A Metropolis version of the EM algorithm

Carlo Gaetan[*]
Dipartimento di Scienze Statistiche, Università di Padova
and
Jian-Feng Yao
IRMAR, Université de Rennes 1

1st February 2002

Abstract

The Expectation Maximisation (EM) algorithm is a popular technique for maximum likelihood in incomplete data models. In order to overcome its documented limitations, several stochastic variants are proposed in the literature. However, none of these algorithms is guaranteed to provide a global maximizer of the likelihood function. In this paper we introduce the MEM algorithm — a Metropolis version of the EM — that can achieve global maximisation of the likelihood.

Keywords: Incomplete data, Simulated annealing, Mixed model.

1 Introduction

The Expectation Maximisation (EM) algorithm (Dempster *et al.*, 1977) is a general iterative method finding the maximum likelihood estimate of the parameters for an underlying distribution from a given data set when data are incomplete. Roughly speaking there are two main situations where the EM algorithm is of central importance. The first occurs when the process we are interested in can not be directly observed and indeed the observed data originates from a non-invertible distortion of that process (missing values, censored data, hidden Markov models). The second occurs when the likelihood function is analytically or numerically intractable but can be thought as the marginal of a data-augmented (thus larger) model which has a much simpler (though unobservable) likelihood function.

Let us briefly recall the algorithm. We are given a parametric family of distributions (P_θ) and the observable vector Y is part of a so-called complete vector $X = (Y, Z)$. Both Y and X have a density function, say $g(y;\theta)$ and $f(x;\theta)$ respectively, with respect to some σ-finite measure dy and dx on the corresponding spaces. Here, θ is a parameter belonging to some subset Θ of the Euclidean space $\mathbb{R}^p$. Let y be the observed data. The objective is to compute the maximum likelihood estimator $\hat{\theta} = \text{argmax}_{\theta\in\Theta} g(y;\theta)$. The EM algorithm maximizes $g(y;\theta)$ by iteratively

[*] Corresponding author. Dipartimento di Scienze Statistiche, Università di Padova, via Battisti, 241 I-35121 Padova (Italy). gaetan@stat.unipd.it

maximizing the conditional expectation of the logarithm of the complete data density, $\log f(x;\theta)$, given y and a present value θ'. More precisely, each iteration of the EM algorithm is decomposed into two steps: an E-step and M-step. At iteration k, given the current estimate θ_{k-1}, the E-step consists in calculating

$$S(\theta, \theta_{k-1}) = E_{\theta_{k-1}}(\log(f(X;\theta)|Y = y) \ , \tag{1·1}$$

where the expectation $E_\theta(\cdot|Y = y)$ is the conditional expectation to $Y = y$ under the probability distribution P_θ (the associated conditional density will be denoted by $h(\cdot|y;\theta)$). The M-step consists in finding

$$\theta_k = \text{argmax}_{\theta\in\Theta} S(\theta, \theta_{k-1}) \ . \tag{1·2}$$

The algorithm repeats these two steps until convergence is reached.

A detailed account of the convergence properties of the sequence (θ_k) can be found in Dempster *et al.* (1977), Wu (1983) and a more recent review is given by Meng and van Dyk (1997). The wide-spread popularity of the EM is largely due to its "monotonicity": the sequence $(g(y;\theta_k))$ is increasing and, under suitable regularity conditions, (θ_k) converges to a stationary point of $l(\theta)$. Note that the monotonicity is guaranteed even for the Generalized EM (GEM) algorithm (Dempster *et al.*, 1977) when θ_k is chosen such that

$$S(\theta_k, \theta_{k-1}) \geq S(\theta_{k-1}, \theta_{k-1}).$$

However, despite these appealing features, the EM algorithm has several limitations. First, it is only a local and deterministic maximizer of the likelihood and its asymptotic behaviour depends heavily on the starting values used. More seriously, apart from some simple models, the algorithm is far from easy to set up: namely the E-step as well as the M-step — or both steps in even worse situations —, can be intractable or numerically inefficient. In order to circumvent these drawbacks, various improvements have been proposed. On one hand, several authors have proposed non stochastic solutions mainly for speeding up the M step. On the other hand, stochastic solutions introduce a simulation step making use of pseudo-random draws at each iteration. This extra-randomness has a double effect. It circumvents the computation in closed form of the conditional expectation in the E step and prevents that the iterations stay near the unstable stationary points.

However, as reviewed in Section 2, none of these suggested improvements is guaranteed to provide a global maximizer of the likelihood function. The aim of this paper is to introduce such a global maximizer by combining the Monte-Carlo imputation principle and ideas from the simulated annealing technique to achieve the global optimization of the likelihood function.

First Section 2 gives a quick discussion on several known stochastic versions of the EM algorithm. This discussion also provides some background for the definition of our algorithm. We give this definition in Section 3, where the basic properties of the MEM algorithm is established. The convergence analysis of the algorithm is done in Section 4. Finally, in Section 5 we provide some experiments in two typical situations found in the statistical literature.

2 Stochastic variants of EM algorithm

In this section we shortly review several known stochastic variants of the EM algorithm, namely the stochastic EM (SEM), the Monte Carlo EM (MCEM) and the stochastic approximation version of the EM (SAEM) algorithms. Note that the motivations for the introduction of a stochastic step at each iteration are not the same for these algorithms. The simulation step of SEM relies on the Stochastic Imputation Principle, meaning that one completes the data y by a sample from $h(\cdot|y;\theta)$, while MCEM and SAEM use pseudo-random draws in order to get a Monte Carlo approximation for the E-step (1·1). However, despite the different motivations, all the three algorithms can be considered as a random perturbation of the deterministic dynamic system generated by EM.

