On the Convergence of the Monte Carlo Maximum Likelihood Method for Latent Variable Models

By R. DOUC[†], O. CAPPÉ[†], E. MOULINES[†] and C.P. ROBERT[‡]

[†] *Dpt. TSI / CNRS URA 820, ENST, Paris*
[‡] *CREST-INSEE, Paris and CNRS UPRES-A 6085, Université de Rouen*

Abstract

While much used in practice, latent variable models raise challenging estimation problems related with the intractability of their likelihoods. Monte Carlo Maximum Likelihood (MCML) is a simulation-based approach to likelihood approximation that has been proposed for complex latent variable models for which deterministic optimization procedures such as the Expectation-Maximization approach are not applicable. It is based on an importance sampling identity for the likelihood ratio, where the importance function is the complete model density at a given parameter value φ. This paper studies the asymptotic performance of the MCML method (in the number of observations n) against the choice of φ and of the number of simulations s_n used in the importance sampling approximation. We provide sufficient conditions for the MCML estimator to converge to the true value of the parameter with n. Our results imply in particular that the initialization parameter φ must be a $\sqrt{n}$-consistent estimate. Otherwise, the number of simulations necessary to attain convergence increases exponentially fast with the sample size.

Keywords

Maximum Likelihood Estimation, Monte Carlo Maximum Likelihood, Simulated Likelihood Ratio, Stochastic Optimization, Stochastic Approximation

Running Headline

Convergence of Monte Carlo Maximum Likelihood

1 Introduction

Monte Carlo Maximum Likelihood (MCML hereafter), as introduced by Geyer and Thompson (1992), is a widely accepted method for maximum likelihood estimation in cases where direct computation and/or maximization of the likelihood is intractable. The method can be used for quite general models but is particularly relevant for latent variable models, possibly with unknown normalizing constants (see Geyer, 1996, Sandmann and Koopman, 1998, Thompson, 1994). The method is based on an importance sampling identity that represents the (observed) likelihood ratio $g_n(\mathbf{y}_{1:n}; \theta)/g_n(\mathbf{y}_{1:n}; \varphi)$ as the expectation of the complete likelihood ratio

$$E_\varphi \left[\left. \frac{f_n(\mathbf{y}_{1:n}, \mathbf{Z}_{1:n}; \theta)}{f_n(\mathbf{y}_{1:n}, \mathbf{Z}_{1:n}; \varphi)} \right| \mathbf{y}_{1:n} \right]$$

for an arbitrary value of the parameter φ (where $\mathbf{y}_{1:n}$ denotes the sample and $\mathbf{Z}_{1:n}$ the corresponding latent variables). The replacement of the expected ratio by a Monte Carlo average, where s_n realizations of $\mathbf{Z}_{1:n}$ are simulated conditionally on $\mathbf{y}_{1:n}$ and the parameter value φ, then provides a simulated approximation to the (observed) likelihood ratio.

Note that the terminology used to describe some related methods is not yet unified: the "Simulated Likelihood Ratio" of Billio *et al.* (1998) is equivalent to the MCML method, while the "Simulated Maximum Likelihood" approach of Danielson and Richard (1993) is distinct. The "Simulated Maximum Likelihood Estimator" considered by Lee (1995) is a variation of MCML with improved convergence properties, but is more limited in scope and more computationally intensive. For a more complete survey of simulation based approaches, see Gouriéroux and Monfort (1993). As indicated above, MCML is a special case of importance sampling ideas (see Geyer, 1996) which is particularly relevant in applications where the sampling density is not chosen from the family of conditional densities associated with the model, as in Sandmann and Koopman (1998).

Geyer (1996) argues that the efficiency of MCML stems from its simplicity, given that the unknown likelihood (ratio) is first approximated using a single round of simulations and the approximation is then maximized via a standard maximization tool. Indeed, when compared with other generally applicable simulation based approaches to maximum likelihood estimation, like the Stochastic Approximation approach of Younes (1988), the Monte Carlo EM of Wei and Tanner (1990), the Stochastic EM of Celeux and Diebolt (1985), or the Stochastic Approximation EM of Lavielle, Delyon and Moulines (1999), a strong incentive for using MCML is that conditional simulations are run *only once* and for a *single fixed value* of φ. Hence, the maximization and the simulation steps are not nested, unlike the algorithms above. In Geyer's (1996) terminology, this classifies MCML as a *stochastic approximation* technique, as opposed to these *stochastic optimization* techniques. There is however clear empirical evidence that the choice of φ has a strong influence on the behavior of MCML (see Geyer, 1996, or Billio *et al*, 1996). While the convergence to the maximum likelihood estimator as s_n goes to infinity clearly holds for a fixed n, as shown by Geyer (1996) and recalled in Section 2, a deeper and thus asymptotic examination of the dependence of the method, and of the convergence of the MCML estimator, on the parameters φ and s_n is thus most timely.

After a brief description of the method in Section 2, a simple latent variable example is discussed in Section 3. This example is truly an illustration, rather than a representative statistical application of the method, since it corresponds to a trivial case where the maximum likelihood estimate is known analytically. For this example, the asymptotic variance of the estimates is found to be extremely sensitive to the parameter value φ used for simulating the unobserved data. This sensitivity is basically exponential in the sample size n, with larger data sets requiring increasingly larger Monte Carlo simulations for the method to actually converge. It is then shown in Section 4 that this seemingly counterintuitive behavior is quite representative of what happens in a large class of latent variable models: the number of simulations s_n has to increase exponentially fast with n for the MCML estimator to be consistent and asymptotically efficient. In Section 5, we attain a more positive result in the sense that the asymptotic covariance matrix of the MCML estimate is bounded (in n) when the initialization parameter φ_n is a $\sqrt{n}$-consistent estimate of the true parameter. In this case, s_n can grow at any rate and even be constant, and the MCML algorithm will still be consistent. The overall conclusion of this paper is therefore that the MCML method should only be used in settings where a preliminary $\sqrt{n}$-consistent estimate of θ is available. A natural candidate is a noninformative Bayes estimate, given that the MCML algorithm can be complemented by a parameter simulation stage to provide a Gibbs sampler for the approximation of Bayes estimates, as in Billio *et al.* (1998).

For simplicity's sake, we focus on models for which the probability distribution of the complete data is known exactly (that is, including the normalizing constant) and assume that exact independent sampling from the conditional distribution of the latent variable is feasible (which is to say that Gibbs sampling applies for this component). Moreover, the results of Sections 4 and 5 are derived under the more restrictive hypothesis that the random variables from the complete model (i.e, the actual observations and the latent variables) are mutually independent. This class of models includes in particular mixture models. Similar results do hold for more general latent variable models and notably for hidden Markov models (where the latent variables are assumed to be Markovian, as in Robert and Titterington, 1998) under the appropriate extensions of technical conditions.

2 Conditional convergence to the maximum likelihood estimate

We first consider the convergence of the MCML estimates to the maximum likelihood estimate as the number of simulations of the latent data increases. Following a remark of Geyer (1994), this type of convergence is referred to as "conditional convergence". In fact, the properties of the MCML estimates are very different in this setting from those in the framework of Sections 4 and 5, when the number n of available observations increases, the number s_n of simulations being then a function of n. Proofs for this section are omitted since Theorems 1 and 2 are basically restatements of Theorems 4 and 7 of Geyer (1994).

2.1 The MCML Algorithm

We begin with a brief description of the method and of the associated notations. For a more complete account of MCML, including its application to unnormalized density functions, see Geyer (1994, 1996).

Let $\mathbf{y}_{1:n} = (y_1, \ldots, y_n)$ denote the observation and $\mathbf{Z}_{1:n} = (Z_1, \ldots, Z_n)$ the associated vector of latent variables. The likelihood function is then

$$g_n(\mathbf{y}_{1:n}; \theta) \triangleq \int f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \theta) \mu_n(\mathbf{dz}_{1:n}),$$

where the complete data density $f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \theta)$ belongs to a parametric family of positive functions, normalized with respect to some dominating measure μ_n, with parameter $\theta \in \Theta \subset \mathbb{R}^d$. The conditional density of the latent variables is denoted by

$$p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n}; \theta) \triangleq \begin{cases} \frac{f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \theta)}{g_n(\mathbf{y}_{1:n}; \theta)} & \text{if } g_n(\mathbf{y}_{1:n}; \theta) > 0, \\ 0 & \text{otherwise,} \end{cases}$$

with respect to the appropriate dominating measure (see Billio *et al.*, 1998, for a description of the conditioning issues and of the dominating measures) and $P_\theta(\cdot|\mathbf{y}_{1:n})$ stands for the associated probability distribution. Furthermore, for a measurable function $\psi(\mathbf{y}_{1:n}, \mathbf{z}_{1:n})$, we denote by

$$E_\theta[\psi(\mathbf{y}_{1:n}, \mathbf{Z}_{1:n})|\mathbf{y}_{1:n}] \triangleq \int \psi(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}) p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n}; \theta) \mu_n(\mathbf{dz}_{1:n}),$$

the conditional expectation under parameter value θ, and similarly use $Var_\theta[\cdot|\mathbf{y}_{1:n}]$ for the conditional variance.

