

# Coupling a stochastic approximation version of EM with a MCMC procedure

E. Kuhn

Université Paris Sud, France.

e-mail: Estelle.Kuhn@math.u-psud.fr

M. Lavielle

Université René Descartes and Université Paris Sud, France.

e-mail: Marc.Lavielle@math.u-psud.fr

7 April 2003

Running headline: *Coupling SAEM with MCMC*

This work was supported by Comity ECOS-Nord, n° V00M03.

## Abstract

The stochastic approximation version of EM (SAEM) proposed by Delyon *et al.* in [5] is a powerful alternative to EM when the E-step is untractable. Convergence of SAEM toward a maximum of the observed likelihood is established when the non observed data are simulated at each iteration under the conditional distribution. We show that this very restrictive assumption can be weakened. Indeed, the results of Benveniste *et al.* for stochastic approximation with Markovian perturbations are used to establish the convergence of SAEM when it is coupled with a Markov chain Monte-Carlo procedure. This result is very useful for many practical applications. Applications to the convolution model and the change-points model are presented to illustrate the proposed method.

**Key-words:** EM algorithm - SAEM algorithm - Stochastic approximation - MCMC algorithm - Convolution model - Change-points model.

## 1 Introduction

A wide class of statistical problems involves observed and non observed data. We can think, for example, in inverse problems such that deconvolution, source separation, change-points detection, etc. Linear and non linear mixed effects models can also be considered as incomplete data models. Estimation of the parameters of these models is a difficult challenge. In particular, the likelihood of the observations cannot usually be maximized in a closed form.

The expectation-maximization (EM) algorithm, proposed by Dempster, Laird and Rubin [6], is a broadly applicable approach for the iterative computation of maximum likelihood estimates, useful in a variety of incomplete-data (or partially-observed-data) statistical problems. The standard incomplete-data scheme considers that the observable incomplete data  $y \sim g(y; \theta)$  results from partial observation of complete data  $(y, z) \sim f(y, z; \theta)$ , where  $g$  and  $f$  are some known density functions. Maximum likelihood estimation of  $\theta$  consists in computing the value of  $\theta$  that maximizes the observed likelihood  $g$ .

The E-step of the EM algorithm computes  $Q(\theta|\theta_k) = E[\log f(y, z; \theta)|y; \theta_k]$  and the M-step determines  $\theta_{k+1}$  as maximizing  $Q(\theta|\theta_k)$ . Then, the observed-data likelihood sequence  $(g(y; \theta_k))$  is nondecreasing along any EM sequence, see [15] for more details. Stochastic versions of EM have been introduced from different perspectives to deal with situations where the E-step is unfeasible in a closed form. Monte Carlo EM (MCEM) replaces this step by a Monte Carlo approximation based on a large number of independent simulations of the missing data, see [12]. The SAEM algorithm, proposed by Delyon, Lavielle and Moulines [5], replaces the E-step by a stochastic approximation. The  $k$ -th step of SAEM generates  $m(k)$  realizations  $z_k(j)$  ( $1 \leq j \leq m(k)$ ) from

$p(z|y; \theta_k)$  and updates  $Q_{k-1}(\theta)$  according to

$$Q_k(\theta) = Q_{k-1}(\theta) + \gamma_k \left( \frac{1}{m(k)} \sum_{j=1}^{m(k)} \log f(y, z_k(j); \theta) - Q_{k-1}(\theta) \right),$$

where  $(\gamma_k)$  is a sequence of positive step sizes decreasing to 0. It is possible to select  $m(k) = 1$  for all  $k$  when simulation of  $z_k$  is heavy. Then the M-step consists in determining  $\theta_{k+1}$  which maximizes  $Q_k(\theta)$ . Precise results of convergence of SAEM are presented in [5] in the case where  $f(y, z; \theta)$  belongs to a regular curved exponential family. In this case, SAEM can be written in terms of the complete-data sufficient statistics. This leads to a general Robbins-Monro-type scheme and the almost sure convergence of the sequence  $(\theta_k)$  to a local maximum of the likelihood is proved under very general assumptions.

Unfortunately, for most non linear models or non Gaussian models, the non observed data cannot be simulated exactly under the conditional distribution. A well known alternative consists in using a Metropolis-Hastings algorithm : that means to introduce a transition probability which has as unique invariant distribution the conditional distribution we want to simulate. In this situation, the assumptions of [5] that ensure the convergence of SAEM are no more satisfied. The aim of this paper is to show that these assumptions can be weakened such that SAEM still converges when it is coupled with a Markov chain Monte Carlo procedure. The results of Benveniste *et al.* [1] for stochastic approximation with Markovian perturbations are used to establish the convergence of this algorithm.

The use of simulated data for estimating parameters is a powerful approach that tends to become popular for a few years. In [16], Yao defines and studies an online stochastic approximation scheme. In this situation, the number of observations goes to infinity and the sequence of estimate converges to the true value of the parameter. In [7], Gu and Kong propose a stochastic version of a Newton-Raphson algorithm for incomplete data estimation. The algorithm proposed by Gu and Zhu in [8] combines a MCMC procedure and a stochastic approximation for spatial models estimation. In these two papers, the information matrix is also estimated by stochastic approximation. In [4], Concordet and Nuñez propose a pseudo simulated maximum likelihood method for estimating the parameters of a non linear mixed effects model.

Application of SAEM to the change-points model was proposed in [10] and to the convolution model in [11], but no results of convergence of the algorithm were given. We show that the convergence results obtained for SAEM are very general and apply to these two examples of application. A Monte Carlo illustrate the performances of the proposed procedure with these two models.

## 2 The stochastic approximation version of EM algorithm

### 2.1 The EM and SAEM algorithms

We observe the data  $y$  in the set  $\mathcal{Y}$  which is a subset of  $\mathbb{R}^n$  and consider the complete data which are obtained by augmenting the observed data  $y$  with the missing data  $z$ . Let  $\mu$  be a  $\sigma$ -finite positive Borel measure on  $\mathbb{R}^l$  and let  $\mathcal{P} = \{f(y, z; \theta), \theta \in \Theta\}$  be a family of probability density functions with respect to  $\mu$  on  $\mathbb{R}^l$ , where  $\Theta$  is a subset of  $\mathbb{R}^p$ . We consider in this paper only models for which the complete data likelihood  $f(y, z; \theta)$  belongs to the curved exponential family. The incomplete data likelihood which is the likelihood of the observed data  $y$  is defined by:

$$g(y; \theta) \triangleq \int_{\mathbb{R}^l} f(y, z; \theta) \mu(dz).$$

Our purpose is to find the value  $\hat{\theta}$  in  $\Theta$  that maximizes the observed likelihood  $g$ .

In the sequel,  $p(z|y; \theta)$  denotes the conditional distribution of the missing data  $z$  given the observed data  $y$ :

$$p(z|y; \theta) \triangleq \begin{cases} f(y, z; \theta)/g(y; \theta) & \text{if } g(y; \theta) \neq 0 \\ 0 & \text{if } g(y; \theta) = 0 \end{cases}.$$

We shall make the following assumptions on the model:

- **(M1)** The parameter space  $\Theta$  is an open subset of  $\mathbb{R}^p$ . The complete data likelihood function is given by:

$$f(y, z; \theta) = \exp \left\{ -\psi(\theta) + \langle \tilde{S}(y, z), \phi(\theta) \rangle \right\},$$

where  $\langle \cdot, \cdot \rangle$  denotes the scalar product,  $\tilde{S}$  is a Borel function on  $\mathbb{R}^l$  in the second variable taking its values in an open subset  $\mathcal{S}$  of  $\mathbb{R}^m$ . Moreover, the convex hull of  $\tilde{S}(\mathbb{R}^l)$  is included in  $\mathcal{S}$ , and, for all  $\theta$  in  $\Theta$ ,

$$\int_{\mathbb{R}^l} |\tilde{S}(y, z)| p(z|y; \theta) \mu(dz) < \infty.$$

- **(M2)** Define  $L : \mathcal{S} \times \Theta \rightarrow \mathbb{R}$  as:

$$L(s; \theta) \triangleq -\psi(\theta) + \langle s, \phi(\theta) \rangle.$$

The functions  $\psi$  and  $\phi$  are twice continuously differentiable on  $\Theta$ .

- **(M3)** The function  $\bar{s} : \Theta \rightarrow \mathcal{S}$  defined as

$$\bar{s}(\theta) \triangleq \int_{\mathbb{R}^l} \tilde{S}(y, z) p(z|y; \theta) \mu(dz)$$

is continuously differentiable on  $\Theta$ .

