An application of MCMC methods for the multiple change-points problem

Marc Lavielle

ELSEVIER

Signal Processing 81 (2001) 39–53

SIGNAL PROCESSING

www.elsevier.nl/locate/sigpro

An application of MCMC methods for the multiple change-points problem

M. Lavielle*,[1], E. Lebarbier

Equipe de Probabilités, Statistique et Modélisation, Université Paris Sud, Bât 425, 91400 Orsay, Cedex, France

Received 1 May 1999; received in revised form 10 April 2000

Abstract

We present in this paper a multiple change-point analysis for which an MCMC sampler plays a fundamental role. It is used for estimating the posterior distribution of the unknown sequence of change-points instants, and also for estimating the hyperparameters of the model. Furthermore, a slight modification of the algorithm allows one to compute the change-points sequences of highest probabilities. The so-called reversible jump algorithm is not necessary in this framework, and a very much simpler and faster procedure of simulation is proposed. We show that different interesting statistics can be derived from the posterior distribution. Indeed, MCMC is powerful for simulating joint distributions, and its use should not be restricted to the estimation of marginal posterior distributions, or posterior means.

Keywords: Change-point detection; Gibbs sampler; Hastings–Metropolis algorithm; Reversible jump; SAEM algorithm

1. Introduction

The subject of change-points analysis has been important in statistics for many years. This significant activity is largely motivated by the big amount of applications in signal processing (EEG, EMG and ECG analysis, geophysics, etc.) [1,3,14,22], and many theoretical results have been obtained in various contexts, (see, for example, [1,4,7,15]).

Among the different approaches, we can mention the on-line (or sequential) detection of change-points. In an off-line context, the unknown sequence of change-points instants can be estimated by minimizing a well-suitable contrast function (see [14]). We shall adopt here a Bayesian approach. Then, the change-point problem consists mainly in estimating the *posterior* distribution of the change-points sequence. That allows, for example, to estimate the probability that a change has occurred at a given instant t. The posterior distribution of the number of changes can also be derived. The maximum a posteriori (MAP) estimator is obtained by maximizing this posterior distribution. When the changes affect the mean of the signal, we show that the MAP estimator is a penalized least-squares estimator, that possesses good statistical properties [15].

An MCMC method is really suitable for estimating the posterior distribution of the change-points sequence. The reversible jump algorithm proposed

[1] Also at Université Paris-V, France.

* Corresponding author. Tel.: + 33-1-6915-5743; fax: + 33-1-6915-7234.

E-mail addresses: marc.lavielle@math.u-psud.fr (M. Lavielle), lebarbier@math.u-psud.fr (E. Lebarbier).

0165-1684/01/$ - see front matter
PII: S0165-1684(00)00189-4

by Green [12], is based on the fact that the dimension of the model can change, according to the number of segments. Unfortunately, this algorithm converges slowly, and many iterations are needed for estimating correctly the posterior distribution.

Another parametrization is shown to be more appropriate than the sequence (τ_k) of change points. It consists in introducing a sequence (r_t) that takes the value 1 at the change-points instants, and 0 between two jumps. The advantage of this parametrization is that the dimension of the sequence r is fixed. When the length of the observed signal is n, the Hastings–Metropolis method proposed in this paper simply consists in sampling sequences of 0 and 1, of fixed length $n-1$. Furthermore, running this algorithm at a *low temperature* allows to estimate the most likely configurations of changes. An hybrid MCMC algorithm, combining the basic Hastings–Metropolis algorithm with the Gibb's sampler [2,11], can be used for estimating also the distribution of the mean sequence. Nevertheless, we show that the posterior expectation of the mean is not appropriate in this context, since it yields a smooth version of the signal, instead of a step function. The distribution of the mean sequence, conditionally to the most likely configuration of change-points, has more sense, and provides a good estimation. At the end, we show that another slight modification of our MCMC algorithm allows to estimate the hyper-parameters of the model. The stochastic approximation of expectation maximization (SAEM) procedure proposed by Delyon et al. [8] merely consists in updating the set of hyper-parameters at each iteration of MCMC. This algorithm converges to a maxima of the likelihood, and provides automatically a "good" *prior* distribution for the unknown sequences.

The paper is organized as follows. Section 2 describes the model of change-points in the mean and details the prior modelling. The Hastings–Metropolis samplers used for estimating the posterior distribution of the change-points instants are described in Section 3. Section 4 addresses the problem of recovering also the sequence of means, and the reversible jump algorithm is presented. Section 5 is dedicated to the SAEM algorithm, for the estimation of the hyper-parameters of the model.

2. Model and notations

Let $y=(y_t,\ t\geqslant 1)$, be a real process such that, for any $t\geqslant 1$,

$$y_t = s(t) + \varepsilon_t, \tag{1}$$

where $(\varepsilon_t,\ t\geqslant 1)$ is a sequence of zero-mean random variables. Here, the function s to recover is assumed to be piecewise constant. Thus, there exists some instants $(\tau_k,\ k\geqslant 0)$, such that the function s is constant between two successive change-points instants. In other words, there exists a sequence $(m_k,\ k\geqslant 1)$ such that, for any $k\geqslant 1$,

$$s(t) = m_k \quad \text{for all } \tau_{k-1}+1\leqslant t\leqslant \tau_k \tag{2}$$

with the convention $\tau_0=0$.

As already suggested by Lavielle [14] or Tourneret et al. [21,22], it is convenient to introduce a change-points process $(r_t,\ t\geqslant 1)$ that takes the value 1 at the change instants and is zero between two changes:

$$r_t = \begin{cases} 1 & \text{if there exists } k \text{ such that } t=\tau_k,\\ 0 & \text{otherwise.}\end{cases} \tag{3}$$

The estimation of the change-points instants reduces to the estimation of the sequence (r_t). Then, the unknown function s will be recovered by estimating the sequences (r_t) and (m_k).

To solve this inverse problem, we shall adopt a Bayesian approach. That means, we have to define the distribution of the non-observed sequences, conditionally to the set of observations. This distribution is usually called the *posterior* distribution and requires to define first the *prior* distribution of (r_t) and (m_k).

Assume that the observed sequence (y_t) is available between instants $t=1$ and n. First, we consider that (r_t) is a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables with parameter λ. Then, for any $r=(r_t,\ 1\leqslant t\leqslant n-1)$ in $\Omega=\{0,1\}^{n-1}$,

$$\pi(r;\lambda) = \lambda^{\sum_{t=1}^{n-1} r_t}(1-\lambda)^{n-1-\sum_{t=1}^{n-1} r_t}. \tag{4}$$

On the other hand, $(s(t),\ 1\leqslant t\leqslant n)$ is modeled as a sequence of i.i.d. Gaussian random variables with

mean μ and variance V. Then,

$$\pi(s(1),\dots,s(n);\mu,V) = \prod_{t=1}^{n}(2\pi V)^{-1/2}\exp\left\{-\frac{1}{2V}(s(t)-\mu)^2\right\}. \quad (5)$$

For a given configuration of change-points r, $\sum_{t=1}^{n-1} r_t$ is the number of change-points. Then, let $K_r = \sum_{t=1}^{n-1} r_t + 1$ be the number of segments, $n_k = \tau_k - \tau_{k-1}$ be the length of segment k and $m = (m_k, 1 \leqslant k \leqslant K_r)$ be the vector of means. Then,

$$\begin{aligned}\pi(m|r;\mu,V) &= \pi(m_1,\dots,m_{K_r}|r;\mu,V)\\ &= \pi(s(1),\dots,s(n),\ s(t)=m_k,\\ &\qquad \tau_{k-1}+1 \leqslant t \leqslant \tau_k,\\ &\qquad 1 \leqslant k \leqslant K_r;\mu,V)\\ &= \prod_{k=1}^{K_r}\left(\frac{2\pi V}{n_k}\right)^{-1/2}\exp\left\{-\frac{n_k}{2V}(m_k-\mu)^2\right\}.\end{aligned} \quad (6)$$

Thus, the m_k's are independent, and m_k is Gaussian with mean μ and variance V/n_k.