The MCEM algorithm

Wei and Tanner (1990) proposed a Monte Carlo implementation of the E step, estimating the expectation in (1·1) by

$$S_k(\theta) = \frac{1}{m_k}\sum_{j=1}^{m_k} \log f(y, \tilde{Z}_{k,j};\theta) \qquad (2\cdot 1)$$

where $\tilde{Z}_{k,1},\ldots,\tilde{Z}_{k,m_k}$ are i.i.d. random samples from the conditional density

$$h(z|y;\theta_{k-1}) = f(x;\theta_{k-1})/g(y;\theta_{k-1}) \ .$$

Then, in the M-step we find

$$\theta_k = \mathrm{argmax}_{\theta\in\Theta} S_k(\theta) \ .$$

There are very few available results concerning the convergence of the MCEM. It is important to note that unlike EM, MCEM does not deterministically increase the likelihood at each iteration. This situation makes the convergence analysis difficult and the central difficulty is how to choose the sequence (m_k) of Monte Carlo replications to guarantee convergence. In their paper Wei and Tanner (1990) roughly recommend to start with small values of m_k and then to increase m_k as θ_k moves closer to the maximizer of $l(\theta)$. Recently, Booth and Hobert (1999) propose a practical rule for (m_k) based on consecutive confidence ellipsoids (see also experiments in Levine and Casella, 2000). However, a well-justified rule for (m_k) guaranteeing the convergence of the MCEM algorithm remains an open problem.

The SEM algorithm

The SEM algorithm (Celeux and Diebolt, 1985) has been the first stochastic version of EM algorithm. Nowadays it appears as a special case of MCEM when $m_k = 1$. The sequence (θ_k) generated by SEM does not converge pointwise. Actually, (θ_k) forms a homogeneous Markov chain which is expected to converge weakly to the unique stationary probability distribution ψ. Pointwise convergence is achieved by considering averaged estimates of the form

$$\tilde{\theta}_n = \frac{1}{n-n_0}\sum_{k=n_0}^{n} \theta_k \ , \qquad (2\cdot 2)$$

where n_0 is the length of the burn-in period in order to reduce the influence of the initial condition. When $n\to\infty$, $\tilde{\theta}_n$ converges to the mean of the stationary distribution ψ which is by definition the SEM estimator. The asymptotic properties of the SEM estimator are studied by Celeux and Diebolt (1993) in the case of finite

Gaussian mixtures and by Chadoeuf *et al.* (2000), who deal with censored Boolean segment processes. Nielsen (2000) gives large sample results for some estimators derived from the sequence (θ_k). Note also that an on-line version of the SEM has been proposed by Yao (2000), where the convergence to a local maximum is established.

The SAEM algorithm

The stochastic approximation EM (SAEM) algorithm has been proposed by Delyon *et al.* (1999) and makes use of a stochastic approximation procedure for estimating the conditional expectation (1·1). The basic idea is similar to the one of MCEM but the Monte Carlo integration is substituted in the E-step by a stochastic averaging procedure, namely

$$S_k(\theta) = S_{k-1}(\theta) + \gamma_k(\log f(y, \tilde{Z}_{k,1}; \theta) - S_{k-1}(\theta)) \ , \tag{2·3}$$

where $\tilde{Z}_{k,1}$ is a random sample from the conditional density $h(z|y;\theta_k)$ and (γ_k) is a decreasing sequence of positive step-sizes. The M-step does not change. One of the interesting features of SAEM is that its convergence analysis can be based on recent results from the stochastic approximation theory. However, pointwise almost sure convergence of the sequence (θ_k) to a local maxima of $g(y;\theta)$ is proved by Delyon *et al.* (1999) under conditions that models from an exponential family essentially satisfy.

3 MEM: a Metropolis version of the EM algorithm

We are interested in a situation where both the E and M steps from the EM algorithm cannot be expressed in closed form.

The proposed MEM algorithm starts like the MCEM by some i.i.d. random draws to get the Monte-Carlo approximation S_k of the conditional expectation of the complete log-likelihood (1·1). Then, instead of a deterministic maximisation M-step of the function S_k yielding the next estimate θ_k, we perform a random move in the parameter space according to a so-called Metropolis rule based on the approximation S_k. Therefore, θ_k will be a doubly random function of the current value θ_{k-1}, in the sense that it depends on both the random draws $\tilde{Z}_{k,j}$ and the random Metropolis-type move. Obviously, we will require these two randomness to be independent.

The idea is inspired by the theory of simulated annealing, see e.g. van Laarhoven and Arts (1987), and the hope is that the used Metropolis-type moves not only mimic a M-step but can also provide a global maximizer of the target function.

We now give the precise definition of the MEM algorithm. Let $(\Theta, \mathcal{A}, m)$ be the parameter space equipped with a probability measure m on a σ-field $\mathcal{A}$. To define a Metropolis rule, we are given a sequence (Q_k) of Markov transition kernels on Θ (*proposal kernels*). Each Q_k is assumed to be *symmetric* w.r.t the reference probability m, that is the measure $m(d\theta)Q_k(\theta, d\theta')$ is symmetric on the product space Θ^2, namely

$$\int_A Q_k(\theta, B) m(d\theta) = \int_B Q_k(\theta, A) m(d\theta) \ , \quad \text{for all} \quad (A, B) \in \mathcal{A}^2.$$

Moreover, let $(m_k)_{k\in\mathbb{N}}$ be an increasing and unbounded sequence of positive integers.

The MEM algorithm:

1. At time $k = 0$, pick a starting value θ_0.
2. At time $k \geq 1$, given the current estimate θ_{k-1},
   - draw m_k i.i.d. samples $\tilde{Z}_{k,1}, \ldots, \tilde{Z}_{k,m_k}$ from the conditional density $h(z|y;\theta_{k-1})$ and define
   $$S_k(\theta) = \frac{1}{m_k} \sum_{j=1}^{m_k} \log f(y, \tilde{Z}_{k,j}; \theta); \qquad (3{\cdot}1)$$
   - *Metropolis updating:* propose a tentative value θ' from the k-th proposal kernel $Q_k(\theta_{k-1}, \cdot)$ and accept it ($\theta_k = \theta'$) with probability
   $$c_k(\theta_{k-1}, \theta', \tilde{Z}_k) = 1 \wedge \exp\left\{m_k[S_k(\theta') - S_k(\theta_{k-1})]\right\}, \qquad (3{\cdot}2)$$
   where $a \wedge b = \min(a, b)$ and $\tilde{Z}_k = (\tilde{Z}_{k,1}, \ldots, \tilde{Z}_{k,m_k})$.
3. Iterate Step 2 until some stopping condition is satisfied.

At time k, the random proposal θ' is automatically accepted if $S_k(\theta') \geq S_k(\theta_{k-1})$; in this respect MEM mimics a step of a GEM algorithm. On the other hand, the random nature of the Metropolis rule implies that even a proposal θ' such that $S_k(\theta') < S_k(\theta_{k-1})$ could be accepted with a positive probability. This feature is of central importance for the MEM algorithm to escape from local maxima.

