time. Consequently, if many fixed points are present, we expect slow convergence of the StEM algorithm for large values of m. On the other hand, if no such fixed points exist the small noise obtained for large values of m will make StEM move rapidly towards the MLE.

One also has to bear in mind that $m > 1$ leads to m M-steps or a possibly more complicated M-step as mentioned in Section 4.2 and obviously results in a slower StE-step (since m times as many simulations are necessary). However, as noted in Section 4.3 the Monte Carlo version may result in an estimator with a smaller conditional bias due to being closer to the EM algorithm. We stress that the bias is a maximum likelihood problem, rather than a StEM-problem; it is due to the MLE being biased for some choices of parametrization, and this is inherited by the estimators derived from the StEM algorithm.

If a Markov chain simulation scheme, such as a Gibbs sampler, is necessary to perform the StE-step, then using m fairly large will typically be a good idea. After burn-in of the Gibbs sampler it will be relatively cheap to obtain multiple simulations compared to the time already spent. The results of Section 4.2 can easily be extended to the case where $\tilde{X}_{i,j}$, $j = 1, \ldots, m$, are not independent given y but only stationary. As in Section 4.2 it is the innovation variance of the Gaussian AR(1) process that is affected and not the parameter $F(\theta_0)$. The new innovation variance will be similar to the variance (36); see Chan and Ledolter (1995) for further details.

If the drift is a problem, the multiple chains approach may be useful. In both cases, we need 'the same' number of simulations (ignoring that the number of iterations needed for convergence may differ), and the resulting estimators are asymptotically equivalent to the Monte Carlo estimator.

The multiple chain approach may also help in assessing convergence of the algorithm as discussed by Gelman and Rubin (1992). The basic idea is to plot all the Markov chains and consider them converged when they look alike. Some numerical measures are also considered; they are implemented in the itsim software written by Gelman.

Other ways of determining convergence exist; most of the methods developed for Markov chain Monte Carlo methods can be used. A recent review of these methods is given by Brooks and Roberts (1998). In the example in Section 4.3, apart from visual inspection of plots of the Markov chain, the itsim software and the gibbsit software written by Lewis (Raftery and Lewis 1992), both available from StatLib, have been used to determine convergence. The results suggest that the chosen burn-in is sufficient. In general more than one method should be used, since these methods are only able to detect lack of convergence, not to prove convergence.

The fraction of missing information determines the speed of convergence (in a neighbourhood of the MLE) for the EM algorithm (cf. Dempster *et al.* 1977). It will be natural to expect the same to be the case for the StEM algorithm. Also the more data are missing, the slower the StE-step will typically turn out to be.

In applications the StEM algorithm appears to converge quickly towards the MLE. The

simulations done for Section 4.3 run in a few seconds, but this example is clearly too simple to give any real indication of run-times. Celeux *et al.* (1996) report simulation experiments with StEM applied to finite mixtures. Their CPU times are in the range of 3–350 seconds depending on sample size and data generating model. Diebolt and Ip (1996) report an application of StEM with a 5 hour CPU time. In their example the dimension of θ is about 100 and the sample size is 1000. There is a large fraction of missing information, and the StE-step requires a Gibbs sampler.

5.2. Unidentified parameters

Unidentified parameters are typically problematical in incomplete data problems: due to the incompleteness, parameters identified in the complete data model may be unidentifiable in the observed data model. As indicated by Diebolt and Ip (1996), the StEM algorithm may be useful for looking at incomplete data problems with unidentified parameters. In this subsection we will discuss large-sample behaviour of the StEM algorithm when some parameters are unidentified.

We shall here only consider the case where unidentified parameters make the observed data information, $I(\theta_0)$, singular. A non-singular information matrix means that the parameter is (at least) locally identified. Hence we consider 'globally' unidentified parameters.

If $I(\theta_0)$ is singular, then the fraction of observed information, $I - F(\theta_0) = I(\theta_0)V(\theta_0)^{-1}$, is also singular. Hence some of the eigenvalues of $I - F(\theta_0)$ are zero. Suppose that $d - r$ of the eigenvalues of $I - F(\theta_0)$ are zero. Then we can write $I - F(\theta_0)^{\mathrm{T}}$ as $\alpha\beta^{\mathrm{T}}$, where α and β are $d \times r$ matrices of full rank. Since $I - F(\theta_0)$ is similar to a symmetric matrix we also have that $\mathrm{rk}[I - F(\theta_0)] = r$. This implies that $\beta^{\mathrm{T}}\alpha$ is non-singular and that the transformation

$$\theta \to \begin{bmatrix} \alpha_\perp^{\mathrm{T}}\theta \\ \beta^{\mathrm{T}}\theta \end{bmatrix} \tag{46}$$

where $\alpha_\perp^{\mathrm{T}}$ is a $(d - r) \times d$ matrix spanning $(\text{span } \alpha^{\mathrm{T}})^\perp$, is a bijection, i.e. a reparametrization (cf. Johansen 1995).