On the other hand, $(\varepsilon_t,\ t \geqslant 1)$ is assumed to be a sequence of independent Gaussian random variables with mean 0 and variance σ^2. Thus, the conditional distribution of the observations is defined by

$$h(y|r,m;\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{K_r}\sum_{t=\tau_{k-1}+1}^{\tau_k}(y_t-m_k)^2\right\}. \quad (7)$$

Let $\theta = (\mu,\ \lambda, V, \sigma^2)$ be the set of hyper-parameters of the model. Then, the prior distribution of (r,m) is given by

$$\pi(r,m;\theta) = \pi(m|r;\ \mu,V)\pi(r;\lambda), \quad (8)$$

the complete likelihood of (y,r,m) is

$$f(y,r,m;\theta) = h(y|r,m;\sigma^2)\pi(m|r;\mu,V)\pi(r;\lambda), \quad (9)$$

and the posterior distribution of (r,m) can be decomposed as

$$p(r,m|y;\theta) = p(r|y;\theta)p(m|y,r;\theta). \quad (10)$$

For a given value of r, the conditional distribution of m is easy to compute. Indeed, let $\bar{y}_k = n_k^{-1}\sum_{t=\tau_{k-1}+1}^{\tau_k} y_t$ be the empirical mean of y in segment k. Then, Eqs. (6) and (7) yield

$$p(m|y,r;\theta) = \prod_{k=1}^{K_r}(2\pi V_k)^{-1/2}\exp\left\{-\frac{1}{2V_k}(m_k-\mu_k)^2\right\}, \quad (11)$$

where

$$V_k = \frac{V\sigma^2}{n_k(V+\sigma^2)} \quad (12)$$

and

$$\mu_k = \frac{V\sigma^2}{V+\sigma^2}\left(\frac{\bar{y}_k}{\sigma^2}+\frac{\mu}{V}\right). \quad (13)$$

Thus, conditionally to the observations, the m_k's remain independent and Gaussian (a short demonstration of these formulae is given in Appendix A).

The following Lemma gives the posterior distribution of r:

Lemma 1. *For any configuration of change-points r, let K_r be the number of segments and let $S_r = \sum_{k=1}^{K_r}\sum_{t=\tau_{k-1}+1}^{\tau_k}(y_t-\bar{y}_k)^2$. Then, the posterior distribution of r is defined by*

$$p(r|y;\theta) = C(y,\theta)\exp\{-\phi S_r - \gamma K_r\} \quad (14)$$

where

$$\phi = \frac{V}{2\sigma^2(\sigma^2+V)},$$

$$\gamma = \frac{1}{2}\log\left(\frac{\sigma^2+V}{\sigma^2}\right) + \log\left(\frac{1-\lambda}{\lambda}\right),$$

and where $C(y,\theta)$ is a normalizing constant.

(The proof of the Lemma is in Appendix A.)

Remark. (1) It is important to insist on the fact that the posterior distribution $p(r|y;\theta)$ is the *joint* distribution of a vector of size $n-1$. Thus, it cannot be used as it stands and should be summarized to some characteristics. Between many others, we can consider the following characteristics:

- For any $1 \leqslant t \leqslant n-1$, the marginal posterior distribution $p(r_t|y;\theta)$ gives the probability to have

a change-point at instant t, conditionally to the observations.

- The MAP estimator is the particular value of r that maximizes $p(r|y;\theta)$. In other words, it is the most likely configuration of change-points, according to the prior and to the observations.
- For any instants τ_a and τ_b, $\mathbb{P}(\sum_{t=\tau_a}^{\tau_b} r_t = k|y;\theta)$ is the probability to have exactly k change-points between these two instants. In particular, when $\tau_a = 1$ and $\tau_b = n-1$, we consider the posterior distribution of the total number of change-points.

(2) The posterior distribution of r can be written

$$p(r|y;\theta) = C(y,\theta)\exp\{-U_\theta(y,r)\}, \tag{15}$$

where $U_\theta(y,r) = \phi S_r + \gamma K_r$ is a penalized contrast, usually called energy function, and which is the sum of two terms: the first term S_r, measures the fidelity to the observations y while the second term K_r corresponds to a penalization term, related to the number of change-points. The coefficients ϕ and γ indicate the relative weights given to these two criteria. A small value of γ in front of ϕ favours configurations with a large number of change-points, while a big value of γ penalizes such configurations. The MAP estimator of r minimizes the energy function $U_\theta(y,r)$. In this particular example, the MAP estimator reduces to a penalized least-squares estimate. We can mention that theoretical results concerning this estimator have been obtained by Lavielle and Moulines [15] under very general conditions.

(3) Unfortunately, the normalizing constant $C(y,\theta)$ in (14) and (15) cannot be computed, since it is the sum over all the possible configurations r of $\exp\{-U_\theta(y,r)\}$, that is, a sum of 2^{n-1} terms. In other words, the posterior distribution of r is known up to this constant and a Markov chain Monte-Carlo method should be used to sample it and estimate some of its characteristics.

3. The Hastings–Metropolis algorithm

3.1. The basic algorithm

The main idea of this algorithm is to generate an ergodic Markov chain $(r^{(i)},\ i \geqslant 0)$ so that $p(\cdot\,|\,y;\theta)$ is its stationary distribution. Then, the ergodic Theorem implies that, for any measurable function f,

$$\bar{f}_N = \frac{1}{N}\sum_{i=1}^{N} f(r^{(i)}) \tag{16}$$

is a strongly consistent estimator of $\mathbb{E}(f(r)|y;\theta)$, i.e. $\bar{f}_N$ converges almost surely to $\mathbb{E}(f(r)|y;\theta)$ when $N \to \infty$ (see, for example, [17]).

An interesting application of this result is the estimation of probabilities of specified events, when f is an indicator function. For example, the marginal posterior distributions of the r_t's and the posterior distribution of the number of segments K_r, can easily be estimated. Indeed, for any $1 \leqslant t \leqslant n-1$ and any $k \geqslant 0$,

$$\frac{1}{N}\sum_{i=1}^{N} r_t^{(i)} \to \mathbb{P}(r_t = 1|y;\theta) \quad \text{a.s.} \tag{17}$$

$$\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{(K_r^{(i)}=k)} \to \mathbb{P}(K_r = k|y;\theta) \quad \text{a.s.}, \tag{18}$$

where $K_r^{(i)} = \sum_{t=1}^{n-1} r_t^{(i)} + 1$ is the number of segments in the configuration $r^{(i)}$.