The MEM sequence (θ_k) forms a time-inhomogeneous Markov chain with transition kernels

$$P_k(\theta, A) = \int_A a_k(\theta, \theta') Q_k(\theta, d\theta') + \chi_A(\theta) \int_\Theta [1 - a_k(\theta, \theta')] Q_k(\theta, d\theta') \qquad (3{\cdot}3)$$

where $A \in \mathcal{A}$ and χ_A is the indicator function of the set A and the function $a_k : \Theta \times \Theta \to [0, 1]$ is the underlying acceptance probability function

$$\begin{aligned} a_k(\theta, \theta') &= \int c_k(\theta, \theta', \tilde{z}_k) \prod_{i=1}^{m_k} h(z_i|y, \theta) dz_1 \ldots dz_{m_k} \\ &= \int \left(1 \wedge \prod_{i=1}^{m_k} \frac{f(y, z_i, \theta')}{f(y, z_i, \theta)}\right) \prod_{i=1}^{m_k} h(z_i|y, \theta) dz_1 \ldots dz_{m_k}. \qquad (3{\cdot}4) \end{aligned}$$

First we prove that each P_k has an invariant probability measure (i.p.m.) $\pi_k(d\theta)$ proportional to $g^{m_k}(y, \theta) m(d\theta)$.

Lemma 1 *Assume that $D_k^{-1} = \int_\Theta g(y, \theta)^{m_k} m(d\theta)) < \infty$. Then, P_k is a reversible kernel with i.p.m. π_k given by*

$$\pi_k(d\theta) = D_k g(y, \theta)^{m_k} m(d\theta)\ . \qquad (3{\cdot}5)$$

Proof. It is sufficient to prove that $\pi_k(\theta)P_k(\theta, d\theta')$ is a symmetric measure on Θ^2, that is

$$\int_{A\times B} P_k(\theta, d\theta')g(y,\theta)^{m_k}m(d\theta) = \int_{B\times A} P_k(\theta, d\theta')g(y,\theta)^{m_k}m(d\theta);.$$

for any $(A, B) \in \mathcal{A}^2$. We have

$$\begin{aligned}
&\int_{A\times B} g(y,\theta)^{m_k}P_k(\theta, d\theta')m(d\theta) \\
&\quad= \int_{A\times B} a_k(\theta,\theta')Q_k(\theta, d\theta')g(y,\theta)^{m_k}m(d\theta) \\
&\quad+ \int_{\Theta\times\Theta} \chi_A(\theta)\chi_B(\theta)[1-a_k(\theta,\theta')]Q_k(\theta, d\theta')g(y,\theta)^{m_k}m(d\theta)\ .
\end{aligned}$$

For the first term on the right-hand, it will be sufficient to prove that $a_k(\theta,\theta')g(y,\theta)^{m_k}$ is a symmetric function, since $Q_k(\theta, d\theta')$ is a symmetric kernel. That is the case because

$$\begin{aligned}
g(y,\theta)^{m_k}a_k(\theta,\theta') &= \int g(y,\theta)^{m_k}\left(1\wedge\prod_{i=1}^{m_k}\frac{f(y,z_i,\theta')}{f(y,z_i,\theta)}\right)\prod_{i=1}^{m_k}h(z_i|y,\theta)dz_1\cdots dz_{m_k} \\
&= \int\left[\prod_{m_k} f(y,z_i,\theta)\wedge\prod_{m_k} f(y,z_i,\theta')\right]dz_1\cdots dz_{m_k}\ .
\end{aligned}$$

For the second term the symmetry is clear and the claim follows immediately. ■

Lemma 1 displays the key feature of the MEM algorithm by inheritance from the Metropolis-type simulated annealing algorithm: as m_k increases to infinity, the invariant density $D_k g^{m_k}(y,\theta)$ concentrates more and more on the set of global maxima of the target function $g(y,\theta)$. Actually when the MEM converges, the support of the limiting distribution is exactly the set of these maxima.

4 Convergence of the MEM algorithm

The MEM chain (θ_k) is not a standard Metropolis chain because in our context, the objective function, namely the observed likelihood function $g(y,\theta)$ is unknown. Therefore, to prove the convergence of the MEM algorithm, we propose to adapt the work of Haario and Sacksman (1991) (hereafter [HS]) about the simulated annealing on a general state space.

4·1 *The behavior of the sequence of i.p.m.* $(\pi_k)_{k=1}^{\infty}$

As in [HS], we first study the sequence of i.p.m's (π_k). We recall the definition (3·5) and without loss of generality we assume $g(y,\theta) > 0$ for all $\theta \in \Theta$. Then we can identify π_k as a Boltzmann distribution

$$\pi_k(d\theta) = D_k e^{-m_k H(y,\theta)}m(d\theta), \tag{4·1}$$

with "energy" function $H(y,\theta) = -\log g(y,\theta)$. Following [HS] we define $\mathcal{L}_H(z)$ as the steepness indicator of the energy function H, namely

$$\mathcal{L}_H(z) = \int_{\mathbb{R}} e^{-zx}\lambda_H(dx),\ z\in\mathbb{C},$$

where the measure λ_H has the distribution function

$$\lambda_H(x) = m\{\theta | H(y,\theta) \leq x\} \ .$$

Theorem 3.2 of [HS] yields the following estimate

$$\sum_{i=n+1}^{k} \|\pi_i - \pi_{i-1}\| \leq \log \frac{\mathcal{L}_H(m_n)}{\mathcal{L}_H(m_k)}, \ 1 \leq n \leq k.$$

4·2 *Estimates for the ergodicity coefficient of P_k*

Let us first recall the definition of Dobrushin's ergodicity coefficient. The norm $\|\lambda\|$ of a probability measure λ on $(\Theta, \mathcal{A})$ is the total variation norm. For a transition kernel $P(\theta, d\theta')$, the Dobrushin contraction coefficient $\delta(P)$ is

$$\delta(P) = \sup_{\lambda \neq \mu} \frac{\|\mu P - \lambda P\|}{\|\mu - \lambda\|},$$

where the sup is taken over all pairs of different probability measures defined on $(\Theta, \mathcal{A})$; see Seneta (1979). We have $0 \leq \delta(P) \leq 1$ and the sub-multiplicativity property $\delta(PP') \leq \delta(P)\delta(P')$. Set

$$\omega(y,z) = \inf_{(\theta,\theta') \in \Theta^2} \frac{f(y,z,\theta')}{f(y,z,\theta)}, \qquad b(y) = \inf_{\theta \in \Theta} \int h(z|y,\theta)\omega(y,z)dz \ .$$