MCML is based on the fact that the observed likelihood ratio may be expressed as the conditional expectation of the complete data likelihood ratio (see Geyer, 1994):

$$l_{\varphi,n}(\theta) \triangleq \frac{g_n(\mathbf{y}_{1:n}; \theta)}{g_n(\mathbf{y}_{1:n}; \varphi)} = \int \frac{f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \theta)}{f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \varphi)} p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n}; \varphi) \mu_n(\mathbf{dz}_{1:n}), \tag{1}$$

where φ is any arbitrary point in the parameter space Θ, as can be seen by a standard importance sampling argument. The method builds on this identity by deriving a Monte Carlo approximation of the likelihood ratio $l_{\varphi,n}(\theta)$,

$$\hat{l}^s_{\varphi,n}(\theta) = \frac{1}{s} \sum_{k=1}^{s} \frac{f_n(\mathbf{y}_{1:n}, \mathbf{Z}^k_{1:n}; \theta)}{f_n(\mathbf{y}_{1:n}, \mathbf{Z}^k_{1:n}; \varphi)} \tag{2}$$

where the $(\mathbf{Z}^k_{1:n})$'s $(k = 1, ..., s)$ are s simulated replications of the complete latent data vector, distributed according to the conditional distribution $p_n(\cdot|\mathbf{y}_{1:n}; \varphi)$, where φ is the same fixed arbitrary point as in (1). The Monte Carlo maximum likelihood (MCML) estimate $\hat{\theta}^s_{\varphi,n}$ is then defined as the maximizer of (2) with respect to θ. One of the advantages of the method is that (2) can be readily adapted to the case where the complete data probability density is only known up to a normalizing constant as in Geyer (1994). For simplicity's sake, we however assume here that the normalizing constant is also known.

Note that we focus on the case where the $\mathbf{Z}^k_{1:n}$'s are independent (in k) generations from $p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n}; \varphi)$. We are thus omitting the natural extension to Markov Chain Monte Carlo simulations. This is an important issue in practice since exact independent simulation is not feasible in many cases. This extension simply requires additional assumptions on the mixing rate of the chain associated with $\mathbf{Z}^k_{1:n}$ so that the rates of convergence are preserved. It will not be considered any further in this paper for simplicity's sake.

2.2 Conditional convergence results

We first recall some convergence results for fixed sample sizes.

Theorem 1 *Let φ be an arbitrary point in Θ and suppose that*

(a) Θ *is compact with a nonempty interior,*

(b) $f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \theta)/f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \varphi)$ *is* μ_n *almost everywhere continuous with respect to* θ *on* Θ*,*

(c) $E_\varphi \left[\sup_{\theta\in\Theta} \left\{ \frac{f_n(\mathbf{y}_{1:n}, \mathbf{Z}_{1:n}; \theta)}{f_n(\mathbf{y}_{1:n}, \mathbf{Z}_{1:n}; \varphi)} \right\} \middle| \mathbf{y}_{1:n} \right] < \infty$,

(d) $l_{\varphi,n}(\theta)$ *has a unique maximum on* Θ, $\hat{\theta}_n$, *which belongs to the interior of* Θ.

Then,

$$\lim_{s\to\infty} \hat{\theta}^s_{\varphi,n} = \hat{\theta}_n \qquad w.\ p.\ 1$$

For a proof of this theorem, see Geyer (1994, Theorem 4), with weaker conditions. Note that condition (c) is not innocuous, as it guarantees that the variance of $\hat{\theta}^s_{\varphi,n}$ is finite, which does not always hold, as pointed out by Billio *et al.* (1998).

Theorem 2 (Geyer, 1994 - Th. 7) *Suppose that*

(a) *The maximum likelihood estimate* $\hat{\theta}_n$ *is unique and belongs to the interior of* Θ,

(b) $\hat{\theta}^s_{\varphi,n}$ *converges in probability to* $\hat{\theta}_n$,

(c) $g_n(\mathbf{y}_{1:n}; \theta) = \int f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \theta)\mu_n(\mathbf{dz}_{1:n})$ *can be differentiated twice under the integral sign w.r.t.* θ,

(d) $s^{1/2}\nabla_\theta \log \hat{l}^s_{\varphi,n}(\hat{\theta}_n) \xrightarrow[s\to\infty]{\mathcal{L}} N(0, V_{\varphi,n}(\mathbf{y}_{1:n}))$,

(e) *The observed information matrix* $D_n(\mathbf{y}_{1:n}) = -\nabla^2_\theta \log g_n(\mathbf{y}_{1:n}; \hat{\theta}_n)$ *is positive definite,*

(f) $\nabla^3_\theta \log l^s_{\varphi,n}(\theta)$ *is bounded in probability uniformly in a neighborhood of* $\hat{\theta}_n$.

Then

$$\sqrt{s}(\hat{\theta}^s_{\varphi,n} - \hat{\theta}_n) \xrightarrow{\mathcal{L}} N(0, \Gamma_{\varphi,n}(\mathbf{y}_{1:n})), \tag{3}$$

where

$$\Gamma_{\varphi,n}(\mathbf{y}_{1:n}) = \left({D_n}^{-1} V_{\varphi,n} {D_n}^{-1}\right)(\mathbf{y}_{1:n}).$$

In the setup of this paper (i.e., with exact independent simulations of the latent variables), the term $V_{\varphi,n}(\mathbf{y}_{1:n})$ can be readily computed as

$$V_{\varphi,n}(\mathbf{y}_{1:n}) = E_\varphi \left[\frac{\nabla_\theta f_n(\mathbf{y}_{1:n}, \mathbf{Z}_{1:n}; \hat{\theta}_n)}{f_n(\mathbf{y}_{1:n}, \mathbf{Z}_{1:n}; \varphi)} \frac{\nabla^T_\theta f_n(\mathbf{y}_{1:n}, \mathbf{Z}_{1:n}; \hat{\theta}_n)}{f_n(\mathbf{y}_{1:n}, \mathbf{Z}_{1:n}; \varphi)} \middle| \mathbf{y}_{1:n} \right] (l_{\varphi,n}(\hat{\theta}_n))^{-2},$$

using assumptions (b) and (c) of Theorem 2. Then, by definition of $\hat{\theta}_n$,

$$\nabla_\theta f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \hat{\theta}_n) = \nabla_\theta p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n}; \hat{\theta}_n) g_n(\mathbf{y}_{1:n}; \hat{\theta}_n),$$

so that

$$V_{\varphi,n}(\mathbf{y}_{1:n}) = E_\varphi \left[\frac{\nabla_\theta p_n(\mathbf{Z}_{1:n}|\mathbf{y}_{1:n}; \hat{\theta}_n)}{p_n(\mathbf{Z}_{1:n}|\mathbf{y}_{1:n}; \varphi)} \frac{\nabla^T_\theta p_n(\mathbf{Z}_{1:n}|\mathbf{y}_{1:n}; \hat{\theta}_n)}{p_n(\mathbf{Z}_{1:n}|\mathbf{y}_{1:n}; \varphi)} \middle| \mathbf{y}_{1:n} \right].$$

Thus

$$V_{\varphi,n}(\mathbf{y}_{1:n}) = E_{\hat{\theta}_n}\left[\left(\frac{p_n(\mathbf{Z}_{1:n}|\mathbf{y}_{1:n};\hat{\theta}_n)}{p_n(\mathbf{Z}_{1:n}|\mathbf{y}_{1:n};\varphi)}\right)\right.$$
$$\left.\nabla_\theta \log p_n(\mathbf{Z}_{1:n}|\mathbf{y}_{1:n};\hat{\theta}_n)\nabla_\theta^T \log p_n(\mathbf{Z}_{1:n}|\mathbf{y}_{1:n};\hat{\theta}_n)\right|\mathbf{y}_{1:n}\Bigg] . \quad (4)$$

Note that $D_n(\mathbf{y}_{1:n})$ is the observed Fisher information matrix; $V_{\varphi,n}(\mathbf{y}_{1:n})$ closely resembles the Fisher information matrix associated with the conditional distribution $P_{\hat{\theta}_n}(\cdot|\mathbf{y}_{1:n})$, save for the presence of the conditional likelihood ratio $p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n};\hat{\theta}_n)/p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n};\varphi)$. Anticipating the results of Section 4, the asymptotic behavior of the method is mainly governed by the fact that for most models of interest the conditional likelihood ratio $p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n};\hat{\theta}_n)/p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n};\varphi)$ has an exponentially diverging behavior as n increases. Before investigating this general behavior of $V_{\varphi,n}(\mathbf{y}_{1:n})$ in Section 4, we first consider an illustrative example.

3 A Simple illustration

3.1 Asymptotic results

Assume that the complete data consist of observed scalar variables Y_i, for $i = 1, \dots, n$ supplemented by corresponding latent variables Z_i (also scalar) such that the complete data distribution is bivariate normal

$$\begin{pmatrix} Y_i \\ Z_i \end{pmatrix} \sim N_2\left(\begin{pmatrix} \theta \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right), \quad (5)$$

i.e. the observed and latent variables are jointly normal. Moreover, the complete data model is assumed to be i.i.d. and the mean of the observation θ is the only parameter of interest, whereas the correlation ρ is known and fixed.

Of course, in this particular example, the (observed) maximum likelihood estimate of θ is $\hat{\theta}_n = (1/n)\sum_{i=1}^n y_i$. Note however that the (complete) maximum likelihood $\hat{\theta}_n - (\rho/n)\sum_{i=1}^n z_i$ usually differs from the observed maximum likelihood estimate (at least if $\rho \neq 0$), so that this example, although extremely simplified, is nontrivial.

The asymptotic covariance terms as defined in Theorem 2 are easily evaluated as

$$D_n(\mathbf{y}_{1:n}) = n, \quad (6)$$

$$V_{\varphi,n}(\mathbf{y}_{1:n}) = \frac{n\rho^2}{(1-\rho^2)}\left[1 + \frac{n\rho^2}{1-\rho^2}(\hat{\theta}_n - \varphi)^2\right]\exp\left(\frac{n\rho^2}{1-\rho^2}(\hat{\theta}_n - \varphi)^2\right). \quad (7)$$

For this particular model, $V_{\varphi,n}(\mathbf{y}_{1:n})$ only depends upon the observed data through $\hat{\theta}_n$ the maximum likelihood estimate. As n increases, $V_{\varphi,n}(\mathbf{y}_{1:n})$ diverges exponentially fast with an exponential rate which is proportional to the squared difference between the initial guess of the parameter value, φ, and the actual maximum likelihood estimate $\hat{\theta}_n$, unless φ also varies with n.