The Hastings–Metropolis algorithm is an iterative procedure. At iteration i, we carry out the following two steps:

- an admissible new value $\tilde{r}$ is drawn from a *proposal kernel* $q(r^{(i)},\tilde{r})$
- $\tilde{r}$ is accepted as the new state, i.e. $r^{(i+1)} = \tilde{r}$, with the following probability:

$$\alpha(r^{(i)},\tilde{r}) = \min\left\{1, \frac{p(\tilde{r}|y;\theta)}{p(r^{(i)}|y;\theta)}\frac{q(\tilde{r},r^{(i)})}{q(r^{(i)},\tilde{r})}\right\}. \tag{19}$$

Remark. (1) If the kernel q is irreducible, then the Markov chain $(r^{(i)})$ is irreducible. Furthermore, the aperiodicity of the chain is ensured if there exists two configurations (r,r') such that $\alpha(r,r') < 1$. Under these conditions, the chain $(r^{(i)})$ is uniformly ergodic, since it takes its values in a finite space.

(2) An initial *burn-in* period is introduced before collecting samples, so that the estimation weakly depends on the initial guess (see [20]). If N_b is the length of this burn-in period, then, the estimator of

$\mathbb{E}(f(r)|r;\theta)$ proposed in (16) is replaced by

$$\bar{f}_N = \frac{1}{N}\sum_{i=N_b+1}^{N_b+N} f(r^{(i)}) \tag{20}$$

(3) For any (r,r'), let $\Delta U(r,r') = U_\theta(y,r') - U_\theta(y,r)$. Then, (15) yields

$$\frac{p(r'|y;\theta)}{p(r|y;\theta)} = \mathrm{e}^{-\Delta U(r,r')}. \tag{21}$$

Since the energy $U_\theta(y,r)$ is a sum of local potentials, a local perturbation of the current state $r^{(i)}$ will affect few terms of this sum and the probability of acceptance $\alpha(r^{(i)}, \tilde{r})$ will be easy to compute.

3.2. The proposal kernels

As it is mentioned just above, any irreducible proposal kernel q can be used. From a practical point of view, it is important to allow more communications between the states of high probability in order to increase the convergence speed. In our example of application, that can be done by using successively the following three kernels at each iteration:

1. q_1 is such that the candidate $\tilde{r}$ is drawn independently of the current state r: $q_1(r,\tilde{r}) = \pi(\tilde{r};\theta)$. Let $\beta = \frac{1}{2}\log((\sigma^2+V)/\sigma^2)$. Then, we obtain
$$\alpha(r,\tilde{r}) = \min\{1, \exp\{-\phi(S_{\tilde{r}} - S_r) - \beta(K_{\tilde{r}} - K_r)\}\}. \tag{22}$$
2. q_2 is such that a new change-point is created or an existing change-point is removed. An instant s is chosen randomly in $\{1,\dots,n-1\}$ and we set $\tilde{r}_t = r_t$ for all $t \neq s$ while $\tilde{r}_s = 1 - r_s$. The acceptance probability turns out to be
$$\alpha(r,\tilde{r}) = \min\{1, \exp\{-\phi[S_{\tilde{r}} - S_r] \pm \gamma\}\}. \tag{23}$$
3. With the third considered kernel q_3, an existing change-point instant is moved. Two instants (s,s') are randomly chosen such that $r_s = 1$ and $r_{s'} = 0$. Then, $\tilde{r}_t = r_t$ for all $t \neq s,s'$ while $\tilde{r}_s = 0$ and $\tilde{r}_{s'} = 1$. In this case, the acceptance probability is
$$\alpha(r,\tilde{r}) = \min\{1, \exp\{-\phi(S_{\tilde{r}} - S_r)\}\}. \tag{24}$$

We propose an example to illustrate this algorithm. We simulate a sequence $y = (y_1,\dots,y_n)$ with $n = 500$. There are four change-points at $\tau_1 = 75$, $\tau_2 = 150$, $\tau_3 = 250$, and $\tau_4 = 400$. The vector of mean is $m = (0.125, 0.5, 0.4, 0.5, 0.125)$. The variance of the additive noise is $\sigma^2 = 0.1$. The observed series and the mean are plotted in Fig. 1(a) and (b).

First of all, because the set of hyper-parameters θ is unknown, it is estimated by using the SAEM procedure described in Section 5. Then, the estimated value $\hat{\theta} = (\hat{\lambda},\hat{\mu},\hat{V},\hat{\sigma}^2) = (0.012, 0{,}346, 2.688, 0.106)$ is used in the MCMC algorithm. We run the MCMC algorithm with 5000 burn-in iterations. The estimations of the marginal posterior probabilities $\{\mathbb{P}(r_t = 1|y;\theta)\}$ obtained after 15 000 and 150 000 iterations are plotted in Fig. 2. The posterior distribution of the number of segments K_r is displayed Fig. 2c (estimated after 150 000 iterations).

First of all, we can remark that the estimations obtained after 15 000 iterations are closed to those obtained after 150 000 iterations. That means that this algorithm converges quite quickly, and only "few iterations" are enough to detect very well the four change-points. Theoretical aspects concerning the convergence control of MCMC methods can be found in [20].

These diagrams can be seen as histograms around each change-points. For example, we obtain a very accurate estimation of the position of the first change-point (at 75) since the estimated posterior distribution of r is very spiky around this instant: the estimates of $\mathbb{P}(r_{75} = 1|y;\theta)$ and $\mathbb{P}(r_{76} = 1|y;\theta)$, obtained with 150 000 iterations, are, respectively, 0.42 and 0.49. On the other hand, the jumps of the mean are smaller at 150 and 250, and the detection of these two change-points is not so accurate: the estimated marginal probabilities are very small around 150 and 250 (around 0.1). Nevertheless, the probability of a change-point is very high in a neighborhood of these two instants. For example, the estimated probability to have a change point in the interval $[135,165]$ (resp. $[235,265]$) is 0.85 (resp. 0.77).

In other words, it is not convenient to apply directly a threshold on the sequence of estimated marginal probabilities $\{\mathbb{P}(r_t = 1|y;\theta)\}$, for detecting the change-points. A first solution consists in estimating the probability to have a change-point in an interval, instead of an instant. Of course, the

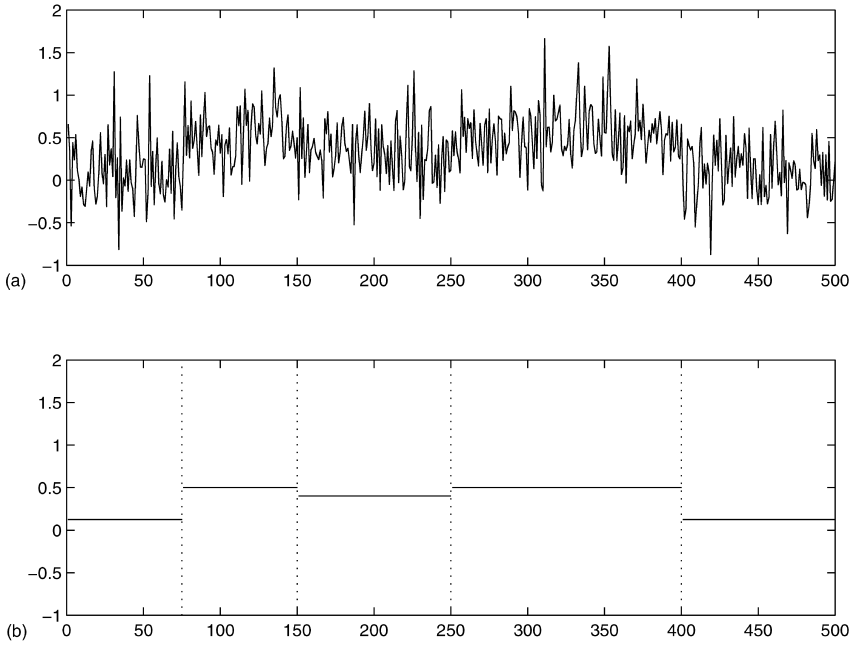

Fig. 1. (a) The observed signal y and (b) the mean of y and the change-point instants.