It is clear that both ω and b are nonnegative and bounded by 1. In the sequel we will need the following

Hypothesis 1 *$b(y) > 0$.*

Remark 1. Assume that 1) Θ is a compact space; 2) for every θ and with respect to the conditional distribution $h(\cdot|y,\theta)$, the function $z \mapsto \omega(y,z)$ is not null everywhere; 3) for all z, the map $\theta \mapsto h(z|y,\theta)$ is continuous. Then $b(y) > 0$. ∎

Let $\Delta(y) = -\log b(y)$ and for $1 \leq n \leq k$, $P^{(n,k)} = P_{n+1} \cdots P_k$ and $Q^{(n,k)} = Q_{n+1} \cdots Q_k$. The Lemma below shows how the kernel $P^{(n,k)}$ inherits contraction from $Q^{(n,k)}$.

Lemma 2 *Under Hypothesis 1 and for all n, k, $1 \leq n \leq k$, we have*

$$1 - \delta(P^{(n,k)}) \geq e^{-\Delta(y)\sum_{j=n+1}^{k} m_j}[1 - \delta(Q^{(n,k)})].$$

Proof. We first show that

$$a_k(\theta,\theta') \geq e^{-\Delta(y)m_k}. \tag{4·2}$$

In fact:

$$\begin{aligned} a_k(\theta,\theta') &= \int \left(1 \wedge \prod_{i=1}^{m_k} \frac{f(y,z_i,\theta')}{f(y,z_i,\theta)}\right) \prod_{i=1}^{m_k} h(z_i|y,\theta)dz_1 \ldots dz_{m_k} \\ &\geq \int \left(1 \wedge \prod_{i=1}^{m_k} \omega(y,z_i)\right) \prod_{i=1}^{m_k} h(z_i|y,\theta)dz_1 \ldots dz_{m_k} \\ &= \left(\int \omega(y,z_i)h(z|y,\theta)dz\right)^{m_k} \\ &\geq e^{-\Delta(y)m_k}. \end{aligned}$$

Set $c = \exp\{-\Delta(y)\sum_{j=n+1}^{k} m_j\}$. Then we have that, for any probability measure μ and $A \in \mathcal{A}$,

$$\begin{aligned}
&\mu P^{(n,k)}(A) \\
&= \int_\Theta \mu(\theta_n) \int_\Theta P_{n+1}(\theta_n, d\theta_{n+1}) \cdots \int_A P_k(\theta_{k-1}, d\theta_k) \\
&\geq \int_\Theta \mu(d\theta_n) \int_\Theta a_{n+1}(\theta_n, \theta_{n+1}) Q_{n+1}(\theta_n, d\theta_{n+1}) \cdots \int_A a_k(\theta_{k-1}, \theta_k) Q_k(\theta_{k-1}, d\theta_k) \\
&\geq c\left(\mu Q^{(n,k)}\right)(A) \ ,
\end{aligned}$$

where the two estimates follow from (3·3) and (4·2), respectively. Then the conclusion follows using standard arguments (see the end of the proof of Lemma 4.1 in [HS]). ∎

Remark 2. Following the above proof, a better estimate could be obtained as follows. Assume the proposal kernels have a density: $Q_k(\theta, \theta') = s_k(\theta, \theta')m(d\theta')$. Set

$$\begin{aligned}
\omega_k(y,z) &= \inf\left\{ \frac{f(y,z,\theta')}{f(y,z,\theta)} \ : \ (\theta,\theta') \in \Theta^2 \ , s_k(\theta,\theta') > 0 \right\} \ , \\
b_k(y) &= \inf_{\theta\in\Theta} \int h(z|y,\theta)\omega_k(y,z)dz \ , \\
\Delta_k(y) &= -\log b_k(y) \ .
\end{aligned}$$

Clearly $\omega_k(y,z) \geq \omega(y,z)$, $\Delta_k(y) \leq \Delta(y)$. Then, the conclusion of Lemma 2 can be refined as

$$1 - \delta(P^{(n,k)}) \geq \exp^{-\sum_{j=n+1}^{k}\Delta_j(y)m_j}[1 - \delta(Q^{(n,k)})]. \qquad (4·3)$$

4·3 *Weak convergence of the MEM*

By analogy with the simulated annealing algorithm, the sequence (m_k) is called an *inverse temperature schedule*. Our main result below states essentially that, if this sequence grows as the logarithm, the MEM algorithm converges to a distribution concentrated in the set of the maxima of the likelihood function $g(y,\theta)$. In the sequel $[x]$ stands for the integer part of x.

Theorem 1 *Assume that the following conditions are satisfied.*

1. *The parameter space Θ is a compact subset of $\mathbb{R}^n$ with a non empty interior and equipped with inherited topology, and let m be the normalized restriction of Lebesgue measure on Θ.*
2. *The log-likelihood function* $\log g(y,\theta)$ *is continuous on Θ taking its maximum value at a finite number of* interior *points, say $\theta_1^*, \ldots, \theta_r^*$. Moreover the Hessian matrix*
$$J(\theta_i^*) = -\left.\frac{\partial^2}{\partial\theta^2}\log g(y,\theta)\right|_{\theta=\theta_i^*}$$
is positive definite at each θ_i^*, $i = 1, \ldots, r$.
3. *Suppose that for some $s \geq 1$ and for all $k \geq 1$,*
$$\delta(Q_{ks+1}Q_{ks+2}\cdots Q_{(k+1)s}) \leq d < 1.$$

4. *There exists $\varepsilon \in (0,1)$ such that the sequence (m_k) satisfies one of the following conditions:*

   (a) $m_k \leq \frac{\log(k+2)}{(1+\varepsilon)s\Delta(y)}$ *and the mapping* $k \to 1/m_k$ *is convex ;*

   (b) $m_k \leq \frac{\log(k+2)}{(1+\varepsilon)s\Delta(y)}$ *and* $\lim_{k\to\infty} \frac{m_{[k-k^{1-\varepsilon/2}]}}{m_k} = 1$.