Now, $\beta^{\mathrm{T}}\theta$ is the identified part of θ, and $\alpha_\perp^{\mathrm{T}}\theta$ is the unidentified part. This is so since the fractions of missing information for the parameters $\beta^{\mathrm{T}}\theta$ and $\alpha_\perp^{\mathrm{T}}\theta$ are

$$\begin{aligned} (\beta^{\mathrm{T}}\mathrm{E}_{\theta_0} I_Y(\theta_0)\beta)(\beta^{\mathrm{T}}V(\theta_0)\beta)^{-1} &= I - \beta^{\mathrm{T}}\alpha, \\ (\alpha_\perp^{\mathrm{T}}\mathrm{E}_{\theta_0} I_Y(\theta_0)\alpha_\perp)(\alpha_\perp^{\mathrm{T}}V(\theta_0)\alpha_\perp)^{-1} &= I. \end{aligned} \tag{47}$$

Hence the fraction of observed information for $\beta^{\mathrm{T}}\theta$ is $\beta^{\mathrm{T}}\alpha$, which is non-singular, whereas it is 0 for $\alpha_\perp^{\mathrm{T}}\theta$.

If we apply the same transformation to the Gaussian AR(1) process $Z_t = F(\theta_0)^{\mathrm{T}} Z_{t-1} + \varepsilon_t$, we obtain

$$\begin{bmatrix} \alpha_\perp^{\mathrm{T}} Z_t \\ \beta^{\mathrm{T}} Z_t \end{bmatrix} = \begin{bmatrix} \alpha_\perp^{\mathrm{T}} F(\theta_0)^{\mathrm{T}} Z_{t-1} \\ \beta^{\mathrm{T}} F(\theta_0)^{\mathrm{T}} Z_{t-1} \end{bmatrix} + \begin{bmatrix} \alpha_\perp^{\mathrm{T}} \varepsilon_t \\ \beta^{\mathrm{T}} \varepsilon_t \end{bmatrix} = \begin{bmatrix} \alpha_\perp^{\mathrm{T}} Z_{t-1} \\ (I - \beta^{\mathrm{T}}\alpha)\beta^{\mathrm{T}} Z_{t-1} \end{bmatrix} + \begin{bmatrix} \alpha_\perp^{\mathrm{T}} \varepsilon_t \\ \beta^{\mathrm{T}} \varepsilon_t \end{bmatrix}. \tag{48}$$

Consequently, since Lemma 3 applies even when $F(\theta_0)$ has eigenvalues equal to 1 and the arguments leading to Proposition 2 may be applied to the identifiable part of the parameter, $\beta^{\mathrm{T}}\theta$, we get that the $\beta^{\mathrm{T}}\tilde{\theta}_n$ is an asymptotically normal with mean $\beta^{\mathrm{T}}\theta_0$ and variance

$$\beta^{\mathrm{T}} I(\theta_0)^{-1}[2I - \{I + F(\theta_0)\}^{-1}]\beta = \beta^{\mathrm{T}} I(\theta_0)^{-1}\beta[2I - (\beta^{\mathrm{T}}\alpha)^{-1}] \tag{49}$$

if $(\beta^{\mathrm{T}}\tilde{\theta}_n(k))_{k\in\mathbb{N}}$ is ergodic and $\sqrt{n}(\beta^{\mathrm{T}}\tilde{\theta}_n - \beta^{\mathrm{T}}\hat{\theta}_n)$ is tight.

Thus if the StEM algorithm is used with some parameters completely unidentified we obtain large-sample results like the ones discussed in the previous sections for the identified part of the parameter and will expect the unidentified part of the parameter to behave as a random walk for large-sample sizes.