3.2 Simulation results

In Theorem 2, terms that decrease at a rate faster than $s^{-1/2}$ are ignored. For finite sample sizes however, these terms may play an important role in the performance of the method. We therefore conducted a series of numerical simulations to illustrate the finite sample behavior of MCML.

The MCML is straightforward in this particular example because the conditional distribution (under which sampling is done) factorizes through a sufficient statistic with known distribution: the Monte Carlo approximation to the likelihood ratio $l_{\varphi,n}(\theta)$ is then

$$\hat{l}^s_{\varphi,n}(\mathbf{y}_{1:n};\theta) = \frac{1}{s}\sum_{k=1}^{s} \exp\left\{-\frac{n}{1-\rho^2}(\theta-\varphi)\left(\frac{\theta+\varphi}{2} + \rho\bar{Z}^k - \hat{\theta}_n\right)\right\}, \tag{8}$$

with $\bar{Z}^k = (1/n)\sum_{i=1}^n Z_i^k$, where $(\mathbf{Z}^k_{1:n})$ denotes the kth simulation of the complete vector of latent variables, and

$$\bar{Z}^k \sim N\left(\rho(\hat{\theta}_n - \varphi), (1-\rho^2)/n\right). \tag{9}$$

[Figure 1 about here.]

Figure 1 displays the dispersion of the algorithm estimates as a function of the number of simulations (from one to one million, on a log scale) for different sample sizes. The quantity displayed is $\sqrt{s}(\hat{\theta}^s_{\varphi,n} - \hat{\theta}_n)$, that is the output of the algorithm recentered around the MLE $\hat{\theta}_n$ and scaled by the asymptotic normalizing factor $s^{-1/2}$. Each boxplot was obtained from 500 independent runs of the algorithm, using fixed data sets (one for each size from 15 to 120) and starting the algorithm from a fixed distance from the MLE ($\varphi - \hat{\theta}_n = -0.1$). The black box on the right of each figure corresponds to the quartiles computed from the asymptotic variance $\Gamma_{\varphi,n}(\mathbf{y}_{1:n})$ assuming normality.

Two interesting features can be observed from Figure 1. The first is that, when comparing the four plots, for a given number of simulations, the normalized estimates $\sqrt{s}(\hat{\theta}^s_{\varphi,n} - \hat{\theta}_n)$ get more and more dispersed as the sample size increases, as expected in the form of $\Gamma_{\varphi,n}(\mathbf{y}_{1:n})$. The normal quartiles, based on the value of $\Gamma_{\varphi,n}(\mathbf{y}_{1:n})$, displayed on the right of each plot show that the observed dispersion of the estimates for a large number of simulations is generally in accordance with the Central Limit Theorem dispersion of Section 2, although the asymptotic spread is clearly not yet achieved after 10^6 simulations for the larger data set (120 observations).

[Figure 2 about here.]

The second noteworthy feature is that each plot, when considered from left to right, shows three different stages. For small numbers of simulations, the normalized dispersion of the estimates is very small. For moderate numbers of simulations, the dispersion increases and the distribution of the normalized estimates is distinctively positively skewed with a heavy positive tail. Finally, for large numbers of simulations the distribution of the normalized estimates is more symmetric and compatible with the Central Limit Theorem. Moreover, the transition between these three stages occurs for numbers of simulations which increase with the sample size. This is almost certainly due to the fact that the higher order terms which are neglected when obtaining the CLT of Theorem 2 also exhibit exponential dependence on the sample size. As a consequence, for moderate numbers of simulations (several hundred to several thousand) and large sample sizes (one hundred observations or more), the asymptotic stage (third one) is not yet reached and the bias term is prevalent. This last point is particularly clear when looking at the dispersion of the unscaled recentered estimates $(\hat{\theta}^s_{\varphi,n} - \hat{\theta}_n)$ displayed on Figure 2 (which corresponds to the larger data set in Figure 1). For moderate numbers of simulations, the predominant effect is thus an important bias of the estimates towards φ, as also observed in Billio *et al.* (1998) on other models. In the extreme case where a single simulation is used, it is easy to check that $\hat{\theta}^1_{\varphi,n} = \hat{\theta}_n - \rho\bar{Z}_1$, where $\bar{Z}_1$ is distributed from (9), so the bias is equal to

$$E(\hat{\theta}^1_{\varphi,n} - \hat{\theta}_n) = \rho^2(\varphi - \hat{\theta}_n),$$

which gives -0.081 in the case of Figs. 1-2. The bias then decreases slowly with the number of simulations as it still amounts to -0.017 after 1000 iterations.

3.3 Comparison with the Stochastic EM approach

For comparison purpose, we consider the application of the Stochastic Expectation-Maximization (or SEM) approach to the same model. In the SEM approach introduced by Celeux and Diebolt (1985), each iteration consists of maximizing the complete data likelihood where the missing data is imputed stochastically by drawing the latent variables according to their conditional distribution given the current estimate of the parameters. The SEM iterates form a Markovian sequence which converges under general conditions to a stationary distribution (Diebolt and Ip, 1996). Precise characterization of this limit law is a difficult issue (Ip, 1994), except in some particular cases such as the simple example considered in this section (see below). Recent results by Nielsen (2000) however suggest that although the stationary distribution of the SEM iterations cannot in general be directly related to the maximum likelihood estimate, it nonetheless provides a mean to construct efficient parameter estimators.

As already noted, the complete data maximum likelihood estimation of θ is given by $\hat{\theta}_n - \rho\bar{Z}$ where $\bar{Z}$ is the normalized conditional mean of the unobserved component which is distributed according to (9) (ϕ being our current guess of the parameter). Denoting the sequence of SEM iterates by $(\tilde{\theta}^s_n)_{s\geq 1}$, it is then easily checked that

$$\tilde{\theta}^{s+1}_n - \hat{\theta}_n = \rho^2(\tilde{\theta}^s_n - \hat{\theta}_n) + U_{s+1},$$

where $(U_s)_{s\geq 2}$ is and iid sequence of zero mean Gaussian random variables with variance $\rho^2(1-\rho^2)/n$; That is, the sequence of SEM iterates forms and AR 1 Gaussian process with stationary distribution

$$N\left(\hat{\theta}_n, \frac{\rho^2}{n(1+\rho^2)}\right). \tag{10}$$

As suggested by Diebolt and Ip (1996), the ergodic average of $(\tilde{\theta}^k_n)_{1\leq k\leq s}$ yields a rate $\sqrt{s}$ estimate of $\hat{\theta}_n$ with asymptotic variance $\rho^2/(n(1-\rho^2))$. The above results are conditional upon a particular outcome of the observations $\mathbf{Y}_{1:n}$. Nielsen (2000) however shows that if $\tilde{\theta}^\infty_n$ denotes the limiting variable distributed according to the stationary distribution of SEM for a given value of n, $\tilde{\theta}^\infty_n$ satisfies an unconditional central limit result, which may be written (in our example)

$$\sqrt{n}\left(\tilde{\theta}^\infty_n - \theta_*\right) \xrightarrow{\mathcal{D}} N\left(0, 1 + \frac{\rho^2}{1+\rho^2}\right) \tag{11}$$

For more general models, it does not hold true that the ergodic average of the SEM iterates converges to the MLE for a fixed value of n. However, the dependence on n observed both in (10) and (11) suggests that it is possible to build efficient parameter estimators from the SEM approach with a reasonable number of simulations – see (Nielsen, 2000) for details. This behavior is of course in sharp contrast with that of the conditional asymptotic variance of the MCML estimate computed in (6)-(7) which diverges as n increases.

4 Asymptotic properties of MCML under fixed initialization

In this section, we show that the behavior observed for the simple example above is characteristic of a large class of models where the method applies. For simplicity's sake, we only consider i.i.d. complete data models, i.e. such that

$$f_n(\mathbf{y}_{1:n}, \mathbf{z}_{1:n}; \theta) = \prod_{i=1}^{n} f(y_i, z_i; \theta).$$

Insight into the following results stems from rewriting (2) as

$$\hat{l}^s_{\varphi,n}(\theta) = l_{\varphi,n}(\theta) \left[\frac{1}{s} \sum_{k=1}^{s} \frac{p_n(\mathbf{z}^k_{1:n}|\mathbf{y}_{1:n}; \theta)}{p_n(\mathbf{z}^k_{1:n}|\mathbf{y}_{1:n}; \varphi)} \right]. \tag{12}$$

In fact, MCML is equivalent to approximating the constant 1 by importance sampling with $p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n}; \theta)$ as target density and $p_n(\mathbf{z}_{1:n}|\mathbf{y}_{1:n}; \varphi)$ as importance (or proposal) density. But, for identifiable regular models, the supports of these two densities tend to separate as n goes to infinity when $\varphi \neq \theta$ (see (9) for the example of Section 3). Thus the importance weights in (12) degenerate, becoming either very small or very large depending on the value of θ, which is a well-known cause of instability for the importance sampling method (Geweke, 1988).

In contrast with the results of Section 2, the results in this section bear on the convergence to the actual value θ_* of the parameter when *both n and the number of simulations* s_n increase. Note that, since we are primarily interested in the growth rate of s_n with n, the former is explicitly written as a function of the latter. For technical simplicity and coherence with Section 5 (where the initialization of MCML varies with n) we also consider that the $\mathbf{Z}_{n,1:n}$'s are simulated independently for each sample size n, hence the notation $Z^k_{n,i}$ where $1 \leq i \leq n$ denotes the observation index, $1 \leq k \leq s_n$ the simulation index and n refers to the sample size. In a sequential setting, this assumption would be quite subefficient, but this is not the problem here, where we are rather focusing on the asymptotic properties of the MCML estimator.