Then, the distributions μ_k of θ_k converge weakly to the probability measure π_∞ defined by

$$\pi_\infty(d\theta) = \frac{\sum_{j=1}^{r} v_j \chi_{\theta_j^*}(d\theta)}{\sum_{j=1}^{r} v_j} \ ,$$

with $v_j = [\det J(\theta_j^*)]^{-1/2}$.

Proof. This is merely a straightforward application of Theorem 7.7 of [HS] taking into account the following:

1. here the cooling schedule is $T_k = 1/m_k$;
2. the smoothness conditions assumed on the target function $g(y,\theta)$ imply that on any small enough neighborhood of a maximum point θ_i^*, the following quadratic expansion holds

   $$-\left[\log g(y,\theta) - \log g(y,\theta_i^*)\right] = (\theta-\theta_i^*)^T J(\theta_i^*)(\theta-\theta_i^*) + \mathrm{o}(||\theta-\theta_i^*||^2) \ .$$

   Then the mixing weights (v_i) for the limiting distribution π_∞ are given by

   $$v_i = m\left(z \in \mathbb{R}^p \ : \ z^T J(\theta_i^*) z \leq 1\right) = \text{ const. } [\det J(\theta_j^*)]^{-1/2} \ . \quad \blacksquare$$

Remark 3. Condition 1, 2 and 4 are standard smoothness assumptions fulfilled by most of usual incomplete data models. In Condition 4, (m_k) can be relaxed along the same lines leading to the refinement of Eq. (4·3) and under the same assumption that the proposals have a density function. Then, the conclusion of Theorem 1 still holds if we substitute $\tilde{\Delta}_k(y)$ for $\Delta(y)$ where

$$\tilde{\Delta}_k(y) = \min\left\{\Delta_j(y) \ : \ 1 + [(k-1)/s]s \leq j \leq s + [(k-1)/s] \ \right\} \ .$$

Here the $\Delta_j(y)$'s are defined in the remark after Lemma 2.

Only Condition 3 is not trivial to check. However note that Θ is compact and let us consider a Gaussian kernel as proposal $Q_k(\theta, d\theta')$ defined by $\theta' = \theta + \varepsilon$, with some independent zero-mean Gaussian random vector ε having a fixed variance-covariance matrix. Then, if we make some projection at the boundary of the domain Θ (see [HS] for details), Condition 3 is satisfied. $\blacksquare$

5 Examples

Example 1: MEM is a global maximizer

The first example that we provide is a rather simple model taken from Arslan *et al.* (1993). Here the likelihood has several well-known local maxima implying that the classical EM algorithm will run to a local maximum nearest to the starting value.

The observed data is $y = (-20, 1, 2, 3)$ assumed to follow a Student's t-distribution with 0.05 degrees of freedom and unknown location parameter θ. The log-likelihood is given by

$$l(\theta) = -0.525 \sum_{i=1}^{4} \log\{0.05 + (y_i - \theta)^2\} \qquad (5{\cdot}1)$$

which does not admit a closed-form solution for the MLE of θ. In the complete-data $x = (y, z)$, the missing variables $z = (z_1, \ldots, z_4)$ are defined so that $Y_i|z_i \sim \mathcal{N}(\theta, 1/z_i)$ independently for $i = 1, \ldots, 4$ and $Z_i \sim \mathcal{G}(0.025, 0.025)$, where $\mathcal{G}(a, b)$ denotes the Gamma distribution with mean a/b. The log complete-data density can be written as

$$\log f(x; \theta) = \text{const} - 0.475 \sum_{i=1}^{4} \log z_i - 0.025 \sum_{i=1}^{4} z_i - 0.5 \sum_{i=1}^{4} z_i (y_i - \theta)^2.$$

It is not difficult to show that the conditional distribution of Z_i given y_i is

$$Z_i|y_i \sim \mathcal{G}\left(0.525, 0.025 + \frac{(y_i - \theta)^2}{2}\right).$$

The log-likelihood function is plotted in Figure 1 and has four local maxima $\hat{\theta}$ located at

$$\hat{\theta}_1 = -19.993, \quad \hat{\theta}_2 = 1.086, \quad \hat{\theta}_3 = 1.997, \quad \hat{\theta}_4 = 2.906.$$

It is easy to show also that the mapping induced by the EM algorithm is defined by

$$\theta_k = \frac{\sum_{i=1} y_i w_i(\theta_{k-1})}{\sum_{i=1} w_i(\theta_{k-1})}, \qquad (5{\cdot}2)$$

where $w_i(\theta) = 1.05 \cdot \left[0.05 + (y_i - \theta)^2\right]^{-1}$.

In our experiment, we have chosen 5 starting values $(-30, -18, 1.5, 2.5, 30)$. For these values the fixed points of the mapping (5·2) are $(-19.993, -19.993, 1.997, 1.997, 1.086)$.

We fixed the number of iterations of MEM equal to 3 000. Then, we used MEM in a kind of deterministic fashion, i.e., before each MEM run, the seed of the pseudo-random generator is set to the same value, regardless of the starting values. Thus for each of the 5 runs the procedure employs the same pseudo-random numbers. The proposal density for θ' at iteration k is a $\mathcal{N}(\theta_{k-1}, 4)$ (see Remark 3) and the temperature schedule is the logarithmic rule $m_k = \log(k + 2)/3$. The Figure 5 shows the behaviour of MEM for each iteration. In particular we note that the algorithm can escape from fixed points that are local maxima. However, our experience with the algorithm suggests that the choice of the variance in the proposal density deserves some care. Setting small values of the variance (e.g. less than or equal to 1) leads to a quite low convergence rate.