Remark. We have not assumed that $\beta^{\mathrm{T}}\theta$ is identifiable, only shown (implicitly) that the corresponding information matrix is non-singular. Hence, $\beta^{\mathrm{T}}\theta$ is locally identifiable. Usually this implies that there is a MLE which is consistent for $\beta^{\mathrm{T}}\theta_0$, where this point is one of possibly many isolated parameter values giving rise to the same distribution. It is not clear whether $\beta^{\mathrm{T}}\tilde{\theta}_n$ is consistent for one of these $\beta^{\mathrm{T}}\theta_0$ values if $\beta^{\mathrm{T}}\theta$ is only locally identifiable. Since the Markov chain $\beta^{\mathrm{T}}\tilde{\theta}_n(k))_{k\in\mathbb{N}}$ is irreducible, it will visit neighbourhoods of all the $\beta^{\mathrm{T}}\theta_0$ values corresponding to the true distribution of the data. Hence we expect that tightness of $\sqrt{n}(\beta^{\mathrm{T}}\tilde{\theta}_n - \beta^{\mathrm{T}}\hat{\theta}_n)$ may be impossible to show, unless $\beta^{\mathrm{T}}\theta$ is actually identified.

5.3. Relaxing the assumptions

The main assumptions are the implicitly assumed feasibility of the two steps – the simulation and the maximization – of the algorithm. These assumptions may be relaxed to some extent.

The M-step may often be replaced by

$$\tilde{\theta}_n(k+1) = \tilde{\theta}_n(k) + V(\tilde{\theta}_n(k))^{-1}\frac{1}{n}\sum_{i=1}^{n} s_{\tilde{X}_i}(\tilde{\theta}_n(k)). \tag{50}$$

This new θ value will typically satisfy (5), which is enough to ensure the conclusion of Lemma 3.

If simulating from the exact conditional distributions is infeasible, one may consider simulating from another distribution giving an unbiased estimator of the conditional expectation required in the E-step of the EM algorithm. This should work as long as the conclusion of Lemma 3 holds partially (the asymptotic parameters in Lemma 3 may change as long as we get the AR(1) structure). In particular, the variance may change. Importance sampling may be an idea worth consideration in this respect. We expect, however, that in most cases the importance weights, which are needed to ensure an unbiased estimator, will be impossible to calculate. The results in Section 4.2.1 can – when the log-likelihood is linear in the simulations – be seen as an example of how the correct conditional

distributions may be replaced by another distribution, here a convolution of m correct conditional distributions scaled by $1/m$ to ensure an unbiased estimate. This convolution idea does not simplify the simulations, however.

The assumption of mutual equivalence of the distributions in the missing data model is rather essential, since it ensures the irreducibility of the Markov chain, $(\tilde{\theta}_n(k))_{k\in\mathbb{N}_0}$ (cf. Lemma 1). We note that mutual equivalence is not strictly necessary for irreducibility, but it is difficult to imagine an example, where the chain is irreducible but mutual equivalence fails. If the Markov chain is not irreducible, there may be absorbing states, which will typically bias the estimates severely. Intuitively, this problem may be avoided either by restarting the Markov chain (according to some fixed distribution) when an absorbing state is reached or by restricting the M-step so that absorbing states are excluded. The first approach was essentially applied by Celeux and Diebolt (1993), who gave asymptotic results for the special case of mixing proportions, where absorption is a problem.

The asymptotic normality and the rate of convergence of the observed data MLE seem essential for asymptotic normality of the StEM estimator. As noted in Section 2.3, the strong consistency of $\hat{\theta}_n$ is not really needed. In the previous subsection we have relaxed the assumption of identifiability of the parameter.

Tanner and Wong (1987) apply an algorithm, which is essentially the StEM algorithm, to estimate a hazard function nonparametrically based on grouped survival data with some promising simulation results. This suggests that it might be possible to extend the results discussed in this paper to suitably nice non-Euclidean parameters.

5.4. Conclusion

When estimating in missing data problems, the EM algorithm is often a useful method. It is derivative-free and, though it typically requires many iterations for convergence, each iteration is often fast (cf. Ruud 1991). When the E-step of the algorithm is infeasible, the StEM algorithm is an alternative. Furthermore, it has some advantages over the EM algorithm: it does not get stuck; it often provides more information about the data (cf. Diebolt and Ip 1996), for instance when parameters cannot be estimated; and even behaves better than the EM algorithm in some cases (cf Celeux *et al.* 1996). Unlike the EM algorithm, StEM always leads to maximization of a complete data log-likelihood in the M-step.

In this paper, asymptotic results have been shown and a number of finite simulation estimators have been discussed. As argued in Section 4 the larger the sample, the larger the simulation burden. Thus for large samples typically only a finite sample of the Markov chain is available for estimation of the unknown parameter. Furthermore, as any simulation is finite even if large, the simulation part of the variance of the estimators should be included in order not to underestimate the variance.