In addition to the notations $P_\theta(\cdot|\mathbf{Y}_{1:n})$ and $E_\theta[\cdot|\mathbf{Y}_{1:n}]$ defined in Section 2, we use $P(\cdot)$ and $E[\cdot]$ to denote respectively the distribution and expectation of functions of $\{Y_n\}_{n\in\mathbb{N}}$, under the true value θ_* of the parameter. The switch from lower case to upper case notation for the observations Y_i is meant to stress the fact that from now on the observations themselves will be considered as random rather than being fixed.

The first item of this section is Theorem 3 which describes the asymptotic behavior of the limiting covariance $\Gamma_{\varphi,n}(\mathbf{Y}_{1:n})$ featured in Theorem 2 (conditional CLT). The obtained asymptotic form shows that the number s_n of simulations should grow exponentially fast with n in order to guarantee that $\Gamma_{\varphi,n}(\mathbf{Y}_{1:n})$ remains bounded. In this case, MCML estimation is indeed consistent (Theorem 4) and asymptotically efficient (Theorem 5) if s_n has a fast enough exponential growth rate. In practice, the perspective of performing such large numbers of simulations is obviously unrealistic. As shown in Section 5, a solution to this shortcoming relies on initializing the MCML algorithm from a $\sqrt{n}$-consistent estimate of θ.

4.1 Asymptotic conditional covariance

In the following, we need further regularity conditions in addition to those of Theorems 1-2. For technical simplicity, we mostly use basic Wald-type regularity conditions. Denote

$$\begin{aligned}
t(y;\varphi,\theta) &\triangleq \int \frac{p(z|y;\theta)^2}{p(z|y;\varphi)}\mu(dz)\,, \\
\tilde{p}(z|y;\varphi,\theta) &\triangleq \frac{1}{t(y;\varphi,\theta)}\frac{p(z|y;\theta)^2}{p(z|y;\varphi)}\,, \\
A(y;\varphi,\theta) &\triangleq \int \left[\nabla_\theta \log p(z|y;\theta)\nabla_\theta^T \log p(z|y;\theta)\right]\tilde{p}(z|y;\varphi,\theta)\mu(dz)\,, \\
b(y;\varphi,\theta) &\triangleq \int \nabla_\theta \log p(z|y;\theta)\tilde{p}(z|y;\varphi,\theta)\mu(dz)\,. \qquad (13)
\end{aligned}$$

We assume that

(H1) The functions

- $(y;\theta) \mapsto \nabla_\theta^2 \log g(y;\theta)$,
- $(y;\theta) \mapsto \log t(y;\varphi,\theta))$,
- $(y;\theta) \mapsto A(y;\varphi,\theta)$,
- $(y;\theta) \mapsto b(y;\varphi,\theta)$,
- $(y;\theta) \mapsto b(y;\varphi,\theta)(b(y;\varphi,\theta))^T$,

satisfy Wald-type conditions in θ_*.

We moreover assume that the model under consideration is regular and in particular that

(H2) $I_g(\theta) \triangleq -E(\nabla_\theta^2 \log g(Y,\theta))$ is positive definite at $\theta = \theta_*$.

Theorem 3 *Under the hypotheses of Theorems 1-2,* **(H1)** *and* **(H2)**,

$$\Gamma_{\varphi,n}(\mathbf{Y}_{1:n}) = \exp(n\delta(\varphi,\theta_*) + o(n))\left[I_g(\theta_*)^{-1}B(\varphi,\theta_*)I_g(\theta_*)^{-1} + o(1)\right] \quad w.p.1, \qquad (14)$$

where

$$\delta(\varphi,\theta_*) \triangleq E(\log t(Y;\varphi,\theta_*)) \geq 0, \qquad (15)$$

$$I_g(\theta_*) \triangleq -E\left(\nabla_\theta^2 \log g(Y;\theta_*)\right), \qquad (16)$$

and

$$B(\varphi,\theta_*) \triangleq E\left[b(Y;\varphi,\theta_*)\right]E\left[b(Y;\varphi,\theta_*)\right]^T.$$

If $E\left[b(Y;\varphi,\theta_*)\right]=0$*, then*

$$\Gamma_{\varphi,n}(\mathbf{Y}_{1:n}) \geq \exp(n\delta(\varphi,\theta_*)+o(n))\left[\frac{1}{n}I_g(\theta_*)^{-1}C(\varphi,\theta_*)I_g(\theta_*)^{-1}\right] \quad w.p.1,$$

where

$$C(\varphi,\theta_*) \triangleq E\left[A(Y;\varphi,\theta_*)-b(Y;\varphi,\theta_*)b(Y;\varphi,\theta_*)^T\right] \geq 0.$$

Using Jensen's inequality, the exponential rate $\delta(\varphi,\theta_*)$ introduced in Theorem 3 can be bounded from below by

$$\delta(\varphi,\theta_*) \geq E\left[K_p(Y;\varphi,\theta_*)\right],$$

where

$$K_p(Y;\alpha,\beta) \triangleq \int \log\frac{p(z|Y;\beta)}{p(z|Y;\alpha)}p(z|Y;\beta)\mu(dz),$$

is the Kullback divergence between the conditional distributions at α and β. Thus, except in cases where the conditional model is non-identifiable (in the sense that there exist values of $\varphi\neq\theta_*$ such that $K_p(Y;\varphi,\theta_*)$ is null a.e.), the limiting variance $\Gamma_{\varphi,n}(\mathbf{Y}_{1:n})$ is dominated by a factor which is, in the most favorable case, of order $\exp(\delta(\varphi,\theta_*)n)/n$. This result implies that the number of Monte Carlo simulations must increase exponentially fast for the variance to decrease with n in (3).

4.2 Consistency and asymptotic efficiency

In this section, we naturally extend the previous result to show that, under some additional assumptions, the MCML procedure is (strongly) consistent when s_n increases exponentially fast with n.

Theorem 4 *If the model is identifiable, under the hypotheses of Theorems 1-2,* **(H2)**, *and*

(H3) $(y,\theta)\mapsto\log g(y,\theta)$ *satisfies a Wald-type condition at* θ_*,

(H4) *The families* $\{\log g(y,\theta),\theta\in\Theta\}$, $\{\|\nabla_\theta\{\log g(y,\theta)\|,\theta\in\Theta\}$ *and* $\{\|\nabla^2_\theta\log g(y,\theta)\|,\theta\in\Theta\}$ *are dominated by integrable functions*

(H5) $\log p(z|y;\theta)$ *satisfies a Wald-type condition for all* $\theta\in\Theta$*, where the exponent* α *and the bounding function* $M()$ *defined in (17) may be chosen such that*

- α *and* $M()$ *do not depend* θ,
- *there exists* $\lambda>0$ *such that*

$$E\left[\log E_\theta(\mathrm{e}^{\lambda M(Z,Y)}|Y)\right]<\infty,$$

for all θ *in* Θ.

Then, if $s_n=\exp(n\gamma)$ *with* $\gamma>\delta(\varphi,\theta_*)$ *as in (15),* $\hat{\theta}^{s_n}_{\varphi,n}$ *converges to* θ_* *with probability one.*

Theorem 5 *Under the hypotheses of Theorem 4 and assuming that*

(H6) *The parametric functions*

- $\log t(y; \varphi, \theta)$,
- $\log \left[\int p(z|y; \theta) \mathrm{e}^{\lambda M(z,y)} \mu(dz)\right]$,
- $\log \left[\int \tilde{p}(z|y; \varphi, \theta) \mathrm{e}^{\lambda M(z,y)} \mu(dz)\right]$,

are dominated by integrable functions independent of θ,

$$\sqrt{n}(\hat{\theta}^{s_n}_{\varphi,n} - \theta_*) \xrightarrow{\mathcal{L}} N(0, I_g(\theta_*)^{-1})$$

Note that the condition $\gamma > \delta(\varphi, \theta_*)$ indeed imply that $\sqrt{n}(\hat{\theta}^{s_n}_{\varphi,n} - \hat{\theta}_n)$ tends to zero in probability (see appendix A.5), and thus the asymptotic efficiency of MCML simply follows as a consequence of the standard efficiency properties of the maximum likelihood estimator. The case where $\hat{\theta}^{s_n}_{\varphi,n} - \hat{\theta}_n$ is exactly of order $n^{-1/2}$, that is when MCML is $\sqrt{n}$-consistent but not necessarily asymptotically efficient, is somewhat artificial (remember that $\delta(\varphi, \theta_*)$ is not known in practice) and has not been investigated.

5 Asymptotic behavior of MCML under consistent initialization

The main message of Section 4 is that MCML, used with an arbitrary value of φ does not perform well for large sample sizes because the number of simulations has to be increased exponentially in order to counter the augmentation of the variance. However, (14) and (15) (see also (7) for the example of Section 3) suggest that s_n may be allowed to grow much more slowly if the parameter value φ used in the simulations stays "close enough to" θ_* (in the sense of the Kullback divergence). Except for the trivial case where $\varphi = \theta_*$, this requirement cannot hold when simulating from a single fixed value of φ as in Section 4. We thus consider in this section that a preliminary sequence φ_n of parameter estimates is available. That is, for a given sample size, we assume that the MCML algorithm is run from an estimate φ_n, rather than an arbitrary fixed value φ.

Our assumptions on this preliminary sequence of estimates are

(H7) The sequence $\{\varphi_n\}_{n\in\mathbb{N}}$ is independent of the observations $\{Y_n\}_{n\in\mathbb{N}}$ used for computing the MCML estimates, and satisfies $\varphi_n \to \theta_*$,

(H8) $\sqrt{n}\|\varphi_n - \theta_*\|$ is bounded from above.