Example 2: a mixed logit-normal model

Our interest in this model originates from the work of McCulloch (1997) where the author considers the computation of the likelihood estimator by the MCEM algorithm in a particular situation. The model is a mixed logit-normal model with a single, normally distributed random effect and a single fixed effect:

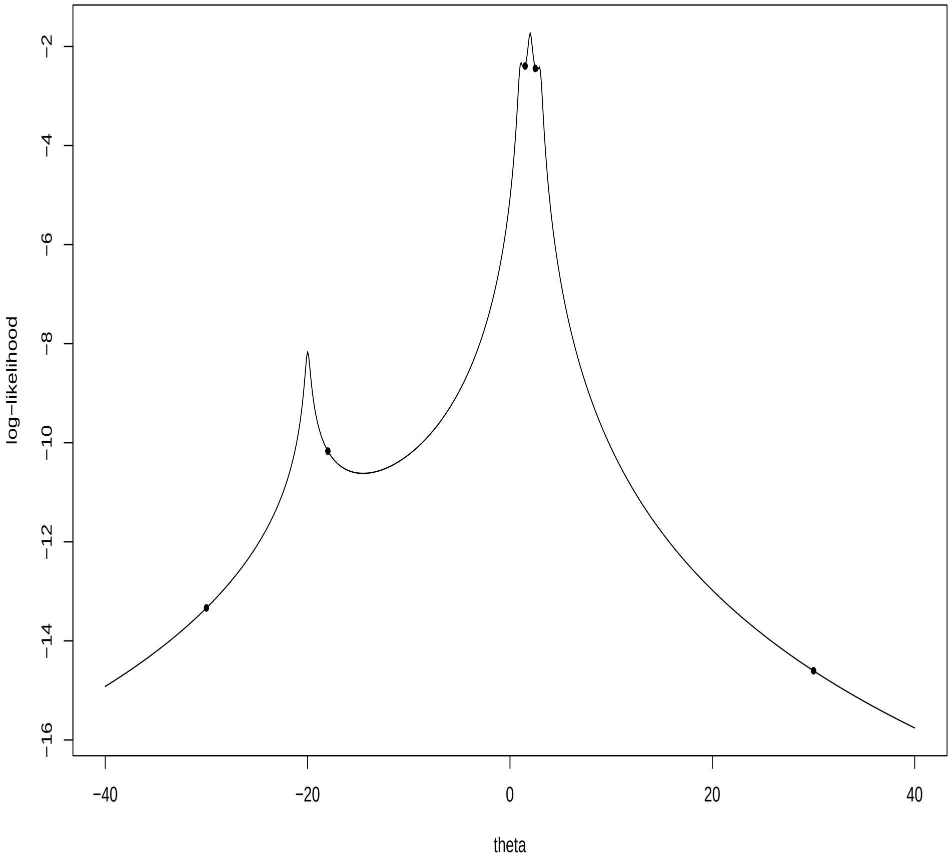


Figure 1: Arslan *et al.* (1993) example: the log-likelihood. The dotted points are the log-likelihood at the starting values $(-30, -18, 1.5, 2.5, 30)$ for the MEM algorithm.

- the random effect is $Z = (Z_j)$, i.i.d. $\mathcal{N}(0, \sigma^2)$;
- the fixed effect is $u = (u_{ij})$;
- conditionally to these effects, the response variables $Y = (Y_{ij})$ are independent Bernoulli variables with parameters (p_{ij}) fulfilling the linear logistic model

$$\log \frac{p_{ij}}{1 - p_{ij}} = \beta u_{ij} + z_j \ . \tag{5·3}$$

Here $i = 1, \ldots, n$ and $j = 1, \ldots, q$.

Let $\theta = (\beta, \sigma^2)$ be the parameters. The missing data are the unobserved random effect Z. For the complete-data $x = (y, z)$, the likelihood is

$$f(x; \theta) = \prod_{j=1}^{q} \xi_{0,\sigma^2}(z_j) \prod_{i=1}^{n} \frac{\exp[y_{ij}(\beta u_{ij} + z_j)]}{1 + \exp[\beta u_{ij} + z_j]} \ .$$

Here ξ_{0,σ^2} is the density function of $\mathcal{N}(0, \sigma^2)$. Thus the observed likelihood function is given through a product of q integrals

$$g(y; \theta) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y, z; \theta) dz_1 \cdots dz_q \ ,$$

which could be computed by numerical integration in this simple case, although the numerical error is far from easy to be controlled when q is not too small. Since this computation becomes infeasible for more complex random effect Z, e.g. $Z_{ij} = U_i + V_j$, it is worth applying EM-family algorithms to find the likelihood estimator.

In his paper McCulloch (1997) proposes to solve this problem by using the MCEM algorithm. It turns out that the conditional density $h(z|y, \theta)$ is as complex as the observed likelihood g, hence also unavailable. Then, at time k, in place of m_k i.i.d samples $\tilde{Z}_{k,1}, \ldots, \tilde{Z}_{k,m_k}$, McCulloch (1997) introduces a Markov chain sampler of length m_k to evaluate the Monte-Carlo mean (2·1). Later this model is also considered by Booth and Hobert (1999) and Levine and Casella (2000) where propose some more efficient versions of the MCEM algorithm are proposed.

Here we follow McCulloch (1997) for this Monte-Carlo step, by using a Metropolis sampler to generate a Markov sample $\tilde{Z}_{k,1}, \ldots, \tilde{Z}_{k,m_k}$ whose invariant distribution is the conditional distribution $h(z|y, \theta_{k-1})$. This sample is then used to evaluate the Monte-Carlo mean (3·1). For the reader's convenience, we will recall this sampler at the end of this example; see also §4.1 of McCulloch (1997).

We use the same setting as in the cited references for the simulation experiment. Namely, $\beta = 5$, $\sigma^2 = 1/2$, $n = 15$, $q = 10$ and $u_{ij} = i/15$. Indeed we use the data y listed in Table 2 of Booth and Hobert (1999) to ease some comparison, since data are not provided in McCulloch (1997)). For these data, the maximum likelihood estimator is found to be $(\widehat{\beta}, \widehat{\sigma}^2) = (6.132, 1.766)$ by numerical integration.

We have chosen the parameters of the MEM algorithm in a close way than those used by Booth and Hobert (1999) in their MCEM experiment. More precisely,

- the starting point is $\theta_0 = (2, 1)$;
- the temperature schedule is a simple-minded logarithmic rule

$$m_k = 100 \times \log(k + e - 1) \ , \qquad k \geq 1;$$

this schedule starts with $m_1 = 100$ which is the initial value used in Booth and Hobert (1999) and belongs to the family for which convergence is guaranteed by Theorem 1.

- at each iteration k, the proposal $\theta' = (\beta', \sigma^{2\prime})$ is defined as

$$\beta' = \beta_{k-1} + \varepsilon_1 \ , \qquad \sigma^{2\prime} = \sigma^2_{k-1} + \varepsilon_2 \ ,$$

where $\theta_{k-1} = (\beta_{k-1}, \sigma^2_{k-1})$ and ε_i are two independent Gaussian variables with mean 0 and given variance 1/10.