**Remark.** *It is obvious that the function  $\bar{s}$  depends on the observation  $y$ . In order to keep notation as simple as possible, we won't denote formally this dependence. By the way, the observation  $y$  are kept fixed throughout the paper.*

- **(M4)** The function  $l : \Theta \rightarrow \mathbb{R}$  defined as the observed data log-likelihood

$$l(\theta) \triangleq \log g(y; \theta) = \log \int_{\mathbb{R}^d} f(y, z; \theta) \mu(dz)$$

is continuously differentiable on  $\Theta$  and

$$\partial_{\theta} \int f(y, z; \theta) \mu(dz) = \int \partial_{\theta} f(y, z; \theta) \mu(dz).$$

- **(M5)** There exists a function  $\hat{\theta} : \mathcal{S} \rightarrow \Theta$ , such that:

$$\forall s \in \mathcal{S}, \quad \forall \theta \in \Theta, \quad L(s; \hat{\theta}(s)) \geq L(s; \theta).$$

Moreover, the function  $\hat{\theta}$  is continuously differentiable on  $\mathcal{S}$ .

Let us define

$$Q(\theta|\theta') = \int_{\mathbb{R}^d} \log f(y, z; \theta) p(z|y; \theta') \mu(dz).$$

In cases where maximization of  $\theta \rightarrow Q(\theta|\theta')$  is much simpler than direct maximization of  $\theta \rightarrow l(\theta)$ , it is useful to apply the EM algorithm which maximizes  $l(\theta)$  by iteratively maximizing  $Q(\theta|\theta')$ . Each iteration of EM can be decomposed in two steps. At iteration  $k$ , the E-step consists in evaluating  $Q(\theta|\theta_k)$ . Then, the M-step consists in computing  $\theta_{k+1}$  by maximizing  $Q(\theta|\theta_k)$ .

Using our notations, the  $k$ -th iteration of the EM algorithm may be expressed as:

$$Q(\theta|\theta_k) = L(\bar{s}(\theta_k); \theta) \tag{1}$$

$$\theta_{k+1} = T(\theta_k) = \hat{\theta}(\bar{s}(\theta_k)). \tag{2}$$

The convergence of this algorithm is due to the fact that increasing  $Q(\theta|\theta_k)$  generates an increase of  $l(\theta_k)$ . This convergence has been studied by many different authors (see [6], [9], [15]) and is ensured, for example, under **(M1)**-**(M5)** when the sequence  $(\theta_k)_{k \geq 0}$  stays within some compact subset of  $\Theta$  (see [5]).

In the SAEM algorithm, the E-step is split into a simulation step (S-step) and an stochastic approximation integration step. At iteration  $k$ , the S-step consists in generating a realization of

the missing data vector  $z_k$  under the conditional distribution  $p(\cdot|y; \theta_k)$  and the integration step in a stochastic averaging procedure:

$$s_k = s_{k-1} + \gamma_k \left( \tilde{S}(y, z_k) - s_{k-1} \right). \quad (3)$$

Then, the complete log-likelihood is maximized in the M-step and  $\theta_{k+1} = \hat{\theta}(s_k)$ .

Some precise results of convergence of this algorithm were obtained in [5]. First, it is assumed that the random variables  $s_0, z_1, z_2, \dots$  are defined on the same probability space  $(\Omega, \mathcal{A}, P)$ . We denote  $\mathcal{F} = \{\mathcal{F}_k\}_{k \geq 0}$  the increasing family of  $\sigma$ -algebra generated by the random variables  $s_0, z_1, z_2, \dots, z_k$ . In addition, we assume that:

- **(SAEM1)** For all  $k$  in  $\mathbb{N}$ ,  $\gamma_k \in [0, 1]$ ,  $\sum_{k=1}^{\infty} \gamma_k = \infty$  and  $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$ .
- **(SAEM2)**  $l : \Theta \rightarrow \mathbb{R}$  and  $\hat{\theta} : \mathcal{S} \rightarrow \Theta$  are  $m$  times differentiable, where  $m$  is the integer such that  $\mathcal{S}$  is an open subset of  $\mathbb{R}^m$ .
- **(SAEM3)**
  1. For all positive Borel function  $\phi$ :

$$E[\phi(z_{k+1})|\mathcal{F}_k] = \int \phi(z)p(z|y; \theta_k)\mu(dz).$$

2. For all  $\theta \in \Theta$ ,  $\int \|\tilde{S}(y, z)\|^2 p(z|y; \theta)\mu(dz) < \infty$ , and the function

$$\Gamma(\theta) \triangleq \text{Cov}_{\theta}[\tilde{S}(y, z)] \triangleq \int_{\mathbb{R}^l} (\tilde{S}(y, z))^2 p(z|y; \theta)\mu(dz) - \left[ \int_{\mathbb{R}^l} \tilde{S}(y, z)p(z|y; \theta)\mu(dz) \right]^2$$

is continuous w.r.t.  $\theta$ .

It was shown in [5], that the sequence  $(\theta_k)_{k \geq 0}$  generated by SAEM converges to a stationary point of the observed likelihood  $g$ , when **(M1)-(M5)** and **(SAEM1)-(SAEM3)** are satisfied, and when the sequence  $(s_k)_{k \geq 0}$  takes its values in a compact subset of  $\mathcal{S}$ .

In many practical situations, it will not be possible to generate the non observed data  $z_k$  exactly under the conditional distribution  $p(\cdot|y; \theta_k)$ . Then, the assumption **(SAEM3)** will not be satisfied. Nevertheless, we will show that **(SAEM3)** can be replaced by a weaker condition that is satisfied when a Markov chain Monte-Carlo procedure is used in the S-step of SAEM.

## 2.2 Coupling the SAEM algorithm with MCMC

For any  $\theta \in \Theta$ , assume that the conditional distribution  $p(\cdot|y; \theta)$  is the unique limiting distribution of a transition probability  $\Pi_{\theta}$ . When the non observed data  $z_k$  cannot be generated

under  $p(\cdot|y; \theta_k)$ , we will use the transition kernel  $\Pi_{\theta_k}$ . Then, the  $k$ -th iteration of the proposed algorithm can be summarized in three steps as follows:

- *Simulation* : using  $z_{k-1}$ , generate a realization  $z_k$  from the transition probability  $\Pi_{\theta_k}(z_{k-1}, \cdot)$ .
- *Stochastic approximation* : update  $s_{k-1}$  according to (3).
- *Maximization* : update  $\theta_k$  according to  $\theta_{k+1} = \widehat{\theta}(s_k)$ .

Usually,  $\Pi_{\theta}$  will be defined as the succession of  $M$  iterations of a MCMC procedure, such as the Metropolis-Hastings algorithm. Then, the S-step of iteration  $k$  consists in simulating  $z_k$  with the transition probability  $\Pi_{\theta_k}(z_{k-1}, dz_k) = P_{\theta_k}^M(z_{k-1}, dz_k)$ , where

$$P_{\theta_k}(z, dz') = q_{\theta_k}(z, z') \min \left\{ \frac{p(z'|y; \theta_k) q_{\theta_k}(z', z)}{p(z|y; \theta_k) q_{\theta_k}(z, z')}, 1 \right\} dz' \quad (4)$$

for  $z' \neq z$  and  $P_{\theta_k}(z, \{z\}) = 1 - \int_{z' \neq z} P_{\theta_k}(z, dz')$ , where  $q_{\theta}(z, z')$  is any aperiodic recurrent transition density.

For example, we can use the marginal distribution  $\pi$  of  $z_k$  as an proposal distribution. Then, writing  $f(y, z; \theta) = \pi(z; \theta) h(y|z; \theta)$ , the acceptance probability only depends on the conditional distribution  $h$  of the observation  $y$ :

$$P_{\theta_k}(z, dz') = \pi(z'; \theta_k) \min \left\{ \frac{h(y|z'; \theta_k)}{h(y|z; \theta_k)}, 1 \right\} dz'. \quad (5)$$

### 3 Convergence result

We will use some result that works for general Robbins-Monro type stochastic approximation procedure. So if we write the recursion (3) into this form, it becomes:

$$s_k = s_{k-1} + \gamma_k h(s_{k-1}) + \gamma_k e_k, \quad (6)$$

where  $h$  stands for the mean field of the algorithm and  $e_k$  is a random perturbation. In our case, we obtain the following expressions for the function  $h$  and the sequence  $(e_k)$ :

$$\begin{aligned} h(s) &= E_{\widehat{\theta}(s)}(\tilde{S}(y, z)) - s = \bar{s}(\widehat{\theta}(s)) - s \\ e_k &= \tilde{S}(z_k) - E[\tilde{S}(z_k)|\mathcal{F}_{k-1}] = \tilde{S}(z_k) - \bar{s}(\widehat{\theta}(s_{k-1})) \end{aligned}$$

where  $E_{\theta}[\Phi(z)] \triangleq \int \Phi(z) p(z|y; \theta) \mu(dz)$ .