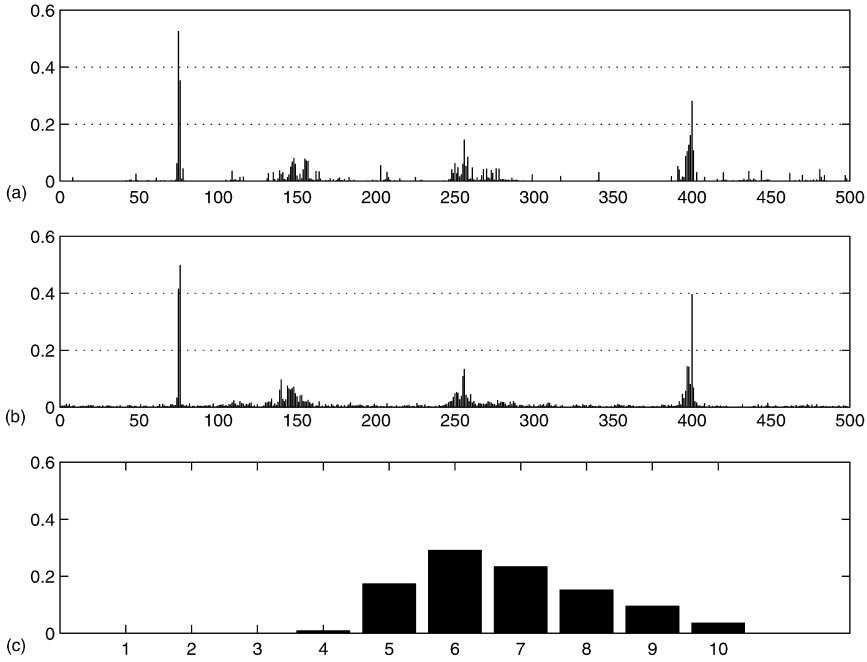

Fig. 2. The posterior distribution of r estimated with the Hastings–Metropolis algorithm. The marginal distributions $\{\mathbb{P}(r_t = 1|y;\theta),\ 1 \leqslant t \leqslant n-1\}$ estimated with: (a) 15 000 iterations, (b) 150 000 iterations and (c) the posterior distribution of the number K_r of segments.

positions of the change-points will not be precisely estimated with this method. To overcome this loss of accuracy, a second approach consists in estimating the configurations of change-points of higher probability. That can easily be done, by using a slight variation of the Hastings–Metropolis algorithm.

3.3. Running the Hastings–Metropolis algorithm at a low temperature

For any $T > 0$, we can consider the distribution $p_T(\cdot\,|y;\theta)$, based on the original distribution $p(\cdot\,|y;\theta)$, and defined as follows:

$$p_T(r|y;\theta) = C_T(y;\theta)\exp\left\{-\frac{U_\theta(y,r)}{T}\right\} \tag{25}$$

$$= C_T(y;\theta)\exp\left\{-\frac{\phi}{T}S_r - \frac{\gamma}{T}K_r\right\}. \tag{26}$$

The role of the parameter T (usually called temperature) is mainly to discriminate the global and the local maxima of the posterior distribution $p(\cdot\,|y;\theta)$. Indeed, any maximum (local or global) of $p(\cdot\,|y;\theta)$ is a minimum of $U_\theta(y,r)$, and also a maximum of $p_T(\cdot\,|y;\theta)$. However, $p_T(r|y;\theta) \to 0$ when $T \to 0$ if r is not a global maximum. Thus, when $T \to 0$, $p_T(\cdot\,|y;\theta)$ converges to the uniform distribution on the set of global maxima of $p(\cdot\,|y;\theta)$.

The Hastings–Metropolis algorithm described above can be used for simulating $p_T(\cdot\,|y;\theta)$. Indeed, from (26), the only modification to introduce in the algorithm is the replacement of the parameters (ϕ,γ) by $(\phi/T,\gamma/T)$.

The simulated annealing algorithm consists in using a sequence of temperatures (T_i) that decreases at each iteration. Then, the Markov chain $(r^{(i)})$ is no longer homogeneous, and converges to the global maxima of $p(\cdot\,|y;\theta)$ (the MAP estimator) if T_i behaves like $T_0/\log(i)$ (see [10,11]). Unfortunately, this schedule of temperature cannot be used in the practice, since it would require a very large number of iterations.

In fact, there exists a very efficient and attractive alternative for detecting the most likely configuration of change-points. It consists merely in running the Hastings–Metropolis algorithm at a fixed low temperature T.

We can observe the influence of a low temperature on the chain's construction during the algorithm: consider two states r and r' such that $U_\theta(y,r) < U_\theta(y,r')$. When T tends to 0, $\exp\{-(U_\theta(y,r') - U_\theta(y,r))/T\}$ tends to 0. So $\alpha(r,r')$ tends to 0 and a move from r to r' has a very low probability. Consequently, when T is a low temperature, the Hastings–Metropolis algorithm will favor the configurations of change-points of highest probabilities.

To set $T = 0$ leads to the so-called iterative conditional modes (ICM) algorithm (see [14]). This deterministic procedure usually leads to a local minima of the posterior distribution of r.

On the other hand, for any $T > 0$, the Markov chain $(r^{(i)})$ simulated with this algorithm remains homogeneous and ergodic: its distribution converges to $p_T(\cdot\,|y;\theta)$.

Fig. 3 shows the results obtained with three temperatures: $T = 0.5$, 0.2 and 0.01. Looking at these results, we can make two main remarks:

1. The false alarms are removed, and only the main events are left. Even a "high" temperature, such as $T = 0.5$, cleans the results. With $T = 0.5$, the estimated probability to have five segments is 0.97. This estimated probability is 1 for $T \leqslant 0.3$. That clearly shows that the most likely configurations are made up of five segments.
2. When the temperature decreases, the posterior distribution becomes more and more concentrated around the MAP estimator of the change-points instants. In this example, the MAP is $\hat{\tau} = (\hat{\tau}_1,\hat{\tau}_2,\hat{\tau}_4,\hat{\tau}_4) = (76, 147, 256, 400)$.

4. The estimation of the mean

Assume now that we are interested in the joint posterior distribution of the mean sequence $m = (m_k)$ and the change-points instants $\tau = (\tau_k)$ (or $r = (r_t)$, according to the parametrization), instead of the posterior distribution of τ. One iteration of the Hastings–Metropolis algorithm will consist now in drawing a candidate $(\tilde{m},\tilde{\tau})$ (or $(\tilde{m},\tilde{r})$), with a new proposal b, and to accept it with a probability $\alpha((m^{(i)},\tau^{(i)}),(\tilde{m},\tilde{\tau}))$. Different approaches can be adopted for choosing a proposal b.