As previously, the simulations $\{Z^k_{n,i}\}_{n\in\mathbb{N},1\leq i\leq n,1\leq k\leq s_n}$ are conditionally independent given the sequence $\{\varphi_n\}_{n\in\mathbb{N}}$ with $Z^k_{n,i}$ depending only upon φ_n. An interesting extension of **(H7)** would of course consist of allowing the sequence $\{\varphi_n\}_{n\in\mathbb{N}}$ to depend upon the observations (up to time index n). A closer look at the proofs in appendix A.5 however shows that such an extension is not tractable with the technique we are using. We thus focus on the simpler case of independent preliminary estimates.

We first show in Section 5.1 that **(H7)** is not sufficient and that **(H8)** is necessary to guarantee that the limiting conditional covariance matrix of Theorem 2 is bounded. We then show that the MCML algorithm initialized with φ_n is consistent for an arbitrary choice of s_n.

5.1 Asymptotic conditional covariance

We first provide an equivalent to Theorem 3 where the leading term is no longer exponential.

Theorem 6 *Under the hypotheses of Theorems 1–2,* **(H2)**, **(H7)**, *and assuming that*

(H9) *The functions*

- $(y;\varphi,\theta) \mapsto \nabla^2_\varphi \log t(y;\varphi,\theta)$,
- $(y;\varphi,\theta) \mapsto A(y;\varphi,\theta)$,
- $(y;\varphi,\theta) \mapsto \nabla_\varphi b(y;\varphi,\theta)$,
- $(y;\varphi,\theta) \mapsto b(y;\varphi,\theta)(b(y;\varphi,\theta))^T$,

satisfy Wald-type conditions at $(\varphi,\theta)=(\theta_*,\theta_*)$,

then

$$\Gamma_{\varphi_n,n}(\mathbf{Y}_{1:n}) = \exp\left(\xi_n^T\left[I_p(\theta_*)+o(1)\right]\xi_n\right)$$
$$\left\{\frac{1}{n}I_g(\theta_*)^{-1}\left[I_p(\theta_*)+o(1)\right]\left[I_p(\theta_*)^{-1}+\xi_n\xi_n^T\right]\left[I_p(\theta_*)+o(1)\right]I_g(\theta_*)^{-1}\right\},$$

where

$$\xi_n \triangleq \sqrt{n}(\hat{\theta}_n-\varphi_n),$$

and

$$I_p(\theta_*) \triangleq E\left[\int \nabla_\theta \log p(z|Y;\theta_*)\nabla_\theta \log p(z|Y;\theta_*)^T p(z|Y;\theta_*)\mu(dz)\right].$$

Interestingly enough, Theorem 6 indicates that the behavior of the limiting conditional covariance matrix $\Gamma_{\varphi_n,n}(\mathbf{Y}_{1:n})$ depends only on $\xi_n=\sqrt{n}(\hat{\theta}_n-\varphi_n)$ as n tends to infinity. As a consequence, the consistency of the φ_n's is not sufficient to guarantee satisfactory convergence properties for the MCML method, since φ_n must converge sufficiently fast, that is with a rate of at least $n^{-1/2}$. Indeed, **(H8)** together with the asymptotic normality of $\hat{\theta}_n$ and the assumption that $\{\varphi_n\}_{n\in\mathbb{N}}$ is independent of the observations imply that $\sqrt{n}(\hat{\theta}_n-\varphi_n)$ converges in distribution, and hence that $\Gamma_{\varphi_n,n}(\mathbf{Y}_{1:n})$ is an $O_p(1)$.

5.2 Consistency

The above remark implies the following result:

Theorem 7 *Under the hypotheses of Theorems 1, 2 and 4,* **(H7)**,

(H10) $(y,\varphi)\mapsto \log t(y;\varphi,\theta_*)$ *satisfies a Wald-type condition at* $\varphi=\theta_*$,

and assuming in addition that the model is identifiable, the MCMLE $\hat{\theta}^{s_n}_{\varphi_n,n}$ *converges to* θ_* *w.p.1.*

The proof of this theorem is similar to that of Theorem 4 in Section 4, the only difference being between Lemma 9 and Lemma 6, where the convergence of

$$\frac{1}{n}\log\left(\frac{1}{s_n}\sum_{k=1}^{s_n}\frac{p_n(\mathbf{Z}^k_{n,1:n}|\mathbf{Y}_{1:n};\theta_*)}{p_n(\mathbf{Z}^k_{n,1:n}|\mathbf{Y}_{1:n};\varphi_n)}\right)$$

to 0 is a consequence of φ_n converging to θ_* rather than of s_n diverging exponentially fast to infinity. Note that the result of Theorem 7 holds even when $s_n = C$, where C is any fixed integer. This generalizes the result observed in Section 3, where

$$\hat{\theta}^1_{\varphi_n,n} = (1-\rho^2)\hat{\theta}_n + \rho^2\varphi_n + \rho\sqrt{\frac{1-\rho^2}{n}}U_n, \qquad U_n \sim N(0,1),$$

with U_n independent from $\hat{\theta}_n$, which implies that $\hat{\theta}^1_{\varphi_n,n}$ is a $\sqrt{n}$-consistent estimate of θ_* under **(H7)** and **(H8)**.

More surprisingly, Theorem 7 does not rely on **(H8)**. This counterintuitive result follows from φ_n being a consistent estimate of θ_*. However, Theorem 6 as well as the example of Section 3 suggests that **(H8)** is indeed necessary when considering the rate of convergence of $\hat{\theta}^{s_n}_{\varphi_n,n}$ to θ_*. At this point, however, we cannot extend Theorem 5 when φ_n depends on n.

6 Conclusion

We have presented results which demonstrate that the MCML method suffers from severe drawbacks in terms of robustness to the choice of the parameter value φ used for simulating the latent variables. The fact that the variance of the MCML estimator increases exponentially fast with the sample size n implies that the validity of the approximation of the likelihood function and in particular of the maximum likelihood estimate are clearly restricted to small values of n, for given values of s_n. Asymptotically the relevance of the method can only be argued in cases where the importance value φ is a consistent estimate of θ_*. In practice, MCML should thus be used in conjunction with another consistent maximum likelihood estimation method, as suggested by Geyer (1996) and Billio *et al.* (1998), like noninformative Bayes estimators. Moreover, this study does not shed any light on the proposal of iterative MCML of Geyer (1996), where the solution of one MCML run is used as the reference value φ for the next MCML run.

More generally, these results suggest that simulation based numerical optimization (or at least stochastic approximation in numerical optimization) can hardly be carried out without somehow restricting the range of plausible values of θ as the sample size increases. Therefore, nesting the maximization (or parameter search) stage and the latent variable simulation stage within one another seems to some extent unavoidable for this type of method.

Appendix

A Proofs of Section 4

Before considering Theorems 3-5, we first state two technical lemmas which are used repeatedly in the sequel.

A.1 Wald-type condition

Definition 1 (Wald-type condition) *Let $\psi : (\mathbb{R}^p \times \Theta \to \mathbb{R}^q)$ denote an integrable parameterized function. ψ satisfies a Wald-type condition at θ, if*

- $E\|\psi(Y;\theta)\| < \infty$,
- *There exist $\rho > 0$ and $\alpha > 0$ such that*

$$\sup_{\|\eta-\theta\|\leq\rho} \frac{\|\psi(y;\eta) - \psi(y;\theta)\|}{\|\eta-\theta\|^\alpha} \leq M(y) \text{ for all } y, \tag{17}$$

where $M(y)$ is a positive Borel function such that $E(M(Y)) < \infty$.

Lemma 1 *Assume that $\psi : (\mathbb{R}^p \times \Theta \to \mathbb{R}^q)$ satisfies a Wald-type condition at θ and let $\{\theta_n\}_{n\geq 0}$ denote a sequence such that $\lim_{n\to\infty} \theta_n = \theta$ w.p.1. Then,*

$$\sum_{i=1}^{n} \psi(Y_i;\theta_n) = nE\left\{\psi(Y;\theta)\right\} + o(n) \quad w.p.1.$$

A.2 Conditional Borel-Cantelli Lemma

Lemma 2 *Let F_n denote a family of Borel functions,*

$$\sum_{n=1}^{+\infty} P_\varphi\left(F_n(\mathbf{Z}_{n,1:n}^{1:s_n}, \mathbf{Y}_{1:n}) \in B \middle| \mathbf{Y}_{1:n}\right) < +\infty$$

implies that, w.p.1, $F_n(\mathbf{Z}_{n,1:n}^{1:s_n}, \mathbf{Y}_{1:n}) \in B^c$ for sufficiently large n's.

This lemma is a simple consequence of the remarks that $Z_{n,i}^k$ and $Z_{n',i'}^{k'}$ are conditionally independent given $\{Y_n\}_{n\in\mathbb{N}}$ whenever $(n,i,k) \neq (n',i',k')$, and that $Z_{n,i}^k$ depends only upon Y_i.

A.3 Asymptotic behavior of the limiting covariance of MCML estimates

The following result is needed in the proof of Theorem 3.

Proposition 1 *Under the assumptions of Theorems 1–2 and* **(H1)**,

$$D_n(\mathbf{Y}_{1:n}) = nI_g(\theta_*) + o(n), \tag{18}$$

almost surely, where $I_g(\theta_*)$ is the Fisher information matrix defined by (16).