The assumption **(SAEM3)1.** means that, given  $\theta_0, \dots, \theta_k$ , the random variables  $z_0, \dots, z_k$

are independent. Coupled with the assumption **(SAEM3)2.**, it allows to prove that the series  $\sum \gamma_k e_k$  converges thanks to the martingal theory. This property plays a key role in the proof of the convergence of the SAEM algorithm in [5]. In this new version of the algorithm, we allow Markovian dependence between  $z_k$  and  $z_{k+1}$  : we suppose that  $z_{k+1}$  is obtained from  $z_k$  due to a Markovian transition depending on  $\theta_{k+1}$ , so that we will need some technical tools presented in [1] to show that the series  $\sum \gamma_k e_k$  still converges.

As announced above, the assumption **(SAEM3)** can be weakened. It is assumed that the random variables  $s_0, z_1, z_2, \dots, z_k, \dots$  are defined on the same probability space  $(\Omega, \mathcal{A}, P)$ . We denote  $\mathcal{F} = \{\mathcal{F}_k\}_{k \geq 0}$  the increasing family of  $\sigma$ -algebra generated by the random variables  $s_0, z_1, z_2, \dots, z_k$ . Since we introduce the transition probability  $\Pi_\theta$  in order to approach the conditional distribution, we have to make some assumptions on it:

- **(SAEM3')**

1. The chain  $(z_k)_{k \geq 0}$  takes its values in a compact subset  $\mathcal{E}$  of  $\mathbb{R}^d$ .
2. For any compact subset  $V$  of  $\Theta$ , there exists a real constant  $L$  such that for any  $(\theta, \theta')$  in  $V^2$

$$\sup_{(x,y) \in \mathcal{E}^2} |\Pi_\theta(x, y) - \Pi_{\theta'}(x, y)| \leq L|\theta - \theta'|.$$

3. The transition probability  $\Pi_\theta$  generates a uniformly ergodic chain which invariant probability is the conditional distribution  $p(\cdot|y; \theta)$ :

$$\exists K_\theta \in \mathbb{R}^+ \quad \exists \rho_\theta \in ]0, 1[ \quad | \quad \forall z \in \mathcal{E} \quad \forall k \in \mathbb{N} \quad \|\Pi_\theta^k(z, \cdot) - p(\cdot|y; \theta)\|_{TV} \leq K_\theta \rho_\theta^k,$$

where  $\|\cdot\|_{TV}$  denotes the total variation norm. We suppose also that:

$$K \triangleq \sup_\theta K_\theta < +\infty \quad \text{and} \quad \rho \triangleq \sup_\theta \rho_\theta < 1.$$

4. The function  $\tilde{S}$  is bounded on  $\mathcal{E}$ .

We also weaken the assumption **(SAEM1)**, we replace it by assumption **(SAEM1')**:

- **(SAEM1')** For all  $k$  in  $\mathbb{N}$ ,  $\gamma_k \in [0, 1]$ ,  $\sum_{k=1}^\infty \gamma_k = \infty$  and there exists  $\lambda$  in  $]\frac{1}{2}, 1]$  such that  $\sum_{k=1}^\infty \gamma_k^{1+\lambda} < \infty$ .

We obtain the following convergence result:

**Theorem 1.** *Assume that assumptions (M1)-(M5), (SAEM1')-(SAEM2) and (SAEM3') hold. Assume in addition the assumption (C): the sequence  $(s_k)_{k \geq 0}$  takes its values in a compact subset of  $\mathcal{S}$ . Then, w.p. 1,  $\lim_{k \rightarrow +\infty} d(\theta_k, \mathcal{L}) = 0$  where  $d(x, A)$  denotes the distance of  $x$  to the closed subset  $A$  and  $\mathcal{L} = \{\theta \in \Theta, \partial_\theta l(y; \theta) = 0\}$  is the set of stationary points of  $l$ .*

**Remarks:**

- For checking assumption **(SAEM3')3**, it is possible to verify some *minoration condition* or *Doebelin's condition* for the transition probability  $\Pi_\theta$  (see chapter 16 of [14]). Otherwise, we have to consider each case individually. Consider for example an independent Metropolis-Hastings algorithm: the transition density  $q_\theta$  mentioned in equation (4) defines an independence sampler:

$$\forall (z, z') \in \mathcal{E}^2 \quad q_\theta(z, z') = q_\theta(z').$$

Then the uniform ergodicity is ensured if the transition  $q_\theta$  satisfies the following inequality (see Theorem 2.1 in [13]):

$$\exists \beta \in \mathbb{R}^+ \quad | \quad \forall z \in \mathcal{E} \quad q_\theta(z) \geq \beta p(z|y; \theta).$$

- In cases where the compactness condition **(C)** is not checked or is difficult to check, it is possible to stabilize the algorithm by using the method of dynamic bounds proposed by Chen *et al.* in [3] and already used in this context by Delyon *et al.* in [5]: if  $s_k$  is outside a given compact set  $\mathcal{K}_k$  of  $\mathcal{S}$ , it is reinitialized in a specific compact set  $\mathcal{K}_0$ .
- It was shown in [5] that, under reasonable conditions, SAEM a.s. avoids traps, i.e., can only converge to a proper local maximizer of the likelihood. This result still holds in our context. Indeed, the paragraph II.3. of [2] about little Markovian perturbations gives some sufficient conditions for avoiding the local minima or saddle points of  $l$ :

$$\liminf \lambda_{\min} \left( E[(\tilde{S}(z_{n+1}) - s^*)(\tilde{S}(z_{n+1}) - s^*)^t | \mathcal{F}_n] \right) > 0 \text{ p.s. on } \Gamma(s^*),$$

where  $\lambda_{\min}(A)$  denotes the smallest eigenvalue of  $A$  and  $\Gamma(s^*)$  is the set of the sequences  $(s_k)$  which converge to  $s^*$ .

**Proof of Theorem 1:**

Theorem 1 is an application of Theorem 2 presented in [5], which gives a general result about convergence of Robbin-Monro type stochastic approximation procedures of the form (6):

**Theorem (Delyon [5]).** *Assume that:*

- **(SA0)** *w.p.1,  $\forall k \geq 0, s_k \in \mathcal{S}$ .*
- **(SA1)**  *$(\gamma_k)_{k \geq 0}$  is a decreasing sequence of positive number such that  $\sum_{k=1}^{\infty} \gamma_k = \infty$ .*
- **(SA2)** *The vector field  $h$  is continuous on  $\mathcal{S}$  and there exists a continuously differentiable function  $V : \mathcal{S} \rightarrow \mathbb{R}$  such that:*
  1.  $\forall s \in \mathcal{S} \quad F(s) = \langle \partial_s V(s), h(s) \rangle \leq 0$ .
  2.  $\text{int}(V(\mathcal{L})) = \emptyset$ , where  $\mathcal{L} \triangleq \{s \in \mathcal{S}, F(s) = 0\}$ .
- **(SA3)** *w.p.1, the closure of  $(\{s_k\}_{k \geq 0})$  is a compact subset of  $\mathcal{S}$ .*
- **(SA4)** *w.p.1,  $\lim_{n \rightarrow \infty} \sum_{k=0}^n \gamma_k e_k$  exists and is finite.*

*Then, w.p.1,  $\limsup d(s_k, \mathcal{L}) = 0$ .*

The assumptions **(SA0-SA3)** don't use the dependence structure of the sequence  $(z_k)$  and are checked under **(M1-M5)** **(SAEM1')** **(SAEM2)** like in the case where the missing data are exactly simulated under the a posteriori distribution (see [5]). Under the assumption **(SAEM3)**, the condition **(SA4)** is checked because the sequence of the partial sum of  $\sum \gamma_k e_k$  is a convergent martingale. For checking **(SA4)** under **(SAEM3')**, we use a result presented in the proposition 7 of Chapter 1 of [1]. The general model of the algorithm considered by Benveniste *et al.* is of the form:

$$s_k = s_{k-1} + \gamma_k H(s_{k-1}, z_k),$$

where the sequence  $(s_k)_{k \geq 0}$  evolves in  $\mathbb{R}^m$  and the sequence  $(z_k)_{k \geq 0}$  in  $\mathbb{R}^l$ . In our algorithm, the function  $H$  is equal to:

$$H(s, z) = \tilde{S}(y, z) - s,$$

where the observation  $y$  are considered as constant.