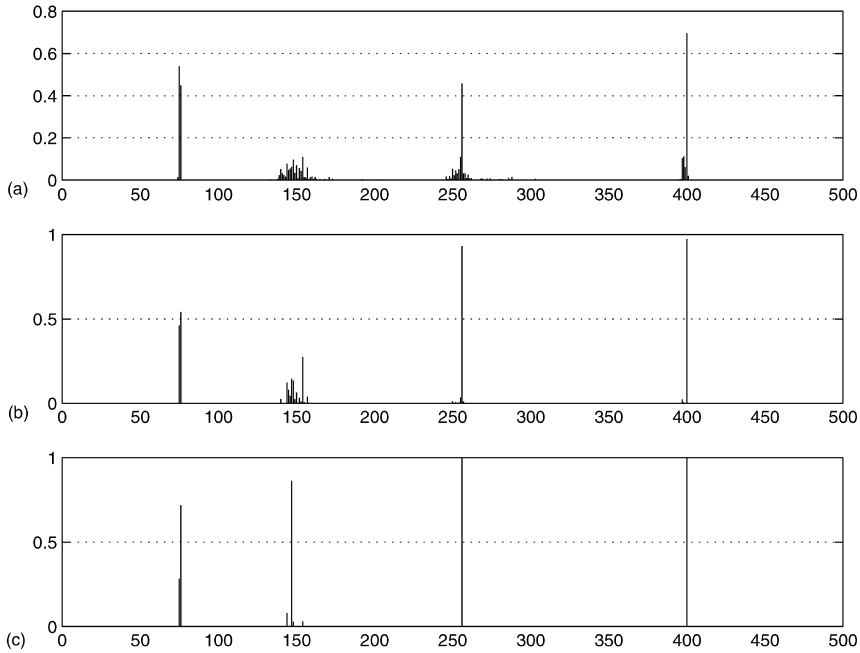

Fig. 3. Running the Hastings–Metropolis algorithm at a low temperature: (a) $T = 0.5$, (b) $T = 0.2$ and (c) $T = 0.01$.

4.1. The reversible jump MCMC algorithm

A first approach is the so-called reversible jump algorithm, proposed by Green [12]. This method is an adaptation of MCMC algorithms, when the dimensionality of the parameter vector is not fixed. Then, the Markov chain can "jump" between models with parameter spaces of different dimensions. As before, different kernels should be used. For example, we can make use of the four following moves:

1. One of the mean m_k is randomly changed. The proposed mean $\tilde{m}_k$ is such that $\log(\tilde{m}_k/m_k)$ is uniformly distributed on $[-\frac{1}{2},\frac{1}{2}]$, in order to avoid big jumps.
2. A change-point is added in segment k. The position $\tilde{\tau}_k$ is drawn uniformly in $[\tau_{k-1}+1,\tau_k-1]$. The mean m_k is split into two means $\tilde{m}_k$ and $\tilde{m}_{k+1}$ such that

$$(\tilde{\tau}_k - \tau_{k-1})\tilde{m}_k + (\tau_k - \tilde{\tau}_k)\tilde{m}_{k+1} = n_k m_k,$$

where $n_k = \tau_k - \tau_{k-1}$ is the length of segment k. This condition is satisfied with

$$\tilde{m}_k = m_k - u\sqrt{\frac{\tilde{n}_{k+1}}{\tilde{n}_k}} \quad \text{and}$$

$$\tilde{m}_{k+1} = m_k + u\sqrt{\frac{\tilde{n}_k}{\tilde{n}_{k+1}}},$$

where u is uniformly distributed on the interval $[-0.2,0.2]$, for the same reason as in the first move.

3. A change-point τ_k is removed. Then, the means m_k and m_{k+1} are replaced by a unique $\tilde{m}_k$ such that

$$(n_k + n_{k+1})\tilde{m}_k = n_k m_k + n_{k+1} m_{k+1}.$$

4. A change-point τ_k is moved. A new position $\tilde{\tau}_k$ is drawn uniformly on $[\tau_{k-1}+1,\tau_{k+1}-1]$ and the means remain unchanged.

At each iteration, an independent random choice is made between these four move types. These have probabilities 0.3 for the moves 1 and 4, and 0.2 for two others.

Following Green [12], the probabilities of acceptance can be computed for these different kernels. The formulae are given in Appendix B. We applied this algorithm on the same series y displayed Fig. 1a. Fig. 4 presents the estimation of the probabilities $\{\mathbb{P}(r_t = 1|y;\theta)\}$ after 15 000 and 150 000 iterations. Comparing Fig. 4a with Fig. 2a, we remark that the reversible jump algorithm converges much more slowly than the Hastings–Metropolis algorithm described in the previous section. Indeed, we are now simulating a pair of variables (m, r) instead of simulating only r. The introduction of a new (continuous) variable to sample slows down the algorithm.

One explanation is the fact that the reversible jump algorithm does not take use of the natural hierarchy

$$p(m, r|y;\theta) = p(r|y;\theta)p(m|r, y;\theta) \tag{27}$$

for its proposal kernels, and many candidates $(\tilde{m},\tilde{r})$ are rejected. Then, a big amount of iterations are required for estimating correctly the posterior distribution of interest.

4.2. An hybrid algorithm

A second approach consists in combining the Hastings–Metropolis algorithm described in Section 3.1 for simulating r, with the Gibb's sampler [19] for simulating m.

The proposal kernels are defined by

$$b((m^{(i)},\tau^{(i)}), (\tilde{m},\tilde{\tau})) = q_i(\tau^{(i)},\tilde{\tau})p(\tilde{m}|\tilde{\tau}, y;\theta), \tag{28}$$

where q_i is one of the proposal kernels defined in Section 3.2. Then, the probability of acceptance is

$$\alpha(m^{(i)},\tau^{(i)},(\tilde{m},\tilde{\tau})) = \min\left\{1,\frac{p(\tilde{m},\tilde{\tau}|y;\theta)}{p(m^{(i)},\tau^{(i)}|y;\theta)}\frac{b((\tilde{m},\tilde{\tau}),(m^{(i)},\tau^{(i)}))}{b((m^{(i)},\tau^{(i)}),(\tilde{m},\tilde{\tau}))}\right\} \tag{29}$$

$$= \min\left\{1,\frac{p(\tilde{\tau}|y;\theta)}{p(\tau^{(i)}|y;\theta)}\frac{q_i(\tilde{\tau},\tau^{(i)})}{q_i(\tau^{(i)},\tilde{\tau})}\right\}. \tag{30}$$

That means that the probability of acceptance does not depend on the mean vectors $m^{(i)}$ and $\tilde{m}$, but only on the configurations of changes $r^{(i)}$ and $\tilde{r}$. In other words, we use the Hastings–Metropolis