The StEM algorithm is a simulation-based method. As a consequence of this, any estimator obtained from a finite sample of the Markov chain involved will contain variation due to the simulations as well as variation due to the data. We note, however, that for $m = 1$ the inflation of the variance is less than 50% (cf. Proposition 4) and that the

simulation part of the variance can be made as small as required by choosing m sufficiently large; it goes down as $1/m$.

In some applications the added variance due to the simulations may be unacceptable. Even in this case the StEM algorithm may be a useful tool. For instance, the output from StEM can be used to give good starting points for other algorithms, for instance the MCEM algorithm with m large. Another possibility is to use a preliminary estimator derived from the StEM algorithm to obtain asymptotically efficient estimators by one iteration of the method of scoring if a suitable estimator of the observed data score function can be obtained. One could consider

$$\hat{\theta}_n^M = \tilde{\theta}_n - \left(\sum_{i=1}^{n} \tilde{I}_{y_i}(\tilde{\theta}_n) \right)^{-1} \sum_{i=1}^{n} \tilde{s}_{y_i}(\tilde{\theta}_n), \tag{51}$$

where $\tilde{s}_{y_i}(\theta) = (1/M)\sum_{j=1}^{M} s_{\tilde{X}_{ij}}(\theta)$ and $\tilde{I}_{y_i}(\theta)$ is an analogous estimator of $I_{y_i}(\theta)$ with M very large. See Hajivassiliou (1997) for more on this idea and Schick (1987) for technical conditions ensuring efficiency.

Acknowledgements

Part of this work was done while the author was visiting the University of Utrecht. The comments of an anonymous referee on a previous version have led to significant changes and a much improved paper.

References

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore, MD: Johns Hopkins University Press.

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

Brooks, S.P. and Roberts, G.O. (1998) Convergence assessment techniques for Markov Chain Monte Carlo. *Statist. Comput.*, **8**, 319–335.

Celeux, G. and Diebolt, J. (1986) The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Quart.*, **2**, 73–82.

Celeux, G. and Diebolt, J. (1993) Asymptotic properties of a stochastic EM algorithm for estimating mixing proportions. *Comm. Statist. Stochastic Models*, **9**, 599–613.

Celeux, G., Chauveau, D. and Diebolt, J. (1996) Stochastic versions of the EM algorithm: An experimental study in the mixture case. *J. Statist. Comput. Simulation*, **55**, 287–314.

Chan, K.S. and Ledolter, J. (1995) Monte Carlo EM estimation for time series models involving counts. *J. Amer. Statist. Assoc.*, **90**, 242–252.

Dempster A.P., Laird N.M. and Rubin D.B. (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B*, **39**, 1–38.

Diebolt, J. and Ip, E.H.S. (1996) Stochastic EM: method and application. In W.R. Gilks, S. Richardson, D.J. Speigelhalter (eds), *Markov Chain Monte Carlo in Practice.* London: Chapman & Hall.

Ethier, S.N. and Kurtz T.G. (1986) *Markov Processes. Characterization and Convergence*. New York: Wiley.

Gelman, A. and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.* **7**, 457–511.

Hajivassiliou, V.A. (1997) Some practical issues in maximum simulated likelihood. *http://econ.1se.ac.uk/~vassilis/*

Ibragimov, I.A. and Has'minskii, R.Z. (1981). *Statistical Estimation. Asymptotic Theory.* New York: Springer-Verlag.

Johansen, S. (1995) *Likelihood-Based Inference in Cointegrated Vector Autoregression Models*. Oxford: Oxford University Press.

Lehmann, E.L. (1983) *Theory of Point Estimation*. New York: Wiley.

Louis, T.A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc. Ser. B*, **44**, 226–233.

Meng, X.-L. and Rubin, D.B. (1991) Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *J. Amer. Statist. Assoc.*, **86**, 899–909.

Meyn, S.P. and Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability.* New York: Springer-Verlag.

Raftery, A.E. and Lewis, S.M. (1992) How many iterations in the Gibbs sampler?. In J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (eds), *Bayesian Statistics, Vol. 4*. Oxford: Oxford University Press.

Roussas, G.G. (1972) *Contiguity of Probability Measures*. London: Cambridge University Press.

Ruud, P.A. (1991) Extensions of estimation methods using the EM algorithm. *J Econometrics*, **49**, 305–341.

Schenker, N. and Welsh, A.H. (1987) Asymptotic results for multiple imputation. *Ann. Statist.*, **16**, 1550–1566.

Schick, A. (1987) A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference*, **16**, 89–105.

Tanner, M.A. and Wong, W.H. (1987) An application of imputation to an estimation problem in grouped lifetime analysis. *Technometrics*, **29**, 23–32.

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

Received January 1998 and revised December 1998