**Proposition (Benveniste et al., [1]).** *Assume the following assumptions:*

- **(A1)**  *$(\gamma_k)_{k \geq 0}$  is a decreasing sequence of positive real numbers such that  $\sum \gamma_k = +\infty$ .*
- **(A2)** *There exists a family  $\{\Pi_{\hat{\theta}(s)}, s \in \mathbb{R}^m\}$  of transition probabilities on  $\mathbb{R}^l$  such that for any Borel subset  $A$  of  $\mathbb{R}^l$ , we have*

$$P(z_k \in A | \mathcal{F}_{k-1}) = \Pi_{\hat{\theta}(s_{k-1})}(z_{k-1}, A).$$

- **(A3)** For any compact subset  $Q$  of  $\mathcal{S}$ , there exists a constant  $C_1$  depending on  $Q$  such that for all  $s$  in  $Q$  and all  $z$  in  $\mathcal{E}$  we have

$$|H(s, z)| \leq C_1.$$

- **(A4)** There exists a function  $h$  on  $\mathcal{S}$  and for each  $s$  in  $\mathcal{S}$  a function  $\nu_s$  on  $\mathbb{R}^l$  such that
  1.  $h$  is locally Lipschitz on  $\mathcal{S}$ : for all  $s$  in  $\mathcal{S}$ , there exist a neighborhood  $\mathcal{V}$  of  $s$  and a real constant  $C$  such that

$$\forall (s', s'') \in \mathcal{V}^2 \quad |h(s') - h(s'')| \leq C|s' - s''|.$$

2.  $(I - \Pi_{\hat{\theta}(s)})\nu_s = H_s - h(s)$  for all  $s$  in  $\mathcal{S}$ , where for all  $z$  in  $\mathcal{E}$ ,  $\Pi_{\hat{\theta}(s)}\nu_s(z) \triangleq \int \nu_s(z')\Pi_{\hat{\theta}(s)}(z, dz')$ .
3. For any compact subset  $Q$  of  $\mathcal{S}$ , there exist real constants  $C_2, C_3$  and  $\lambda$  in  $]\frac{1}{2}, 1]$ , such that for all  $s$  and  $s'$  in  $Q$  and for all  $z$  in  $\mathcal{E}$

$$|\nu_s(z)| \leq C_2$$

$$|\Pi_{\hat{\theta}(s)}\nu_s(z) - \Pi_{\hat{\theta}(s')}\nu_{s'}(z)| \leq C_3|s - s'|^\lambda.$$

- **(A5)** For any compact subset  $Q$  of  $\mathcal{S}$  and any positive real  $q$ , there exists a real number  $\mu_q(Q)$  such that, for all  $n$ , for all  $z_0$  in  $\mathbb{R}^l$  and all  $s_0$  in  $\mathbb{R}^m$

$$E_{z_0, s_0} \left( (1 + |z_k|^q) \mathbb{1}(s_{k-1} \in Q, k \leq n) \right) \leq \mu_q(Q),$$

where  $E_{z_0, s_0}$  denotes the expectation under the distribution of  $(z_k, s_k)_{k \geq 0}$  for the initial conditions  $(z_0, s_0)$ .

Then, for any compact subset  $Q$  of  $\mathcal{S}$ , denoting  $\tau(Q) = \inf\{k, s_k \notin Q\}$ , if the constant  $\lambda$  from **(A4)** verifies  $\sum \gamma_k^{1+\lambda} < +\infty$ , then, on  $\{\tau(Q) = \infty\}$ , the series  $\sum_k \gamma_k e_k$  converges a.s. and in  $L^2$ .

**Remark.** The proposition of Benveniste et al. requires  $\sum \gamma_k^{1+\lambda} \leq 1$ , but we only need  $\sum \gamma_k^{1+\lambda} < +\infty$  to obtain the convergence of the series  $\sum_k \gamma_k e_k$ .

In order to apply this proposition in our case, we have to check the assumptions **(A1–A5)**:

- **(A1)** is implied by **(SAEM1')**.
- **(A2)** is verified with the transition probability  $\Pi_{\hat{\theta}(s)}$ .
- **(A3)** is verified since  $\tilde{S}$  is bounded on  $\mathcal{E}$ .

- **(A5)** is verified since the sequence  $(z_k)_{k \geq 0}$  takes its values in a compact subset  $\mathcal{E}$  of  $\mathbb{R}^l$ .

Consider now the mean field of the algorithm defined by:

$$h(s) \triangleq \bar{s}(\widehat{\theta}(s)) - s.$$

We will show that this function  $h$  satisfies **(A4)**. The assumptions **(M3)** and **(M5)** imply that the function  $h$  is continuously differentiable on  $\mathcal{S}$ . So  $h$  is locally lipschitz on  $\mathcal{S}$  and assumption **(A4)1** is checked.

We define:

$$\nu_s(z) \triangleq \sum_{k \geq 0} \Pi_{\widehat{\theta}(s)}^k (H(s, z) - h(s)) = \sum_{k \geq 0} \left( \Pi_{\widehat{\theta}(s)}^k \tilde{S}(y, z) - p_{\widehat{\theta}(s)} \tilde{S} \right),$$

where  $p_{\widehat{\theta}(s)} \tilde{S} \triangleq \int \tilde{S}(y, z) p_{\widehat{\theta}(s)}(z|y) dz$ .

Since  $\Pi_{\widehat{\theta}(s)}$  is uniformly ergodic, there exist  $K_{\widehat{\theta}(s)} \in \mathbb{R}^+$  and  $\rho_{\widehat{\theta}(s)} \in ]0, 1[$ , such that, for any  $k \in \mathbb{N}$ , for any  $z$  in  $\mathcal{E}$

$$\sup_{\|u\| \leq 1} \left| \Pi_{\widehat{\theta}(s)}^k u(z) - p_{\widehat{\theta}(s)} u \right| \leq K_{\widehat{\theta}(s)} \rho_{\widehat{\theta}(s)}^k.$$

Since  $\tilde{S}$  is bounded on  $\mathcal{E}$ , the series defining  $\nu_s$  is convergent. Moreover, we have:

$$(I - \Pi_{\widehat{\theta}(s)}) \nu_s = \tilde{S} - p_{\widehat{\theta}(s)} \tilde{S},$$

which proves **(A4)2**.

Under assumption **(SAEM3')3** and since  $\tilde{S}$  is bounded on  $\mathcal{E}$ , we obtain the first inequality of **(A4)3**. We will now prove the second inequality of **(A4)3**:

$$\begin{aligned} \Pi_{\widehat{\theta}(s)} \nu_s(z) - \Pi_{\widehat{\theta}(s')} \nu_{s'}(z) &= \sum_{k \geq 1} \left( \Pi_{\widehat{\theta}(s)}^k \tilde{S}(y, z) - p_{\widehat{\theta}(s)} \tilde{S} \right) - \sum_{k \geq 1} \left( \Pi_{\widehat{\theta}(s')}^k \tilde{S}(y, z) - p_{\widehat{\theta}(s')} \tilde{S} \right) \\ &= \nu_s(z) - \nu_{s'}(z) + p_{\widehat{\theta}(s')} \tilde{S} - p_{\widehat{\theta}(s)} \tilde{S} \end{aligned}$$

so we have

$$|\Pi_{\widehat{\theta}(s)} \nu_s(z) - \Pi_{\widehat{\theta}(s')} \nu_{s'}(z)| \leq |\nu_s(z) - \nu_{s'}(z)| + |p_{\widehat{\theta}(s)} \tilde{S} - p_{\widehat{\theta}(s')} \tilde{S}|.$$

We will use the following technical lemma (the proof is in the appendix) to prove the second inequality of **(A4)3**:

**Lemma 1.** *If we assume **(SAEM2)**, **(SAEM3')** and **(C)**, then for any compact subset  $\mathcal{Q}$  of  $\mathcal{S}$ , there exist  $K_1$  and  $K_2$  in  $\mathbb{R}^+$  such that, for any  $\alpha \in ]0, 1[$ , for any  $(s, s', z) \in \mathcal{Q}^2 \times \mathcal{E}$ ,*

$$|p_{\widehat{\theta}(s)} \tilde{S} - p_{\widehat{\theta}(s')} \tilde{S}| \leq K_1 |s - s'|^\alpha \quad \text{and} \quad |\nu_s(z) - \nu_{s'}(z)| \leq K_2 |s - s'|^\alpha.$$

Proposition 7 of [1] can be applied to prove **(SA4)** and Theorem 2 of [5] proves the convergence of the sequence of estimates  $(\theta_k)_{k \geq 0}$ .  $\square$

## 4 Applications

### 4.1 The deconvolution problem

In a convolution model, the observation  $\mathbf{y} = (y_{L+1}, \dots, y_n)$  is the linear convolution of an non observed input sequence  $\mathbf{z} = (z_1, \dots, z_n)$  with additive noise  $\boldsymbol{\varepsilon}$ , e.g.,

$$y_t = \sum_{l=0}^L \varphi_l z_{t-l} + \sigma \varepsilon_t, \quad L+1 \leq t \leq n \quad (7)$$

where  $\boldsymbol{\varphi} = (\varphi_0, \dots, \varphi_L)$  is the convolution filter.