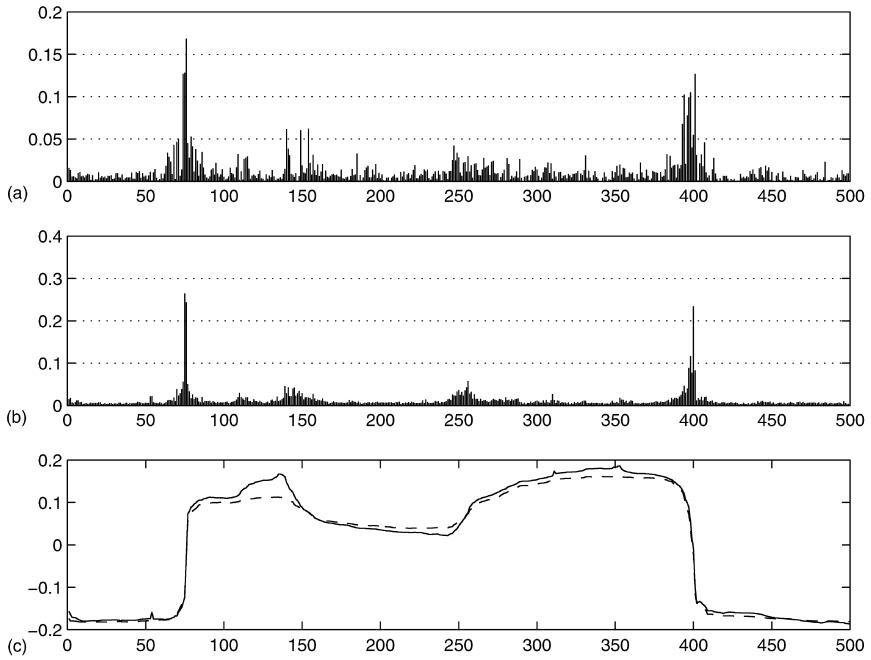

Fig. 4. The posterior distributions of (r, m) estimated with the Reversible Jump algorithm. The marginal distributions $\{\mathbb{P}(r_t = 1|y;\theta),\ 1 \leqslant t \leqslant n-1\}$ estimated with: (a) 15 000 iterations, (b) 150 000 iteration, (c) (-) the posterior mean of m; (- -) the posterior mean of m, conditionaly to $K_r = 5$.

algorithm described in Section 3.1 for generating the sequence $(r^{(i)})$, while $m^{(i)}$ is drawn at iteration i with the conditional distribution $p(m|r^{(i)}, y;\theta)$. This algorithm was shown to converge much more faster than the Reversible Jump algorithm, since a very good approximation of the marginal posterior probabilities $\{\mathbb{P}(r_t = 1|y;\theta)\}$ is obtained after only 15 000 iterations (see Fig. 2a).

4.3. What can we do with a joint distribution?

These algorithms produce ergodic Markov chains $(s^{(i)}, r^{(i)})$ that converge to the joint posterior distribution $p(s, r|y;\theta)$. Once more, this joint distribution cannot be described completely, but should be reduced to some interesting and tractable characteristics. Most of the times, MCMC is only used for estimating the posterior mean of the non observed variable. In our context, that would mean to estimate $\mathbb{E}(s(t)|y;\theta)$, $1 \leqslant t \leqslant n$, by the empirical mean $N^{-1}\sum_{i=N_b+1}^{N_b+N} s(t)^{(i)}$ (or eventually, by using the Rao–Blackell version described in [6,19]). This estimated posterior mean is displayed Fig. 4.

Unfortunately, this posterior mean is uninteresting in our context. Indeed, let $\Omega = \{0,1\}^{n-1}$ be the set of possible configurations of change-points. Then, the marginal posterior distribution $p(s|y;\theta)$ is the sum of the joint posterior distributions $p(s, r|y;\theta)$, over *all* the possible configurations:

$$p(s|y;\theta) = \sum_{r\in\Omega} p(s, r|y;\theta) \tag{31}$$

and the posterior mean can also be decomposed as weighted sum of conditional means, over *all* the possible configurations:

$$\mathbb{E}(s|y;\theta) = \sum_{r\in\Omega} \mathbb{E}(s|r, y;\theta)p(r|y;\theta). \tag{32}$$

That means, we are mixing configurations without any change-points with configurations with one, two, 10 or more change-points. Then, the meaning of this posterior mean is not obvious at all.

Green proposes to estimate this posterior mean, conditionally to a given number of segments (or to a given number of change-points). Such an example is presented in Fig. 4, for five segments. The estimated mean remains smooth, since we are now integrating over all the configurations with five segments. Actually, this curve can be seen as a smooth version of the original data. It is a little bit embarrassing to obtain a smooth function, when we are looking for a step function.

Another approach consists in estimating m, conditionally to a given configuration of change-points r. That is very easy, since the conditional distribution $p(s|r, y;\theta)$ is a Gaussian distribution with known parameters. For example, it seems natural to consider the most likely configuration of change-points, that is, the MAP estimate of r. Then, conditionally to this particular configuration, $(m_1, m_2, m_3, m_4, m_5)$ is Gaussian with mean (0.106, 0.534, 0.338, 0.548, 0.114) and variance $(13, 14, 9, 7, 10)\times 10^{-3}$.

5. Estimation of θ using SAEM algorithm

The implementation of an MCMC algorithm as described above, assumes that the set of parameters of the model is known. Recall that these hyper-parameters are, respectively, the prior proportion of change-points λ, the parameters μ and V of the Gaussian distribution for the vector of means m, and σ^2 the variance of the additive noise. Instead of setting the hyper-parameters θ to a particular value, as it is usually done in a Bayesian framework, we propose to estimate θ.

The maximum likelihood estimator (MLE) of θ maximizes the likelihood of the observed data $g(y;\theta)$. Unfortunately, the MLE cannot be computed in a closed form in a context of incomplete data. The SAEM is well suitable for computing the MLE in this kind of situation, see [5,13] for some examples of application. This stochastic version of the expectation maximization (EM), [9] algorithm just consists in updating the estimate of the hyper-parameters θ at each iteration of the MCMC algorithm described above. This update is based on a stochastic approximation of the minimal sufficient statistics of the complete data model (r, y). Thus, the first thing to do is to write the complete likelihood $f(r, y;\theta)$ in a standard exponential form. We have the following Lemma:

Lemma 2. *For any configuration of changes* r, *let* $\bar{y}_k = n_k^{-1}\sum_{t=\tau_{k-1}+1}^{\tau_k} y_t$, $\bar{y} = n^{-1}\sum_{t=1}^{n} y_t$ *and*

$S_r = \sum_{k=1}^{K_r} \sum_{t=\tau_{k-1}+1}^{\tau_k} (y_t - \bar{y}_k)^2$. *Then, the likelihood of the complete data is defined by*

$$f(y,r;\theta) = (2\pi\sigma^2)^{-n/2} \left(\frac{\sigma^2+V}{\sigma^2}\right)^{-K_r/2} \lambda^{K_r-1}(1-\lambda)^{n-K_r} \times \exp\left\{ -\frac{1}{2(V+\sigma^2)} \times \left(\sum_{t=1}^{n} (y_t-\mu)^2 + \frac{V}{\sigma^2} S_r \right) \right\} \quad (33)$$

and the maximum likelihood estimator of θ (*i.e. the value of* θ *that maximizes the complete likelihood* $f(y,r;\cdot)$) *is* $\tilde{\theta} = (\tilde{\mu},\tilde{\lambda},\tilde{\sigma}^2,\tilde{V})$, *where*

$$\tilde{\mu} = \bar{y}, \quad (34)$$

$$\tilde{\lambda} = \frac{K_r - 1}{n-1}, \quad (35)$$

$$\tilde{\sigma}^2 = \frac{S_r}{n-K_r}, \quad (36)$$

$$\tilde{V} = \frac{\sum_{t=1}^{n} (y_t - \bar{y})^2 - S_r}{K_r} - \tilde{\sigma}^2. \quad (37)$$

(The proof of this Lemma is in Appendix A.)