Proof. Since $\hat{\theta}_n$ minimizes $g_n(\mathbf{Y}_{1:n};\theta)$ in a point which belongs to the interior of Θ,

$$D_n(\mathbf{Y}_{1:n}) = -\nabla_\theta^2 \log g_n(\mathbf{Y}_{1:n}, \hat{\theta}_n) = -\sum_{i=1}^{n} \nabla_\theta^2 \log g(Y_i;\hat{\theta}_n). \tag{19}$$

Lemma 1 along with **(H1)** complete the proof. □

The proof of Theorem 3 then goes as follows:

Proof. (Theorem 3) The variance $V_{\varphi,n}(\mathbf{Y}_{1:n})$, defined in (4), can be rewritten as

$$\begin{aligned} V_{\varphi,n}(\mathbf{Y}_{1:n}) &= \prod_{i=1}^{n} t(Y_i;\varphi,\hat{\theta}_n) \\ &\times \left[\sum_{i=1}^{n} A(Y_i;\varphi,\hat{\theta}_n) - \sum_{i=1}^{n} b(Y_i;\varphi,\hat{\theta}_n) b(Y_i;\varphi,\hat{\theta}_n)^T \right. \\ &+ \left. (\sum_{i=1}^{n} b(Y_i;\varphi,\hat{\theta}_n))(\sum_{j=1}^{n} b(Y_j;\varphi,\hat{\theta}_n))^T \right]. \end{aligned} \tag{20}$$

The product $\prod_{i=1}^{n} t(Y_i;\varphi,\hat{\theta}_n)$ can be rewritten as

$$\exp \sum_{i=1}^{n} \log t(Y_i;\varphi,\hat{\theta}_n).$$

The result of Theorem 3 is then obtained by applications of Lemma 1 for the functions defined in **(H1)**. In the particular case where $E(b(Y;\varphi,\theta_*)) = 0$, the term between brackets in (20) can be bounded from below by its first two terms which are of order n. Finally, $C(\varphi,\theta_*)$ is easily seen to be positive since $A(Y;\varphi,\theta_*) - b(Y;\varphi,\theta_*)b(Y;\varphi,\theta_*)^T$ is the covariance matrix of $\nabla_\theta^T \log p(Z|Y;\theta)$ under the probability measure $\tilde{p}(z|y;\varphi,\theta)\mu(dz)$. □

A.4 Consistency of MCML

Denote

$$Q_n(\theta) \triangleq \frac{\hat{l}_{\varphi,n}^{s_n}(\theta)}{l_{\varphi,n}(\theta)}, \tag{21}$$

and

$$T_n(\theta) \triangleq \frac{1}{n} \log Q_n(\theta). \tag{22}$$

Assumption **(H3)** implies that $\frac{\log l_{\varphi,n}(\theta)}{n}$ converges w.p.1 to

$$L(\theta) \triangleq -K_g(\theta,\theta_*) + K_g(\theta_*,\varphi), \tag{23}$$

where $K_g(\alpha,\beta)$ denotes the Kullback divergence between α and β. For an identifiable model, θ_* is the unique minimizer of $K_g(\theta,\theta_*)$. The proof of Theorem 4 thus proceeds as follows: first, we show that $T_n(\theta)$ is bounded from above by 0 (w.p.1) uniformly in Θ for sufficiently large values of n (Lemma 5), and, second, we show that $T_n(\theta_*)$ converges to 0 w.p.1.

Lemma 3 *For any θ in Θ,*

$$\overline{\lim_{n\to\infty}} T_n(\theta) \le 0 \qquad \textit{w.p.1.}$$

Proof.

$$Q_n(\theta) = \frac{1}{s_n} \sum_{k=1}^{s_n} \frac{p_n(\mathbf{Z}^k_{n,1:n}|\mathbf{Y}_{1:n};\theta)}{p_n(\mathbf{Z}^k_{n,1:n}|\mathbf{Y}_{1:n};\varphi)},$$

and thus $E_\varphi(Q_n(\theta)|\mathbf{Y}_{1:n}) = 1$. As a consequence, $P_\varphi(Q_n(\theta)/(n(\log n)^{1+c})|\mathbf{Y}_{1:n})$ is summable for any $c > 0$. Application of the conditional Borel-Cantelli Lemma then shows that $Q_n(\theta)/n(\log n)^{1+c} = o(1)$ w.p.1 and hence that $\overline{\lim}_{n\to\infty} T_n(\theta) \leq 0$ w.p.1. □

The following lemma ensures that under some additional regularity conditions on $p(z|y;\theta)$, $T_n()$ can be bounded from above by an arbitrary positive constant, uniformly in an open neighborhood of θ.

Lemma 4 *Under* **(H5)**, *for all $\theta \in \Theta$ and all $\epsilon > 0$, there exist $\eta_{\theta,\epsilon} > 0$ and $N_{\theta,\epsilon} \in \mathbb{N}$ such that for all $n \geq N_{\theta,\epsilon}$,*

$$\sup_{\theta' \in B(\theta,\eta_{\theta,\epsilon})} T_n(\theta') < \epsilon. \tag{24}$$

Proof. Let ϵ be a strictly positive real number, and θ an arbitrary point of Θ. For sufficiently small values of η, **(H5)** implies

$$\forall (z,y) \in \mathbb{R}^2,\ \frac{p(z|y;\theta')}{p(z|y;\theta)} \leq \exp\left(\eta^\alpha M(z,y)\right), \tag{25}$$

for any $\theta' \in B(\theta,\eta)$, the open ball of radius η centered in θ. Then,

$$T_n(\theta') \leq \frac{1}{n} \log \left[\frac{1}{s_n} \sum_{k=1}^{s_n} \prod_{i=1}^{n} \frac{p(Z^k_{n,i}|Y_i;\theta)}{p(Z^k_{n,i}|Y_i;\varphi)} \exp\left(\eta^\alpha M(Z^k_i, Y_i)\right) \right]. \tag{26}$$

Denoting by $D_n(\theta,\eta)$ the term between brackets in (26),

$$\frac{1}{n} \log D_n(\theta,\eta) = \frac{1}{n} \log \frac{D_n(\theta,\eta)}{E_\theta(D_n(\theta,\eta)|\mathbf{Y}_{1:n})} + \frac{1}{n} \log E_\theta(D_n(\theta,\eta)|\mathbf{Y}_{1:n}). \tag{27}$$

The first term in the r.h.s. of (27) can be shown to verify

$$\overline{\lim_{n\to\infty}} \frac{1}{n} \log \frac{D_n(\theta,\eta_{\theta,\epsilon})}{E_\theta(D_n(\theta,\eta_{\theta,\epsilon})|\mathbf{Y}_{1:n})} \leq 0,$$

proceeding as in the proof of Lemma 3. The second term in the r.h.s. of (27) writes

$$\begin{aligned} \frac{1}{n} \log E_\theta(D_n(\theta,\eta)|\mathbf{Y}_{1:n}) &= \frac{1}{n} \sum_{i=1}^{n} \log E_\theta(\exp(\eta^\alpha M(Z,Y_i))|Y_i), \\ &= E\left[\log E_\theta(\exp(\eta^\alpha M(Z,Y))|Y)\right] + o(1), \end{aligned}$$

that is, converges to 0 as $\eta \to 0$. □

The proof of the following lemma is omitted since it is a direct corollary of Lemma 4 under compactness of Θ.

Lemma 5 *Under assumptions* **(H5)** *and* **(H7)**, $\overline{\lim}_{n\to\infty} \sup_{\theta\in\Theta} T_n(\theta) \leq 0$ *w.p.1.*

Lemma 6 $T_n(\theta_*) \to 0$ *w.p.1.*

Proof. In the course of proving Lemma 3, we have already seen that $E_\varphi(Q_n(\theta_*)|\mathbf{Y}_{1:n}) = 1$, moreover

$$
\begin{aligned}
Var_\varphi[Q_n(\theta_*)|\mathbf{Y}_{1:n}] &\leq \frac{1}{s_n}\int \frac{p_n^2(\mathbf{Z}_{n,1:n}|\mathbf{Y}_{1:n};\theta_*)}{p_n(\mathbf{Z}_{n,1:n}|\mathbf{Y}_{1:n};\varphi)}\mu_n(\mathbf{Z}_{n,1:n}),\\
&\leq \frac{1}{s_n}\exp\left(\sum_{i=1}^n \log t(Y_i;\varphi,\theta_*)\right), \qquad (28)
\end{aligned}
$$

where $t()$ was defined in (13). For $s_n = \exp(n\gamma)$, **(H1)** implies that the upper bound in (28) is summable, and hence that $Q_n(\theta_*) \to 1$ w.p.1 by application of the conditional Borel-Cantelli Lemma. □

Proof. (Theorem 4) First note that $\frac{1}{n}\log l_{\varphi,n}(\theta) \to L(\theta)$ where $L(\theta)$ is defined by (23). Now, by definition of $\hat{\theta}^{s_n}_{\varphi,n}$

$$
\frac{1}{n}\log \hat{l}^{s_n}_{\varphi,n}(\hat{\theta}^{s_n}_{\varphi,n}) \geq \frac{1}{n}\log \hat{l}^{s_n}_{\varphi,n}(\theta_*), \qquad (29)
$$

which is equivalent to

$$
\begin{aligned}
L(\hat{\theta}^{s_n}_{\varphi,n}) \geq L(\theta_*) &- L(\theta_*) + \frac{1}{n}\log l_{\varphi,n}(\theta_*) - \frac{1}{n}\log l_{\varphi,n}(\theta_*) + \frac{1}{n}\log \hat{l}^{s_n}_{\varphi,n}(\theta_*)\\
&+ L(\hat{\theta}^{s_n}_{\varphi,n}) - \frac{1}{n}\log l_{\varphi,n}(\hat{\theta}^{s_n}_{\varphi,n}) + \frac{1}{n}\log l_{\varphi,n}(\hat{\theta}^{s_n}_{\varphi,n}) - \frac{1}{n}\log \hat{l}^{s_n}_{\varphi,n}(\hat{\theta}^{s_n}_{\varphi,n}). \qquad (30)
\end{aligned}
$$

Thus,

$$
L(\hat{\theta}^{s_n}_{\varphi,n}) \geq L(\theta_*) - 2\sup_{\theta\in\Theta}|L(\theta) - \frac{1}{n}\log l_{\varphi,n}(\theta)| + T_n(\theta_*) - \sup_{\theta\in\Theta} T_n(\theta). \qquad (31)
$$

Lemmas 5 and 6 then show that $L(\theta^{s_n}_{\varphi,n}) \xrightarrow[n\to\infty]{} L(\theta_*)$ by application of the uniform strong law of large numbers. □