This kind of model is commonly used in seismic deconvolution, fMRI data analysis, or statistical signal analysis.

We shall make the following assumptions, concerning the random sequences  $\mathbf{z}$  and  $\boldsymbol{\varepsilon}$  : (i)  $(z_t, 1 \leq t \leq n)$  is a sequence of independent and identically distributed random variables, distributed according to some distribution function  $\pi$ , and taking their values in some compact of  $\mathbb{R}$ . (ii)  $(\varepsilon_t, L+1 \leq t \leq n)$  is a sequence of independent standardized Gaussian variables, (iii)  $z_t$  and  $\varepsilon_{t'}$  are independent for all pairs of time indexes  $1 \leq t \leq n$  and  $L+1 \leq t' \leq n$ .

This assumption together with equation (7) specifies completely the log-likelihood of the observed data samples. Let  $\mathcal{M}(\mathbf{z})$  be the  $(n-L) \times (L+1)$  matrix such that  $\mathcal{M}_{ij}(\mathbf{z}) = z_{L+1+i-j}$ . Then, the complete log-likelihood is, up to a constant,

$$\log f(\mathbf{y}, \mathbf{z}; \theta) = -\frac{n-L}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \|\mathbf{y} - \mathcal{M}(\mathbf{z}) \boldsymbol{\varphi}\|^2 + \log \pi(\mathbf{z}). \quad (8)$$

Deconvolution consists in recovering the input sequence  $\mathbf{z}$  from the observation  $\mathbf{y}$ . Of course, deconvolution requires an accurate estimation of the convolution filter  $\boldsymbol{\varphi}$  and the noise variance  $\sigma^2$ . SAEM will be very useful for estimating these parameters. Furthermore, the marginal distribution  $\pi$  of the input sequence  $\mathbf{z}$  can also be estimated, whenever it belongs to the exponential family and depends on a unknown parameter  $\psi$  :

$$\pi(\mathbf{z}; \psi) = C(\psi) \exp \left\{ - \left\langle \tilde{S}_\pi(\mathbf{z}), \psi \right\rangle \right\}. \quad (9)$$

Here, the vector of parameters of the model is  $\theta = (\psi, \boldsymbol{\varphi}, \sigma^2)$  and the minimal sufficient statistics are  $\tilde{S}(\mathbf{y}, \mathbf{z}) = (\mathcal{M}(\mathbf{z})^t \mathcal{M}(\mathbf{z}), \mathcal{M}(\mathbf{z})^t \mathbf{y}, \tilde{S}_\pi(\mathbf{z}))$ . At step  $k$ , we used the following procedure for generating  $\mathbf{z}_k$  from  $\mathbf{z}_{k-1}$ , using  $\theta_k = (\psi_k, \boldsymbol{\varphi}_k, \sigma_k^2)$  :

- i) a permutation  $p_k$  of  $\{1, 2, \dots, n\}$  is randomly chosen,

ii) for  $i = 1, 2, \dots, n$  :

1. Let  $j = p_k(i)$ . Set  $\tilde{z}_t = z_{k-1,t}$  for any  $t \neq j$  and generate  $\tilde{z}_j \sim \pi(\cdot; \psi_k)$ .
2. Compute

$$\alpha = \frac{1}{2\sigma_k^2} (\|\mathbf{y} - \mathcal{M}(\tilde{\mathbf{z}})\boldsymbol{\varphi}_k\|^2 - \|\mathbf{y} - \mathcal{M}(\mathbf{z})\boldsymbol{\varphi}_k\|^2). \quad (10)$$

3. Generate  $u \sim \text{Expo}(1)$ . Set  $\mathbf{z}_k = \tilde{\mathbf{z}}$  if  $\alpha < u$  and  $\mathbf{z}_k = \mathbf{z}_{k-1}$  otherwise.

We can easily show that  $\Pi_{\theta_k}(\mathbf{z}_{k-1}, \cdot)$  is the transition probability of an ergodic Markov chain, that converges uniformly to the conditional distribution  $p(\cdot|\mathbf{y}; \theta_k)$ . For estimating  $(\psi, \boldsymbol{\varphi}, \sigma^2)$ , we define a sequence  $(s_{k,1}, s_{k,2}, s_{k,3})$  according to (3) :

$$\begin{aligned} s_{k,1} &= s_{k-1,1} + \gamma_k (\mathcal{M}(\mathbf{z}_k)^t \mathcal{M}(\mathbf{z}_k) - s_{k-1,1}) \\ s_{k,2} &= s_{k-1,2} + \gamma_k (\mathcal{M}(\mathbf{z}_k)^t \mathbf{y} - s_{k-1,2}) \\ s_{k,3} &= s_{k-1,3} + \gamma_k (\tilde{S}_\pi(\mathbf{z}_k) - s_{k-1,3}). \end{aligned}$$

Then, the maximization step yields

$$\begin{aligned} \boldsymbol{\varphi}_{k+1} &= (s_{k,1})^{-1} s_{k,2} \\ \sigma_{k+1}^2 &= \frac{1}{n-L} (\mathbf{y}^t \mathbf{y} - (s_{k,2})^t \boldsymbol{\varphi}_{k+1}) \\ \psi_k &= \text{Argmax } C(\psi) e^{-\langle s_{k,3}, \psi \rangle}. \end{aligned}$$

All the assumptions of Theorem 1 are satisfied whenever  $\pi$  has a bounded support, and the SAEM algorithm converges to a (local) maximum of the observed likelihood. To assume that the input variables  $\mathbf{z}$  are bounded is not a restrictive assumption for a practical point of view. Indeed, any non bounded distribution is truncated in the practice.

We used the convolution model described in (7) for simulating an observed series  $\mathbf{y}$  of length  $n = 500$ . In this example, the input sequence  $\mathbf{z}$  are independent  $Beta(a, b)$  random variables on  $[0, 1]$  with  $a = b = 3$ . So the statistic  $\tilde{S}_\pi(\mathbf{z})$  is  $(S_1(\mathbf{z}), S_2(\mathbf{z})) = (\sum \log(z_j), \sum \log(1 - z_j))$ . The convolution filter is  $\boldsymbol{\varphi} = (1, -3, 2, 6, 2, -3, 1)$ . We choose  $\sigma^2 = 0.2286$  in order to ensure a Signal to Noise ratio equal to 10dB, *i.e.*  $\text{Var}(\mathcal{M}(\mathbf{z})\boldsymbol{\varphi}) = 10\text{Var}(\sigma\boldsymbol{\varepsilon})$ .

Table 1 gives the estimation of  $\theta = (a, b, \boldsymbol{\varphi}, \sigma^2)$ . A Monte-Carlo based on 100 replications was used for estimating the mean and the standard deviation of two estimators. First, the maximum likelihood estimator  $\hat{\theta}_f$ , which maximizes the complete likelihood  $f(\mathbf{y}, \mathbf{z}; \theta)$  assuming

that the input series  $\mathbf{z}$  is known, was computed as follows:

$$\begin{aligned}\hat{\varphi}_f &= (\mathcal{M}(\mathbf{z})^t \mathcal{M}(\mathbf{z}))^{-1} \mathcal{M}(\mathbf{z})^t \mathbf{y} \\ \hat{\sigma}_f^2 &= \frac{1}{n-L} \left( \mathbf{y}^t \mathbf{y} - \mathbf{y}^t \mathcal{M}(\mathbf{z}) \hat{\varphi}_f \right) \\ (\hat{a}_f, \hat{b}_f) &= \text{Arg max}_{a,b} \frac{1}{B(a,b)} e^{-(a-1)S_1(\mathbf{z}) - (b-1)S_2(\mathbf{z})},\end{aligned}$$

where  $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$  and  $\Gamma(a) = \int_0^{+\infty} e^{-x} x^{a-1} dx$ .