Remark. (1) The maximum likelihood of μ is $\bar{y}$ and therefore does not depend on the non-observed data r. Thus, $\bar{y}$ is also the value of μ that maximizes the observed likelihood $g(y;\theta)$. The SAEM algorithm will be used for estimating the others parameters λ, σ^2 and V.

(2) Just like in an analysis of variance (ANOVA) context, the maximum likelihood of σ^2 is the empirical residual variance $(n-K_r)^{-1} \sum_{k=1}^{K_r} \sum_{t=\tau_{k-1}+1}^{\tau_k} (y_t - \bar{y}_k)^2$, that is, the total sum of residual squares S_r divided by the degrees of freedom $n-K_r$ (see [18, pp. 185–186]).

(3) Considering the observed series y as a constant, the maximum likelihood estimator of (λ,σ^2,V) only depends on the missing sequence r via the two statistics K_r and S_r. As we shall see just below, the SAEM algorithm is based on this remark.

The proposed SAEM algorithm is an iterative algorithm that requires an initial configuration of change-points $r^{(0)}$ and an initial guess $\theta^{(0)}$. Then, at iteration i, a simulation step and an estimation step are performed as follows:

- *Simulation step*: a new configuration $r^{(i)}$ is generated with M iterations of the MCMC algorithm, using the current values of the hyper-parameters $\theta^{(i-1)}$ and the current configuration $r^{(i-1)}$.
- *Estimation step*: $\theta^{(i)}$ is updated by using the new configuration $r^{(i)}$, and according to the two following steps:

1. *Stochastic approximation*: update the approximation of the sufficient statistics as follows:

$$s_1^{(i)} = s_1^{(i-1)} + a_i(K_{r^{(i)}} - s_1^{(i-1)}), \quad (38)$$

$$s_2^{(i)} = s_2^{(i-1)} + a_i(S_{r^{(i)}} - s_2^{(i-1)}), \quad (39)$$

where $K_{r^{(i)}}$ and $S_{r^{(i)}}$ are the sufficient statistics of the complete model, computed at the point $(y,r^{(i)})$ and where (a_i) is a sequence of decreasing stepsizes such that $\sum a_i = \infty$ and $\sum a_i^2 \leqslant \infty$.

2. *Maximization step*: compute $\theta^{(i)} = (\lambda^{(i)},\sigma^{2(i)},V^{(i)})$ by maximizing the complete likelihood (see (35)–(37)):

$$\lambda^{(i)} = \frac{s_1^{(i)} - 1}{n-1}, \quad (40)$$

$$\sigma^{2(i)} = \frac{s_2^{(i)}}{n - s_1^{(i)}}, \quad (41)$$

$$V^{(i)} = \frac{\sum_{t=1}^{n} (y_t - \bar{y})^2 - s_2^{(i)}}{s_1^{(i)}} - \sigma^{2(i)}. \quad (42)$$

Remarks. (1) We choose a decreasing sequence (a_i) in order to obtain a pointwise convergence of the sequence $(\theta^{(i)})$ to a value $\theta^\star$ (see [10] for results concerning stochastics algorithms and the many references therein). A satisfactory schedule consists in setting $a_i = 1$ during some iterations (about 10 iterations in the practice), for converging quickly to a neighborhood of $\theta^\star$ and then, (a_i) decreases as $(1/i)$.

(2) It was shown by Delyon et al. [8] that SAEM onverges to a (local or global) maximum of the observed data likelihood $g(y;\theta)$ under very general conditions, but assuming exact and independent simulations of the missing data at each iterations. Here, the sequence of missing data $(r^{(i)})$ is a Markov chain, and this result does not apply directly. Nevertheless, by using the results of Metivier and

Priouret [16] for this kind of situation, we can show that the algorithm described above converges to a maximum of $g(y;\theta)$ if the sequence of parameters $(\theta^{(i)})$ belongs to a compact set. Then, the slight technical stabilization device proposed by Delyon et al. [8] for the SAEM algorithm ensures the compactness of $(\theta^{(i)})$, and its convergence to a maximum of the observed likelihood.

We propose in Fig. 5 a numerical example of this algorithm. A series y of length 1000 was simulated with the following parameters: $\lambda = 0.01$, $V = 1$ and $\sigma^2 = 0.1$. The change-points sequence r and the vector of means m were obtained using (4) and (6). Then, y was obtained using (7). This series is displayed Fig. 5(a) together with the sequence of means and the change-points. The number of iterations of MCMC to perform before updating $\theta^{(i)}$ was fixed to $M = 1000$. The sequence of step-sizes (a_i) was such that $a_i = 1$ for $1 \leqslant i \leqslant 10$, and $a_i = 1/(i - 10)$ for $i \geqslant 11$. The sequences $(\lambda^{(i)})$, $(V^{(i)})$ and $(\sigma^{2^{(i)}})$ are displayed Fig. 5(b)–(d). The algorithm quickly converges, and after 30 iterations, the estimated parameters are $\lambda^{(30)} = 0.007$, $V^{(30)} = 1.469$ and $\sigma^{2^{(30)}} = 0.091$.

6. Conclusion

We have proposed an attractive methodology for the change-points problem, in a Bayesian context. The probabilistic model makes use of an non-observed sequence r, and a MCMC algorithm can be used for estimating the posterior distribution of this change-points process r. Numerical experiments have clearly shown that this procedure is much more faster than the Reversible Jump algorithm. Furthermore, the hyperparameters of the model are estimated, rather than arbitrarily chosen. We have also seen that a slight modification of the sampler allows to select the most likely configurations of change-points.

The main advantage of this method is the ability to perform automatically different tasks. We think that this kind of approach should not be restricted to the problem of detecting change-points in a

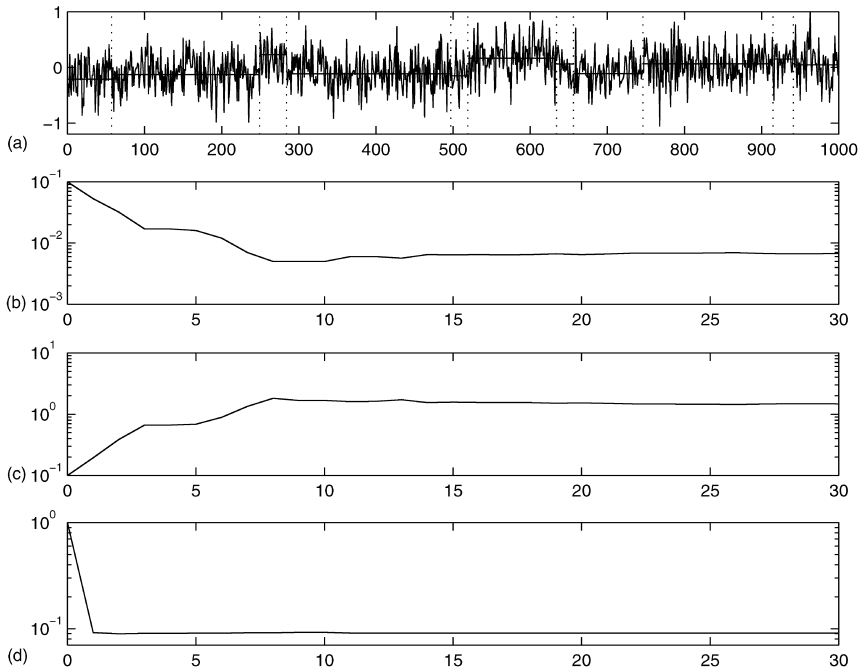

Fig. 5. Estimation of the hyper-parameters. (a) the signal y, the mean m and the change-point instants τ, simulated with $\lambda = 0.01$, $V = 1.5$ and $\sigma^2 = 0.1$, (b), (c) and (d) the sequences of estimates $(\lambda^{(i)})$, $(V^{(i)})$ and $(\sigma^{2^{(i)}})$.

signal contaminated by an additive (or a multiplicative) noise. Indeed, it should be interesting and useful to extend this approach for detecting changes in the spectrum of a signal, for example.