A.5 Asymptotic efficiency of MCML

From the proof of Lemma 6, we know that $Q_n(\theta_*) \to 1$ w.p.1. In order to prove theorem 5 however, a much stronger version of the same result is needed:

Lemma 7 *There exists a compact neighborhood K of θ_* such that*

$$
\sup_{\theta\in K}|Q_n(\theta) - 1| \to 0 \quad \textit{w.p.1}
$$

Proof. By continuity of $\theta \mapsto \delta(\varphi,\theta)$, there exists a closed ball K centered in θ_* such that for all $\theta \in K$, $\delta(\varphi,\theta) < \gamma' < \gamma$. Let $\mathcal{G}_n$ denote a ρ_n-net covering K and $\{\theta_j\}_{1\leq j\leq \#\{\mathcal{G}_n\}}$ the associated grid points. The grid spacing ρ_n is set to $\rho_n = n^{-2/\alpha}$ where α is defined in **(H5)**. $\mathcal{G}_n$ is set so as to minimize the number of grid points while covering K, and thus $\#\{\mathcal{G}_n\} \equiv \rho_n^{-d}$ where d is the dimension of the parameter vector θ.

$$
P_\varphi\left[\sup_{\theta\in K}|Q_n(\theta) - 1| \geq \epsilon \,\middle|\, \mathbf{Y}_{1:n}\right] \leq \#\{\mathcal{G}_n\}\max_{\theta_j\in\mathcal{G}_n} P_\varphi\left[\sup_{\theta\in B(\theta_j,\rho_n)}|Q_n(\theta) - 1| \geq \epsilon \,\middle|\, \mathbf{Y}_{1:n}\right], \qquad (32)
$$

where $B(\theta_j, \rho_n)$ denotes the ball of radius ρ_n centered in θ_j. Applying **(H5)** in θ_j (see the proof of Lemma 4) yields,

$$Q^-_{n,j} \le Q_n(\theta) \le Q^+_{n,j} \quad \text{for } \theta \in B(\theta_j, \rho_n), \tag{33}$$

where

$$Q^-_{n,j} = \frac{1}{s_n} \sum_{k=1}^{s_n} \prod_{i=1}^{n} \frac{p(Z^k_{n,i}|Y_i; \theta_j)}{p(Z^k_{n,i}|Y_i; \varphi)} \mathrm{e}^{-\rho_n^\alpha M(Z^k_{n,i}, Y_i)},$$

and

$$Q^+_{n,j} = \frac{1}{s_n} \sum_{k=1}^{s_n} \prod_{i=1}^{n} \frac{p(Z^k_{n,i}|Y_i; \theta_j)}{p(Z^k_{n,i}|Y_i; \varphi)} \mathrm{e}^{\rho_n^\alpha M(Z^k_{n,i}, Y_i)}.$$

We now consider the behavior of $|Q^+_{n,j} - 1|$ in more details (identical results are obtained for $|Q^-_{n,j} - 1|$):

$$E_\varphi \left[Q^+_{n,j} \middle| \mathbf{Y}_{1:n} \right] = \prod_{i=1}^{n} \int p(z|Y_i; \theta_j) \mathrm{e}^{\rho_n^\alpha M(z, Y_i)} \mu(dz).$$

Applications of Jensen's inequality yield,

$$1 \le E_\varphi \left[Q^+_{n,j} \middle| \mathbf{Y}_{1:n} \right] \le \exp \left(\frac{n\rho_n^\alpha}{\lambda} \frac{1}{n} \sum_{i=1}^{n} \log \left[\int p(z|Y_i; \theta_j) \mathrm{e}^{\lambda M(z, Y_i)} \mu(dz) \right] \right), \tag{34}$$

where λ is the constant defined in **(H5)**. By application of the uniform law of large numbers, the upper bound in (34) is equivalent to

$$\exp \left(\frac{n\rho_n^\alpha}{\lambda} E \left[\log E_{\theta_j} \left(\mathrm{e}^{\lambda M(Z,Y)} \middle| Y \right) \right] \right),$$

and thus converges to 1, uniformly on K, as a consequence of the choice of the grid spacing ρ_n and of **(H5)**-**(H6)**.

Thus, for sufficiently large values of n,

$$\begin{aligned} P_\varphi \left[|Q^+_{n,j} - 1| \ge \epsilon \middle| \mathbf{Y}_{1:n} \right] &\le P_\varphi \left[\left| Q^+_{n,j} - E_\varphi \left[Q^+_{n,j} \middle| \mathbf{Y}_{1:n} \right] \right| \ge \epsilon' \middle| \mathbf{Y}_{1:n} \right] \\ &\le \frac{1}{\epsilon'^2} Var_\varphi \left[Q^+_{n,j} \middle| \mathbf{Y}_{1:n} \right], \end{aligned}$$

where ϵ' is any positive real number smaller than ϵ. Now,

$$Var_\varphi \left[Q^+_{n,j} \middle| \mathbf{Y}_{1:n} \right] \le \frac{\prod_{i=1}^n t(Y_i; \varphi, \theta_j)}{s_n} \prod_{i=1}^{n} \int \tilde{p}(z|Y_i; \varphi, \theta_j) \mathrm{e}^{\rho_n^\alpha M(z, Y_i)} \mu(dz), \tag{35}$$

where the function t and $\tilde{p}$ are as defined in (13). Proceeding as in the case of $E_\varphi \left[Q^+_{n,j} \middle| \mathbf{Y}_{1:n} \right]$, one obtains

$$1 \le \prod_{i=1}^{n} \int \tilde{p}(z|Y_i; \varphi, \theta_j) \mathrm{e}^{\rho_n^\alpha M(z, Y_i)} \mu(dz) \le \exp \left(\frac{n\rho_n^\alpha}{\lambda} \frac{1}{n} \sum_{i=1}^{n} \log \left[\int \tilde{p}(z|Y_i; \varphi, \theta_j) \mathrm{e}^{\lambda M(z, Y_i)} \mu(dz) \right] \right),$$

and thus the rightmost term of (35) converges uniformly to 1. Hence, for n sufficiently large,

$$Var_{\varphi}\left[Q_{n,j}^{+}\middle|\mathbf{Y}_{1:n}\right]=O\left(\frac{1}{s_n}\exp\left[n\left(\frac{1}{n}\sum_{i=1}^{n}\log t(Y_i;\varphi,\theta_j)\right)\right]\right).$$

From the uniform law of large numbers, $\frac{1}{n}\sum_{i=1}^{n}\log t(Y_i;\varphi,\theta_j)$ converges to $\delta(\varphi,\theta_j)$, and thus $Var_{\varphi}\left[Q_{n,j}^{+}\middle|\mathbf{Y}_{1:n}\right]=O(\mathrm{e}^{-n(\gamma-\gamma')})$ where $\gamma>\gamma'$.

Using analog results concerning $Q_{n,j}^{-}$, (32) implies that

$$P_{\varphi}\left[\sup_{\theta\in K}|Q_n(\theta)-1|\geq\epsilon\middle|\mathbf{Y}_{1:n}\right]=O(n^{2d/\alpha}\mathrm{e}^{-n(\gamma-\gamma')}),$$

which is summable as required. □

Proof. (Theorem 5) Denote $\epsilon_n\triangleq\sqrt{n}(\hat{\theta}_{\varphi,n}^{s_n}-\hat{\theta}_n)$. We will show that $\epsilon_n\to 0$ w.p.1, which is sufficient to prove Theorem 5 since

$$\sqrt{n}(\hat{\theta}_{\varphi,n}^{s_n}-\theta_*)=\epsilon_n+\sqrt{n}(\hat{\theta}_n-\theta_*).$$

Since $\hat{\theta}_{\varphi,n}^{s_n}$ is the maximizer of $\hat{l}_{\varphi,n}^{s_n}(\theta)$,

$$\log\hat{l}_{\varphi,n}^{s_n}(\hat{\theta}_{\varphi,n}^{s_n})\geq\log\hat{l}_{\varphi,n}^{s_n}(\hat{\theta}_n).$$

Equivalently,

$$\begin{aligned}\log Q_n(\hat{\theta}_{\varphi,n}^{s_n})-\log Q_n(\hat{\theta}_n) &\geq \log l_{\varphi,n}(\hat{\theta}_n)-\log l_{\varphi,n}(\hat{\theta}_{\varphi,n}^{s_n}),\\ &\geq -\frac{1}{n}\epsilon_n^T\nabla_\theta^2\log l_{\varphi,n}(t_n\hat{\theta}_n+(1-t_n)\hat{\theta}_{\varphi,n}^{s_n})\epsilon_n,\end{aligned}\tag{36}$$

for some constant t_n in $[0,1]$. Since $(y,\theta)\mapsto\nabla_\theta^2\log g(y,\theta)$ satisfies a Wald type condition at θ_*,

$$-\frac{1}{n}\nabla_\theta^2\log l_{\varphi,n}(t_n\hat{\theta}_n+(1-t_n)\hat{\theta}_{\varphi,n}^{s_n})\to I_g(\theta_*).$$

Because $I_g(\theta_*)$ is positive definite, there exists $M>0$ such that, for n sufficiently large,

$$\log Q_n(\hat{\theta}_{\varphi,n}^{s_n})-\log Q_n(\hat{\theta}_n)\geq M\|\epsilon_n\|^2.\tag{37}$$