On the other hand, the estimator  $\hat{\theta}_g$  maximizes the incomplete likelihood  $g(\mathbf{y}; \theta)$ , considering that the input series  $\mathbf{z}$  is unknown.

We computed  $\hat{\theta}_g$  with 100 iterations of SAEM, using  $\gamma_k = 1$  for  $1 \leq k \leq 30$  and  $\gamma_k = 1/(k-29)$  for  $k \geq 31$ .

The initialisation is the uniform distribution for  $\mathbf{z}$  ( $a_0 = b_0 = 0$ ). The initial guess for the filter is a spike located at  $j = 4$ . That ensures that the algorithm recovers the good phase of the convolution filter. For a different initialization, the algorithm can converge to a local maximum of the likelihood that cannot be compared with the true filter  $\varphi^*$ . For example, using as initial guess a spike at  $j = 2$ , a simulation gives  $\hat{\varphi} = (0.73, 6.65, 1.42, -3.51, -0.58, 2.54, -1.32)$ . We remark that the phase of the true filter is not recovered, but the estimated filter and the true filter have both almost the same transfert function. The problem of convergence to the global maximum of the likelihood is beyond the scope of this paper (see [11] for a simulated annealing version of this algorithm).

The results presented in Table 1 confirm that  $\hat{\theta}_f$  is a more accurate estimate of  $\theta$  than  $\hat{\theta}_g$ . Nevertheless, we can remark that, when  $\mathbf{z}$  is not observed, SAEM provides a good estimation of  $\theta$ .

**TABLE 1 ABOUT HERE**

## 4.2 The change-points problem

We use the model considered in [10]. We observe a real sequence  $\mathbf{y} = (y_i, 1 \leq i \leq n)$ , such that, for any  $1 \leq i \leq n$ ,

$$y_i = f(t_i) + \sigma \varepsilon_i, \quad (11)$$

where  $(t_i, 1 \leq i \leq n)$  is a sequence of known observation times and  $(\varepsilon_i, 1 \leq i \leq n)$  is a sequence of independent zero-mean Gaussian variables with unit variance. The function  $f$  to recover is piecewise constant. Thus, there exists a sequence of instants  $(\tau_j, j \geq 0)$  among the sequence

$(t_i, 1 \leq i \leq n)$  and a sequence  $(m_j, j \geq 1)$  such that, for any  $j \geq 1$ ,

$$f(t) = m_j \text{ for all } \tau_{j-1} < t \leq \tau_j \quad (12)$$

with the convention  $\tau_0 = 0$ .

We introduce a latent sequence of independent identically distributed Bernoulli random variables  $(z_i, 1 \leq i \leq n-1)$  that takes the value 1 at the change instants and 0 between two changes :

$$z_i = \begin{cases} 1 & \text{if there exists } j \text{ such that } t_i = \tau_j \\ 0 & \text{otherwise} \end{cases} . \quad (13)$$

Let  $\lambda$  be the parameter of the Bernoulli and for any  $\mathbf{z} = (z_i, 1 \leq i \leq n-1)$  in  $\Omega = \{0, 1\}^{n-1}$ , let  $K_{\mathbf{z}} = \sum_{i=1}^{n-1} z_i + 1$  be the number of segments (*i.e.*  $K_{\mathbf{z}} - 1$  is the number of change-points) defined by  $\mathbf{z}$ . Then,

$$\pi(\mathbf{z}; \lambda) = \lambda^{K_{\mathbf{z}}-1} (1 - \lambda)^{n-K_{\mathbf{z}}} . \quad (14)$$

Conditionally to the change-points sequence, the vector  $\mathbf{m} = (m_j, 1 \leq j \leq K_{\mathbf{z}})$  is assumed to be Gaussian:

$$\pi(\mathbf{m}|\mathbf{z}; \mu, V) = \prod_{j=1}^{K_{\mathbf{z}}} \left( \frac{2\pi V}{n_j} \right)^{-\frac{1}{2}} \exp \left\{ -\frac{n_j}{2V} (m_j - \mu)^2 \right\}, \quad (15)$$

where  $n_j = \sum_{i=1}^n \mathbb{1}_{] \tau_{j-1}, \tau_j ]}(t_i)$  is the length of segment  $j$  for  $1 \leq j \leq K_{\mathbf{z}}$ .

On the other hand,  $(\varepsilon_i, 1 \leq i \leq n)$  is assumed to be a sequence of independent zero-mean and unit variance Gaussian random variables. Thus, the conditional distribution of the observations is defined by:

$$h(\mathbf{y}|\mathbf{z}, \mathbf{m}; \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{j=1}^{K_{\mathbf{z}}} \sum_{l=N_{j-1}+1}^{N_j} (y_l - m_j)^2 \right\}, \quad (16)$$

where  $N_j = \sum_{l=1}^j n_l$  for  $1 \leq j \leq K_{\mathbf{z}}$  and  $N_0 = 0$ .

Let  $\theta = (\mu, \lambda, V, \sigma^2)$  be the set of hyper-parameters of the model. For any configuration of changes  $\mathbf{z}$ , let  $\bar{y}_j = n_j^{-1} \sum_{l=N_{j-1}+1}^{N_j} y_l$ ,  $\bar{y} = n^{-1} \sum_{l=1}^n y_l$  and  $C_{\mathbf{z}} = \sum_{j=1}^{K_{\mathbf{z}}} \sum_{l=N_{j-1}+1}^{N_j} (y_l - \bar{y}_j)^2$ . Then, after some calculus, it can be shown (see [10]) that the likelihood of  $(\mathbf{y}, \mathbf{z})$  is defined by

$$\begin{aligned} f(\mathbf{y}, \mathbf{z}; \theta) &= (2\pi\sigma^2)^{-\frac{n}{2}} \left( \frac{\sigma^2 + V}{\sigma^2} \right)^{-\frac{K_{\mathbf{z}}}{2}} \lambda^{K_{\mathbf{z}}-1} (1 - \lambda)^{n-K_{\mathbf{z}}} \\ &\times \exp \left\{ -\frac{1}{2(V + \sigma^2)} \left( \sum_{i=1}^n (y_i - \mu)^2 + \frac{V}{\sigma^2} C_{\mathbf{z}} \right) \right\}. \end{aligned} \quad (17)$$

Here, the minimal sufficient statistics to be approximated is  $(K_{\mathbf{z}}, C_{\mathbf{z}})$  according to (3) :

$$\begin{aligned} s_{k,1} &= s_{k-1,1} + \gamma_k(K_{\mathbf{z}} - s_{k-1,1}) \\ s_{k,2} &= s_{k-1,2} + \gamma_k(C_{\mathbf{z}} - s_{k-1,2}). \end{aligned}$$

Then, the maximization step yields

$$\begin{aligned} \mu_k &= \bar{y} \\ \lambda_k &= \frac{s_{k,1} - 1}{n - 1} \\ \sigma_k^2 &= \frac{s_{k,2}}{n - s_{k,1}} \\ V_k &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - s_{k,2}}{s_{k,1}} - \sigma_k^2. \end{aligned}$$

Here,  $\mathbf{z}$  takes a finite number of values. Then, any irreducible proposal kernel  $q$  can be used to generate a geometrically ergodic kernel. In this application, we used alternatively the three following kernels: *i*) a new vector  $\tilde{\mathbf{z}}$  is drawn independently of the current value  $\mathbf{z}_{k-1}$  with the marginal distribution  $\pi(\cdot; \psi_{k-1})$ , *ii*) a new change-point is created, or an existing change-point is removed, *iii*) an existing change-point is shifted (see [10] for more details concerning the MCMC procedure).

The maximum likelihood estimate of  $\mu$  is  $\bar{y}$ . For estimating the others hyper-parameters, the SAEM algorithm can be used since it is easy to check that the assumptions of Theorem 1 are still satisfied.

We used the change-point model described above for simulating an observed series  $\mathbf{y}$  of length  $n = 500$ . We set the values of the hyper-parameters to  $\lambda^* = 0.02$ ,  $V^* = 40$  and  $\sigma^{2*} = 1$ .