Our purpose is to sketch the behaviour of the MEM algorithm in this important situation rather than to provide an extensive simulation experiment. Note that the complexity of the MEM algorithm as well as for the MCEM algorithm is proportional not to the total number K of iterations, but the total number of Monte-Carlo replications $C = m_1 + \cdots + m_K$. We have run the MEM algorithm up to time $K = 10\ 000$. This quite large time is counterbalanced by the low increasing rate of the Monte-Carlo replication number m_k which varies from $m_1 = 100$ to $m_K = 921$ (for the cited MCEM in Booth and Hobert (1999), $K = 41$ is quite small, while m_k is a step function increasing from 100 to 17 536).

We got 20 independent runs of the MEM algorithm with $K = 10\ 000$. Figure 5 displays one of such runs (the others are very similar). As we can see, the MEM sequence (θ_k) approaches well the likelihood estimator, although residual fluctuations are present for large k. We believe, that in the current context, this is due to the Monte-Carlo sampling error from the Markov chain sampler used to evaluate the Monte-Carlo mean (3·1) as m_k remains relatively small. A simple way to get rid of these fluctuations is to consider the averaged sequence

$$\bar{\theta}_k = \frac{\theta_1 + \cdots + \theta_k}{k}$$

which is also displayed. Note that this average sequence can be recursively computed. It is clear that the average MEM estimator converges quickly to the likelihood estimator. It is this sequence that should be used in practice.

We conclude this example by recalling the *McCulloch's Metropolis sampler* for sampling from the conditional distribution $h(z|y;\theta)$ (McCulloch (1997)). At iteration k of the MEM algorithm, we generate m_k values $\tilde{Z}_{k,1}, \ldots, \tilde{Z}_{k,m_k}$ using this procedure. Let $\theta = (\beta, \sigma^2) = (\beta_{k-1}, \sigma^2_{k-1})$ be the current estimate and set $z(t) = \tilde{Z}_{k,t} = (z_1(t), \ldots, z_q(t))$ for $t = 1, \ldots, m_k$.

1. Initialize the chain with $z(0) = (z_1(0), \ldots, z_q(0))$ by q i.i.d. draws from $\mathcal{N}(0, \sigma^2)$.
2. For $t = 1, \ldots, m_k$ and for $j = 1, \ldots, q$
   - propose an update z' for $z_j(t-1)$ by an independent draw from $\mathcal{N}(0, \sigma^2)$;
   - accept this update, i.e. set $z_j(t) = z'$ (otherwise $z_j(t) = z_j(t-1)$) with probability $1 \wedge\ \alpha(z_j(t-1), z')$ where

$$\alpha(z, z') = \exp\left[(z' - z) y_{+j}\right] \prod_{i=1}^{n} \frac{1 + \exp[\beta u_{ij} + z]}{1 + \exp[\beta u_{ij} + z']} \ , \qquad y_{+j} = \sum_{i=1}^{n} y_{ij} \ .$$

Therefore the chain $(z(t))$ is an ergodic Markov chain with invariant density $h(\cdot|y, \theta)$.

References

Arslan, O., Constable, P.D.L. and Kent, J.T. (1993). Domains of conference for the EM algorithm: a cautionary tale in a location estimation problem. *Statistical Computing*, **3**, 103-108.

Booth J. G. and Hobert J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society B*, **89**, 265-285

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

Celeux, G. and Diebolt, J. (1993).Asymptotic properties of the stochastic EM algorithm for estimating mixing proportions. *Communications in Statistics - Stochastic Models*, **9**, 599–613.

Chadoeuf, J, Senoussi, R. and Yao, J.F. (2000). Parametric estimation of a Boolean segment process with stochastic restoration estimation. *Journal of Computational and Graphical Statistics*, **9**, 390-402.

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

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

Haario, A. and Sacksman, E. (1991). Simulated annealing in general state space. *Advances in Applied Probability*, **8**, 1177-1182.

Levine R. A. and Casella G. (2000). Implementation of the Monte Carlo EM algorithm. *Journal of Computational and Graphical Statistics*, **10**, 422-439

McCulloch C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association*, **92**, 162-170.

Meng X. and van Dyk D. (1997). The EM algorithm - an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society B*, **59**, 511-567

Nielsen, S. (2000). The stochastic EM algorithm: estimation and asymptotic results, *Bernoulli* **6**, 457–489.

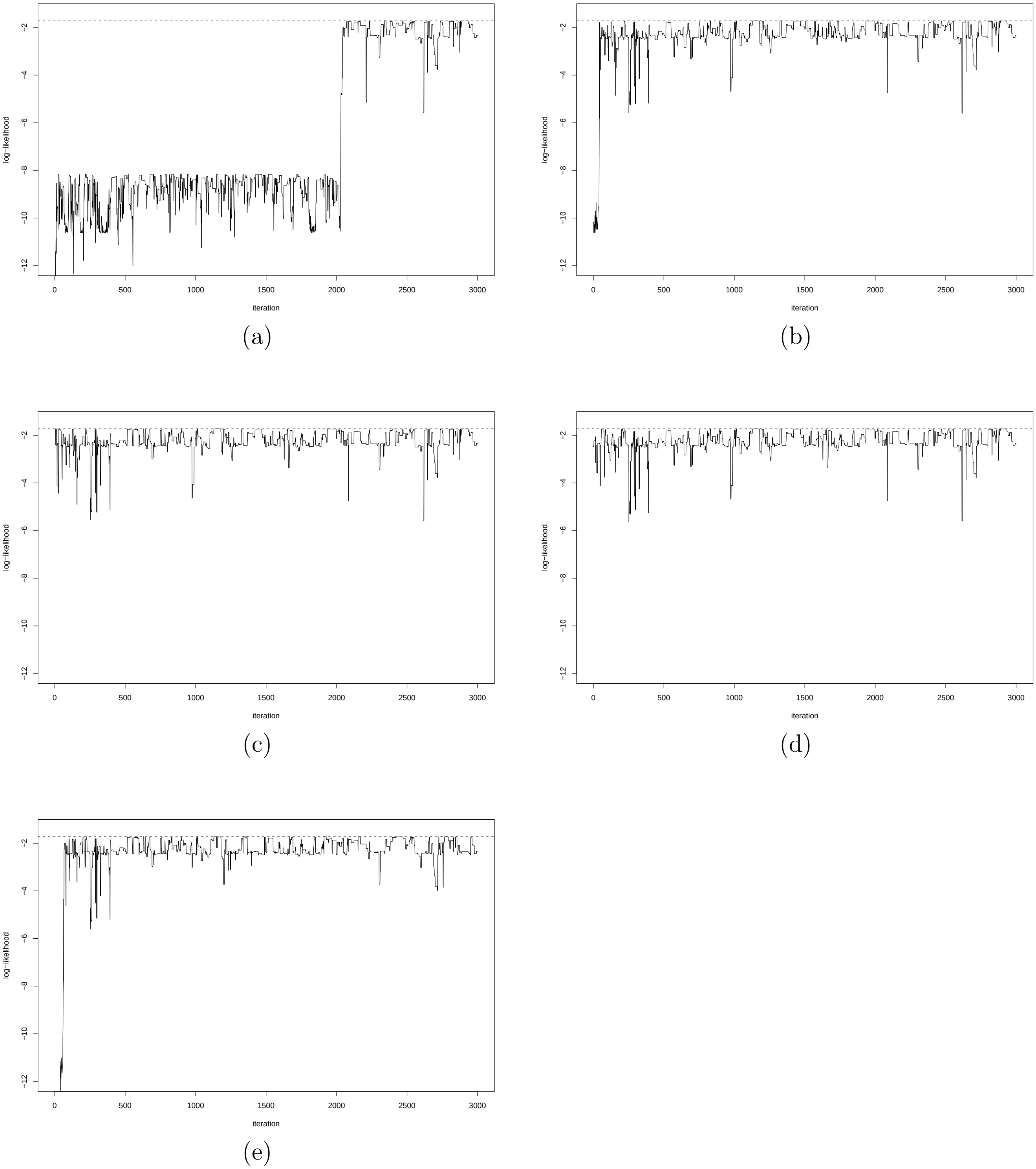
Seneta E. (1979). Coefficients of ergodicity : structure and applications. *Advances in Applied Probability*, **11**, 576-590