The proof is completed by application of Lemma 7. □

B Proofs of Section 5

B.1 Asymptotic behavior of the limiting covariance

Proof. (Theorem 6) Starting from $V_{\varphi,n}(\mathbf{Y}_{1:n})$ as given by (20), where the functions t, A and b are defined in (13), repeated applications of Lemma 1 for the functions defined in **(H9)** yield

$$\prod_{i=1}^{n}t(Y_i;\varphi_n,\hat{\theta}_n)=\exp\left\{\xi_n^T\left[E\left(\nabla_\varphi^2\log t(Y_i;\theta_*,\theta_*)\right)+o(1)\right]\xi_n\right\},$$

$$\sum_{i=1}^{n} A(Y_i;\varphi,\hat{\theta}_n) = n\left[E\left(A(Y;\theta_*,\theta_*)\right) + o(1)\right],$$

$$\sum_{i=1}^{n} b(Y_i;\varphi,\hat{\theta}_n) b(Y_i;\varphi,\hat{\theta}_n)^T = n\left\{E\left[b(Y;\theta_*,\theta_*)b(Y;\theta_*,\theta_*)^T\right] + o(1)\right\},$$

and

$$\left[\sum_{i=1}^{n} b(Y_i;\varphi,\hat{\theta}_n)\right]\left[\sum_{j=1}^{n} b(Y_j;\varphi,\hat{\theta}_n)\right]^T =$$
$$n\left\{E\left[\nabla_\varphi b(Y;\theta_*,\theta_*)\right] + o(1)\right\}\xi_n\xi_n^T\left\{E\left[\nabla_\varphi b(Y;\theta_*,\theta_*)\right] + o(1)\right\}^T.$$

Theorem 6 follows from

$$\begin{aligned}
E\left[\nabla_\varphi^2 \log t(Y_i;\theta_*,\theta_*)\right] &= 2I_p(\theta_*),\\
E\left[A(Y;\theta_*,\theta_*)\right] &= I_p(\theta_*),\\
E\left[b(Y;\theta_*,\theta_*)b(Y;\theta_*,\theta_*)^T\right] &= 0,\\
E\left[\nabla_\varphi b(Y;\theta_*,\theta_*)\right] &= I_p(\theta_*).
\end{aligned}$$

□

B.2 Consistency

The proofs for this section closely follow those of Section A.4 where the notation introduced in (22) now stands for

$$T_n(\theta) = \frac{1}{n}\left(\log \hat{l}^{s_n}_{\varphi_n,n}(\theta) - \log l_{\varphi_n,n}(\theta)\right).$$

Lemmas 5 and 6 of Section A.4 are now to be replaced respectively by Lemmas 8 and 9. Only the proof of Lemma 9 is given because it significantly differs from that of Lemma 6.

Lemma 8 *Under* **(H7)**, $\overline{\lim}_{n\to\infty} \sup_{\theta\in\Theta} T_n(\theta) \le 0$ *w.p.1.*

Lemma 9 *Under* **(H7)** *and* **(H10)**, $T_n(\theta_*) \to 0$ *w.p.1.*

Proof. Denoting

$$R_n \triangleq \left(\prod_{k=1}^{s_n} \frac{p_n(\mathbf{Z}^k_{n,1:n}|\mathbf{Y}_{1:n};\varphi_n)}{p_n(\mathbf{Z}^k_{n,1:n}|\mathbf{Y}_{1:n};\theta_*)}\right) \Bigg/ \left(E_{\varphi_n}\left(\left.\frac{p_n(\mathbf{Z}^k_{n,1:n}|\mathbf{Y}_{1:n};\varphi_n)}{p_n(\mathbf{Z}^k_{n,1:n}|\mathbf{Y}_{1:n};\theta_*)}\right| \mathbf{Y}_{1:n}\right)\right)^{s_n},$$

one obtains

$$T_n(\theta_*) \ge -\frac{1}{n}\log(R_n)^{\frac{1}{s_n}} - \frac{1}{n}\sum_{i=1}^{n} \log E_{\varphi_n}\left(\left.\frac{p(Z|Y_i;\varphi_n)}{p(Z|Y_i;\theta_*)}\right| Y_i\right). \tag{38}$$

Moreover,

$$\underline{\lim}_{n\to\infty} -\frac{1}{n}\log R_n^{\frac{1}{s_n}} \ge 0,$$

since $E(R_n) = 1$ (see the proof of Lemma 3). The second term in the r.h.s. of (38) writes

$$\frac{1}{n}\sum_{i=1}^{n}\log E_{\varphi_n}\left(\left.\frac{p(Z|Y_i;\varphi_n)}{p(Z|Y_i;\theta_*)}\right| Y_i\right) = \frac{1}{n}\sum_{i=1}^{n}\log t(Y_i;\varphi_n,\theta_*).$$

Thus, $\varphi_n \to \theta_*$ and **(H10)** imply that the second term of (38) converges with probability one to $E(\log t(Y;\theta_*,\theta_*)) = 0$. Hence,

$$\underline{\lim}_{n\to\infty} T_n(\theta_*) \geq 0\,,$$

which, together with Lemma 8, completes the proof. □

References

Billio, M., Monfort, A., Robert, C.P. (1998). The simulated likelihood method. Tech. report 9821, CREST, INSEE, Paris.

Danielson, J., Richard, J.F. (1993). Quadratic acceleration for simulated maximum likelihood evaluation. *J. Applied Econometrics*, **8**, 153-173.

Celeux, G., Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Quart.*, **2**, 73-82.

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

Geweke, J. (1988) Antithetic acceleration of Monte Carlo integration in Bayesian inference. *J. Econometrics* **38**, 73-90.

Gu, M.G., Kong, F.H. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proc. National Academy of Sciences* **95**(13).

Geyer, C.J., Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data, (with discussion). *J. Roy. Statist. Soc.* (Ser. B), **54** 657-699.

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

Geyer, C.J. (1996). Estimation and optimization of functions. In *Markov Chain Monte Carlo in Practice*, (W. R. Gilks, S. Richardson and D. J. Spiegelhalter eds.). Chapman and Hall, London.

Gouriéroux, C., Monfort, A. (1993) Simulation-based inference: A survey with special reference to panel data models. *J. Econometrics*, **59**, 5-33.

Ip, E. H. S. (1994) A stochastic EM estimator in the presence of missing data - Theory and applications. Technical report #304, Department of Statistics, Stanford University.

Lavielle, M., Delyon, B., Moulines., E. (1999) On a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, **27** 94-128.

Lee, L.-F. (1995). Statistical inference with simulated likelihood functions. Working Paper No. 96/1, Hong Kong University of Science and Technology, Department of Economics.

Nielsen, S. F. (2000) The stochastic EM algorithm: Estimation and asymptotic results. To appear in *Bernoulli*.

Sandmann, G., Koopman, S.J. (1998). Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *J. Econometrics*, **87**(2), 271 - 301.

Thompson, E.A. (1994). Monte Carlo likelihood in genetic mapping. *Statistical Science*, **9**(3), 355–366.

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

Younes, L. (1988). Estimation and annealing for Gibbsian fields. *Annales de l'Institut Henri Poincaré*, **24**(2) 269-294.

Corresponding author

Olivier Cappé
ENST, Dpt. TSI
46 rue Barrault, 75634 Paris cedex 13, France
tel +33 1 45 81 71 11 *fax* +33 1 45 88 79 35 *email* `cappe@tsi.enst.fr`

List of Figures

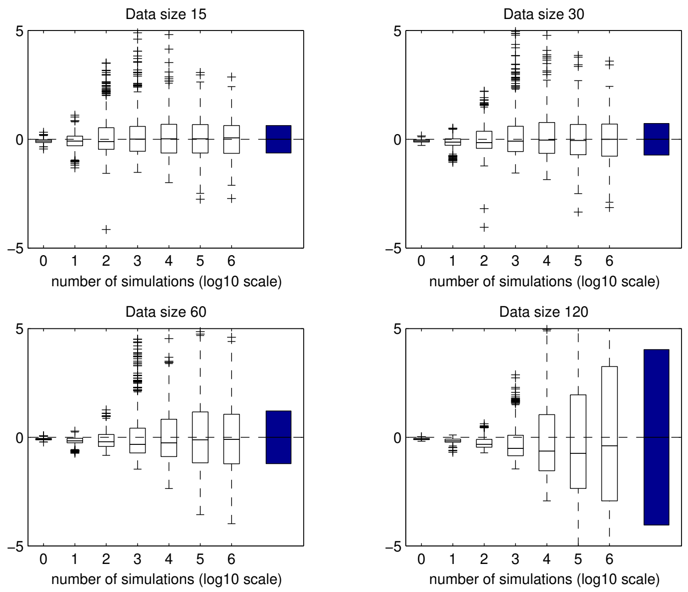

Figure 1: MCML estimates, recentered around the maximum likelihood estimate and rescaled by $s^{-1/2}$, as a function of the number of iterations s (on a log scale) for different numbers of observation (500 runs of the algorithm, $\rho = 0.9$, $\varphi - \hat{\theta}_n = -0.1$). The black box on the right features the quartiles corresponding to the asymptotic normal approximation.

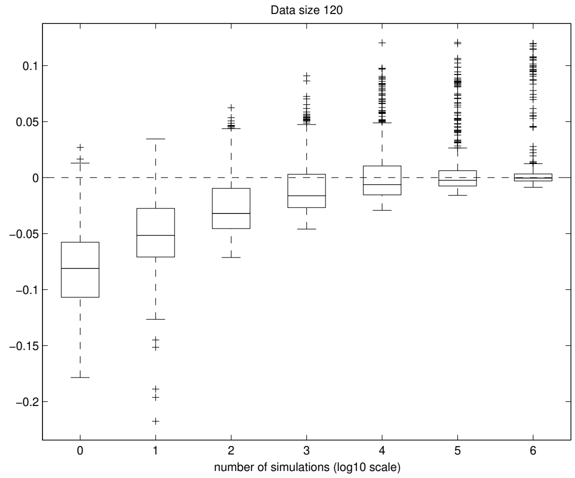

Figure 2: Unscaled MCML estimates, recentered around the maximum likelihood estimate, as a function of the number of iterations (500 runs of the algorithm, $\rho = 0.9$, $\varphi - \hat{\theta}_n = -0.1$, 120 observations).