Table 2 gives the estimation of  $\theta = (\lambda, V, \sigma^2)$ . A Monte-Carlo based on 100 replications was used for estimating the mean and the standard deviation of  $\hat{\theta}_f$ , which maximizes the complete likelihood  $f(\mathbf{y}, \mathbf{z}; \theta)$  and  $\hat{\theta}_g$  that maximizes the incomplete likelihood  $g(\mathbf{y}; \theta)$ . We computed  $\hat{\theta}_g$  with 100 iterations of SAEM, using  $\gamma_k = 1$  for  $1 \leq k \leq 30$  and  $\gamma_k = 1/(k - 29)$  for  $k \geq 31$ .

In this example, SAEM still produces a good estimation of  $\theta$ . In particular, the noise variance  $\sigma^2$  is estimated with almost the same accuracy, when the change-points are known and when they are unknown.

**TABLE 2 ABOUT HERE**

## A Appendix

We first prove a technical lemma which will be necessary to prove Lemma 1.

**Lemma 2.** *If we assume (SAEM2), (SAEM3') and (C), then for any compact subset  $\mathcal{Q}$  of  $\mathcal{S}$ , there exists a real constant  $M$  such that:*

$$\forall (s, s', y, z) \in \mathcal{Q}^2 \times \mathcal{Y} \times \mathcal{E} \quad \forall k \in \mathbb{N} \quad |\Pi_{\hat{\theta}(s)}^k \tilde{S}(y, z) - \Pi_{\hat{\theta}(s')}^k \tilde{S}(y, z)| \leq Mk|s - s'|.$$

**Proof:**  $\forall (s, s') \in \mathcal{Q}^2 \quad \forall (y, z) \in \mathcal{Y} \times \mathcal{E} \quad \forall k \in \mathbb{N}$ ,

$$\begin{aligned} & |\Pi_{\hat{\theta}(s)}^k \tilde{S}(y, z) - \Pi_{\hat{\theta}(s')}^k \tilde{S}(y, z)| \\ & \leq \sum_{i=0}^{k-1} \left| \Pi_{\hat{\theta}(s)}^{k-i} \Pi_{\hat{\theta}(s')}^i \tilde{S}(y, z) - \Pi_{\hat{\theta}(s)}^{k-1-i} \Pi_{\hat{\theta}(s')}^{i+1} \tilde{S}(y, z) \right| \\ & = \sum_{i=0}^{k-1} \left| \Pi_{\hat{\theta}(s)}^{k-1-i} \left( \Pi_{\hat{\theta}(s)} - \Pi_{\hat{\theta}(s')} \right) \Pi_{\hat{\theta}(s')}^i \tilde{S}(y, z) \right| \\ & = \sum_{i=0}^{k-1} \left| \int \int \int \Pi_{\hat{\theta}(s')}^i(z, u) \left( \Pi_{\hat{\theta}(s)}(u, v) - \Pi_{\hat{\theta}(s')}^i(u, v) \right) \Pi_{\hat{\theta}(s)}^{k-1-i}(v, w) \tilde{S}(y, w) dudvdw \right| \\ & \leq \|\tilde{S}\|_{\infty} \sum_{i=0}^{k-1} \int \int \int \Pi_{\hat{\theta}(s')}^i(z, u) \left| \Pi_{\hat{\theta}(s)}(u, v) - \Pi_{\hat{\theta}(s')}^i(u, v) \right| \Pi_{\hat{\theta}(s)}^{k-1-i}(v, w) dudvdw. \end{aligned}$$

The assumption (SAEM2) ensures that the set  $\hat{\theta}(\mathcal{Q})$  is compact, so that the assumption (SAEM3')**2** ensures the existence of real constants  $L$  and  $\tilde{L}$  such that:

$$\begin{aligned} |\Pi_{\hat{\theta}(s)}^k \tilde{S}(y, z) - \Pi_{\hat{\theta}(s')}^k \tilde{S}(y, z)| & \leq \|\tilde{S}\|_{\infty} L |\hat{\theta}(s) - \hat{\theta}(s')| \sum_{i=0}^{k-1} \int \int \int \Pi_{\hat{\theta}(s')}^i(z, u) \Pi_{\hat{\theta}(s)}^{k-1-i}(v, w) dudvdw \\ & \leq \|\tilde{S}\|_{\infty} \tilde{L} |s - s'| \sum_{i=0}^{k-1} \int \int \int \Pi_{\hat{\theta}(s')}^i(z, u) \Pi_{\hat{\theta}(s)}^{k-1-i}(v, w) dudvdw \end{aligned}$$

since  $\hat{\theta}$  is continuously differentiable. Moreover  $\mathcal{E}$  is compact, so we obtain:

$$|\Pi_{\hat{\theta}(s)}^k \tilde{S}(y, z) - \Pi_{\hat{\theta}(s')}^k \tilde{S}(y, z)| \leq \|\tilde{S}\|_{\infty} \tilde{L} M(\mathcal{E}) k |s - s'|$$

where  $M(A)$  denotes the Lebesgue measure of the set  $A$ .  $\square$

**Proof of lemma 1:** Assumptions (SAEM3')**1** and (SAEM3')**3** imply the existence of constants  $K \in \mathbb{R}^+$  and  $\rho \in ]0, 1[$  such that for all  $(s, s')$  in  $\mathcal{Q}^2$ , for all  $(y, z)$  in  $\mathcal{Y} \times \mathcal{E}$  and for all  $k$  in

$\mathbb{N}$ , we have

$$\begin{aligned} |p_{\hat{\theta}(s)}\tilde{S} - p_{\hat{\theta}(s')} \tilde{S}| &\leq |p_{\hat{\theta}(s)}\tilde{S} - \Pi_{\hat{\theta}(s)}^k \tilde{S}(y, z)| + |\Pi_{\hat{\theta}(s)}^k \tilde{S}(y, z) - \Pi_{\hat{\theta}(s')}^k \tilde{S}(y, z)| + |\Pi_{\hat{\theta}(s')}^k \tilde{S}(y, z) - p_{\hat{\theta}(s')} \tilde{S}| \\ &\leq 2\|\tilde{S}\|_\infty K\rho^k + Mk|s - s'|. \end{aligned}$$

We choose  $k \approx \log(|s - s'|)/\log(\rho)$  so that the two terms have approximatively the same weight. Then, there exists a constant  $K_1$  in  $\mathbb{R}^+$  such that

$$\forall (s, s') \in \mathcal{Q}^2 \quad \text{s.t.} \quad |s - s'| < 1, \quad |p_{\hat{\theta}(s)}\tilde{S} - p_{\hat{\theta}(s')} \tilde{S}| \leq \frac{K_1}{\log(\rho)} \log(|s - s'|)|s - s'|.$$

Since we have

$$\forall \alpha \in ]0, 1[ \quad 1[\lim_{|s-s'|\rightarrow 0} \log(|s - s'|)|s - s'|^{1-\alpha} = 0$$

we can deduce that  $\forall \alpha \in ]0, 1[, \exists \eta > 0$  s.t.

$$\forall (s, s') \in \mathcal{Q}^2 \quad \text{s.t.} \quad |s - s'| \leq \eta, \quad |p_{\hat{\theta}(s)}\tilde{S} - p_{\hat{\theta}(s')} \tilde{S}| \leq K_1|s - s'|^\alpha \quad (18)$$

which proves the first inequality of Lemma 1 for all pair  $(s, s')$  in a compact  $\mathcal{Q}$  such that  $|s - s'| \leq \eta$ .