Wei, G.C.G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation. *Journal of the American Statistical Association*,**85**, 699-704.

Wu, C.F.J. (1983). On the convergence properties of the EM algorithm, *Annals of Statistics*, **11**, 95–103.

van Laarhoven, P.J.M. and Arts E.H.L. (1987). *Simulated Annealing: Theory and Applications*, Reidel, Dordrecht.

Yao J. (2000). On recursive estimation in incomplete data models. *Statistics*, **34**, 27-51.



(a)

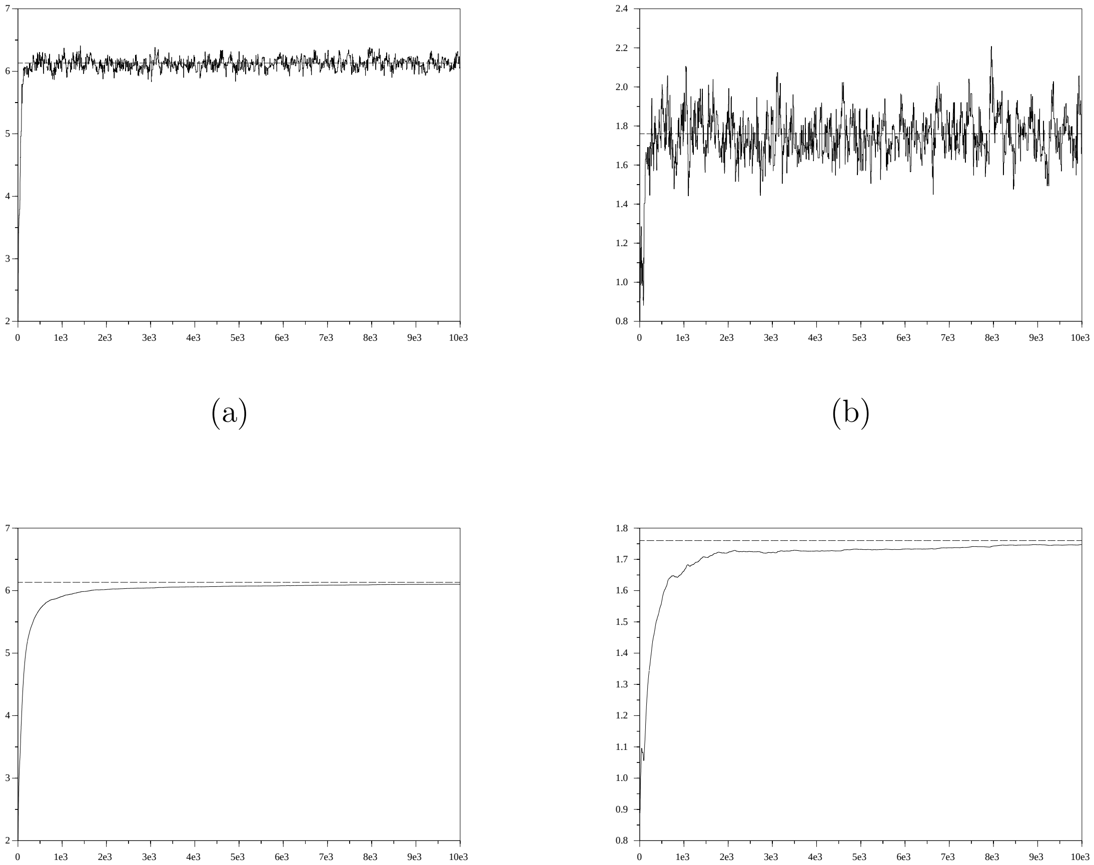
(b)

(c)

(d)

(e)

Figure 2: One run of the MEM algorithm up to 3 000 iterations from (a) $\theta_0 = -30$, (b) $\theta_0 = -18$, (c) $\theta_0 = 1.5$, (d) $\theta_0 = 2.5$, (e) $\theta_0 = 30$. Dashed horizontal lines are the global maximum of (5·1).



(a)

(b)

(c).

(d)

Figure 3: One run of the MEM algorithm up to time 10 000 where the horizontal lines are the values of the likelihood estimator $\hat{\beta} = 6.132$, $\hat{\sigma}^2 = 1.766$. (a) regression coefficient estimator β_k. (b) variance component estimator σ_k^2. (c) averaged regression coefficient estimator $\bar{\beta}_k$. (d) averaged variance component estimator $\bar{\sigma}_k^2$.