Concerning the second inequality of Lemma 1, let us define

$$a_{k,s,s'}(y, z) = \left| \left( \Pi_{\hat{\theta}(s)}^k \tilde{S}(y, z) - p_{\hat{\theta}(s)} \tilde{S} \right) - \left( \Pi_{\hat{\theta}(s')}^k \tilde{S}(y, z) - p_{\hat{\theta}(s')} \tilde{S} \right) \right|.$$

On one hand, since  $\tilde{S}$  is bounded and  $\Pi_{\hat{\theta}(s)}$  is uniformly ergodic, there exist  $K \in \mathbb{R}^+$  and  $\rho \in ]0, 1[$  such that, for any  $k \in \mathbb{N}$  and any  $(s, s', y, z) \in \mathcal{Q}^2 \times \mathcal{Y} \times \mathcal{E}$ ,

$$\begin{aligned} a_{k,s,s'}(y, z) &\leq |\Pi_{\hat{\theta}(s)}^k \tilde{S}(y, z) - p_{\hat{\theta}(s)} \tilde{S}| + |\Pi_{\hat{\theta}(s')}^k \tilde{S}(y, z) - p_{\hat{\theta}(s')} \tilde{S}| \\ &\leq 2\|\tilde{S}\|_\infty K\rho^k. \end{aligned}$$

On the other hand, for any  $\beta \in ]0, 1[$ , there exists  $\eta > 0$  such that, for any  $k \in \mathbb{N}$  and any  $(s, s', y, z) \in \mathcal{Q}^2 \times \mathcal{Y} \times \mathcal{E}$  such that  $|s - s'| \leq \eta$ ,

$$\begin{aligned} a_{k,s,s'}(y, z) &\leq |\Pi_{\hat{\theta}(s)}^k \tilde{S}(y, z) - \Pi_{\hat{\theta}(s')}^k \tilde{S}(y, z)| + |p_{\hat{\theta}(s)} \tilde{S} - p_{\hat{\theta}(s')} \tilde{S}| \\ &\leq Mk|s - s'| + K_1|s - s'|^\beta \leq \max(M, K_1)(k + 1)|s - s'|^\beta. \end{aligned}$$

Then, using a convexity argument, there exist  $L_1 \in \mathbb{R}^+$ ,  $L_2 \in \mathbb{R}^+$  and  $L_3 \in \mathbb{R}^+$  such that, for any  $a \in ]0, 1[$ , for any  $\beta \in ]0, 1[$ , there exists  $\eta > 0$  such that, for any  $k \in \mathbb{N}$  and any  $(s, s', y, z) \in \mathcal{Q}^2 \times \mathcal{Y} \times \mathcal{E}$  such that  $|s - s'| \leq \eta$ ,

$$\begin{aligned} a_{k,s,s'}(y, z) &\leq \min(L_1\rho^k, L_2(k + 1)|s - s'|^\beta) \\ &\leq L_3\rho^{(1-a)k}|s - s'|^{\beta a}(k + 1)^a. \end{aligned}$$

Thus, for any  $\beta \in ]0, 1[$ , for any  $a \in ]0, 1[$ , there exists  $\eta > 0$  such that, for any  $k \in \mathbb{N}$  and any  $(s, s', y, z) \in \mathcal{Q}^2 \times \mathcal{Y} \times \mathcal{E}$  such that  $|s - s'| \leq \eta$ ,

$$\begin{aligned} |\nu_s(z) - \nu_{s'}(z)| &\leq \sum_{k \geq 0} a_{k,s,s'}(y, z) \\ &\leq L_3 \left( \sum_{k \geq 0} \rho^{(1-a)k} (k+1)^a \right) |s - s'|^{a\beta}. \end{aligned}$$

So we obtain the following conclusion: for any  $a \in ]0, 1[$ , for any  $\alpha \in ]0, a[$ , there exist real constants  $K_2$  and  $\eta$  such that for all  $z$  in  $\mathcal{E}$

$$\forall (s, s') \in \mathcal{Q}^2 \quad \text{s.t.} \quad |s - s'| < \eta \quad |\nu_s(z) - \nu_{s'}(z)| \leq K_2 |s - s'|^\alpha. \quad (19)$$

Since  $\mathcal{Q}$  is compact, we can recover it with a finite number  $N$  of balls of diameter  $\eta$  and obtain so the inequality (18) and (19) with the same constants multiplied by  $N$  for any pair  $(s, s')$  in  $\mathcal{Q}^2$ .  $\square$

## References

- [1] Albert Benveniste, Michel Métivier, and Pierre Priouret, *Adaptive algorithms and stochastic approximations*, Springer-Verlag, Berlin, 1990, Translated from the French by Stephen S. Wilson.
- [2] O. Brandiere and M. Duflo, *Les algorithmes stochastiques contournent-ils les pièges ?*, Ann. Inst. Henri Poincaré (1995), 395–427.
- [3] Han Fu Chen, Guo Lei, and Ai Jun Gao, *Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds*, Stochastic Process. Appl. **27** (1988), no. 2, 217–231.
- [4] Didier Concordet and Olivier G. Nunez, *A simulated pseudo-maximum likelihood estimator for nonlinear mixed models*, Comput. Statist. Data Anal. **39** (2002), no. 2, 187–201.
- [5] Bernard Delyon, Marc Lavielle, and Eric Moulines, *Convergence of a stochastic approximation version of the EM algorithm*, Ann. Statist. **27** (1999), no. 1, 94–128.
- [6] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Roy. Statist. Soc. Ser. B **39** (1977), no. 1, 1–38, With discussion.

- [7] Ming Gao Gu and Fan Hui Kong, *A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems*, Proc. Natl. Acad. Sci. USA **95** (1998), no. 13, 7270–7274 (electronic).
- [8] Ming Gao Gu and Hong-Tu Zhu, *Maximum likelihood estimation for spatial models by Markov chain Monte Carlo stochastic approximation*, J. R. Stat. Soc. Ser. B Stat. Methodol. **63** (2001), no. 2, 339–355.
- [9] Kenneth Lange, *A gradient algorithm locally equivalent to the EM algorithm*, J. Roy. Statist. Soc. Ser. B **57** (1995), no. 2, 425–437.
- [10] M. Lavielle and E. Lebarbier, *An application of MCMC methods to the multiple change-points problem*, Signal Processing **81** (2001), 39–53.
- [11] M. Lavielle and E. Moulines, *A simulated annealing version of the EM algorithm for non-Gaussian deconvolution*, Statistics and Computing **7** (1997), no. 4, 229–236.
- [12] Xiao-Li Meng and Donald B. Rubin, *Maximum likelihood estimation via the ECM algorithm: a general framework*, Biometrika **80** (1993), no. 2, 267–278.
- [13] K.L. Mengersen and R.L. Tweedie, *Rates of convergence of the Hastings and Metropolis algorithms.*, Ann. Stat. **24** (1996), no. 1, 101–121 (English).
- [14] S.P. Meyn and R.L. Tweedie, *Markov chains and stochastic stability.* (English).
- [15] C.-F. Jeff Wu, *On the convergence properties of the EM algorithm*, Ann. Statist. **11** (1983), no. 1, 95–103.
- [16] Jian-Feng Yao, *On recursive estimation in incomplete data models.*, Statistics **34** (2000), no. 1, 27–51 (English).

	$\theta^*$	$\theta_0$	E $\hat{\theta}_f$	std $\hat{\theta}_f$	E $\hat{\theta}_g$	std $\hat{\theta}_g$
$a$	3	0	3.0075	0.1846	2.8615	0.4014
$b$	3	0	3.0220	0.1838	2.6396	0.3712
$\varphi_1$	1	0	1.0002	0.1020	1.0277	0.5820
$\varphi_2$	-3	0	-3.0058	0.0997	-2.6564	0.4612
$\varphi_3$	2	0	2.0100	0.1046	1.8809	0.8416
$\varphi_4$	6	1	5.9810	0.0963	5.5637	0.4024
$\varphi_5$	2	0	2.0034	0.1128	1.8963	0.7713
$\varphi_6$	-3	0	-2.9985	0.0955	-2.7920	0.5766
$\varphi_7$	1	0	1.0009	0.1047	0.8361	0.5523
$\sigma^2$	0.2286	1	0.2266	0.0157	0.2615	0.0593

Table 1. Estimation of  $\theta = (a, b, \varphi, \sigma^2)$ :  $\theta^*$  is the true value of  $\theta$ ,  $\theta_0$  is the initialization,  $\hat{\theta}_f$  is the estimation obtained by maximizing  $f(\mathbf{y}, \mathbf{z}; \theta)$  and  $\hat{\theta}_g$  is the estimation obtained by maximizing  $g(\mathbf{y}; \theta)$ .

	$\theta^*$	$\theta_0$	E $\hat{\theta}_f$	std $\hat{\theta}_f$	E $\hat{\theta}_g$	std $\hat{\theta}_g$
$\lambda$	0.02	0.05	0.0201	0.0051	0.0235	0.0105
$V$	40	10	38.1	15.8	32.5	19.0
$\sigma^2$	1	5	1.01	0.07	1.01	0.08

Table 2. Estimation of  $\theta = (\lambda, V, \sigma^2)$ :  $\theta^*$  is the true value of  $\theta$ ,  $\theta_0$  is the initialization,  $\hat{\theta}_f$  is the estimation obtained by maximizing  $f(\mathbf{y}, \mathbf{z}; \theta)$  and  $\hat{\theta}_g$  is the estimation obtained by maximizing  $g(\mathbf{y}; \theta)$ .

\*\*\*\*\*

Corresponding author :

Marc LAVIELLE

Universit Paris-Sud, Bt 425

91400 Orsay, FRANCE

Marc.Lavielle@math.u-psud.fr

\*\*\*\*\*