Appendix A

Proof of formulae (11)–(13). Using Eqs. (6) and (7), we have

$$\begin{aligned} &p(m|y,r;\theta)h(y|r;\theta) \\ &\quad = h(y|r,m;\sigma^2)\pi(m|r;\mu,V) \\ &\quad = (2\pi\sigma^2)^{-n/2}\mathrm{e}^{-(1/2\sigma^2)\sum_{k=1}^{K_r}\sum_{t=\tau_{k-1}+1}^{\tau_k}(y_t-m_k)^2} \\ &\qquad \times \prod_{k=1}^{K_r}\left(\frac{2\pi V}{n_k}\right)^{-1/2}\mathrm{e}^{-(n_k/2V)}(m_k-\mu)^2 \\ &\quad = \prod_{k=1}^{K_r}(2\pi V_k)^{-1/2}\mathrm{e}^{-(1/2V_k)(m_k-\mu_k)^2} \\ &\qquad \times (2\pi\sigma^2)^{-n/2}\prod_{k=1}^{K_r}\left(\frac{V}{n_kV_k}\right)^{-1/2} \\ &\qquad \times \exp\left\{-\frac{1}{2}\left(\frac{\sum_{t=\tau_{k-1}+1}^{\tau_k}y_t^2}{\sigma^2}+\frac{n_k\mu^2}{V}-\frac{\mu_k^2}{V_k}\right)\right\}, \end{aligned} \tag{A.1}$$

where

$$\mu_k = \frac{V\sigma^2}{V+\sigma^2}\left(\frac{\bar{y}_k}{\sigma^2}+\frac{\mu}{V}\right)$$

and

$$V_k = \frac{V\sigma^2}{n_k(V+\sigma^2)}.$$

By identification, we obtain the distribution of m conditional on a given configuration r and the observations y

$$p(m|y,r;\theta) = \prod_{k=1}^{K_r}(2\pi V_k)^{-1/2}\mathrm{e}^{-(1/2V_k)}(m_k-\mu_k)^2,$$

where μ_k and V_k are, respectively, the posterior mean and variance of m_k.

Proof of Lemma 1. First, remark that according to (A.1)

$$\begin{aligned} h(y|r;\theta) &= (2\pi\sigma^2)^{-n/2}\prod_{k=1}^{K_r}\left(\frac{V}{n_kV_k}\right)^{-1/2} \\ &\qquad \times \exp\left\{-\frac{1}{2}\left(\sum_{t=\tau_{k-1}+1}^{\tau_k}\frac{y_t^2}{\sigma^2}+\frac{n_k\mu^2}{V}-\frac{\mu_k^2}{V_k}\right)\right\} \\ &= (2\pi\sigma^2)^{-n/2}\left(\frac{\sigma^2+V}{\sigma}\right)^{-K_r/2}\exp\left\{-\frac{1}{2(V+\sigma^2)}\right. \\ &\qquad \left.\times\left(\sum_{t=1}^{n}(y_t-\mu)^2+\frac{V}{\sigma^2}S_r\right)\right\}. \end{aligned} \tag{A.2}$$

We can then obtain Lemma 1 using (4) and (A.2). Indeed,

$$\begin{aligned} p(r|y;\theta) &= \frac{h(y|r;\theta)\pi(r;\lambda)}{g(y;\theta)} \\ &= C(y;\theta)\exp\{-\phi S_r-\gamma K_r\}, \end{aligned} \tag{A.3}$$

where

$$\phi = \frac{V}{2\sigma^2(\sigma^2+V)},$$

$$\gamma = \frac{1}{2}\log\left(\frac{\sigma^2+V}{\sigma^2}\right)+\log\left(\frac{1-\lambda}{\lambda}\right).$$

Proof of Lemma 2. Since $f(y,r;\theta) = h(y|r;\theta)\pi(r;\lambda)$, we can directly obtain (33) from (4) and (A.2). Then, the expression of the maximum likelihood estimate of θ is obtained by maximizing $f(y,r;\theta)$ with respect to θ.

Appendix B

Probabilities of acceptance for the Reversible Jump algorithm. Let (m,τ) be the current state and $(\tilde{m},\tilde{\tau})$ be the proposed candidate. The probability of acceptance is

$$\begin{aligned} \alpha((m,\tau),(\tilde{m},\tilde{\tau})) = \min\left\{1, \frac{p(\tilde{m},\tilde{\tau}|y;\theta)}{p(m,\tau|y;\theta)}\times\frac{j((\tilde{m},\tilde{\tau}),(m,\tau))q_2(u')}{j((m,\tau),(\tilde{m},\tilde{\tau}))q_1(u)}\right. \\ \left.\times\left|\frac{\partial(\tilde{m},\tilde{\tau},u')}{\partial(m,\tau,u)}\right|\right\} \end{aligned}$$

$$= \min\left\{1, \frac{h(y|\tilde{m},\tilde{\tau};\sigma^2)}{h(y|m,\tau;\sigma^2)} \times \frac{\pi(\tilde{m},\tilde{\tau};\theta)}{\pi(m,\tau;\theta)} \times \frac{j((\tilde{m},\tilde{\tau}),(m,\tau))q_2(u')}{j((m,\tau),(\tilde{m},\tilde{\tau}))q_1(u)} \times \left|\frac{\partial(\tilde{m},\tilde{\tau},u')}{\partial(m,\tau,u)}\right|\right\}$$

$$= \min\{1, A\},$$

where

- $j((\tilde{m},\tilde{\tau}),(m,\tau))$ (resp. $j((m,\tau),(\tilde{m},\tilde{\tau}))$) is the probability of choosing the move from $(\tilde{m},\tilde{\tau})$ to (m,τ) (resp. from (m,τ) to $(\tilde{m},\tilde{\tau})$).
- u (resp. u') is generated from the proposal density $q_1(u)$ (resp. $q_2(u')$) such that $(\tilde{m},\tilde{\tau},u') = f(m,\tau,u)$ where f is a specific invertible function.
- the final term is the Jacobian arising from the change of variables from (m,τ,u) to $(\tilde{m},\tilde{\tau},u')$.

We can compute A for the different moves:

Move 1:

$$A = \frac{h(y|\tilde{m},\tilde{\tau};\theta)}{h(y|m,\tau;\theta)} \times \exp\left\{-\frac{1}{2V}(\tilde{m}_k^2 - m_k^2)(\tau_k - \tau_{k-1})\right\} \times \frac{\tilde{m}_k}{m_k}.$$

Move 2:

$$A = \frac{h(y|\tilde{m},\tilde{\tau};\theta)}{h(y|m,\tau;\theta)} \times \frac{\lambda}{1-\lambda} \times (2\pi)^{-1/2}\left(\frac{(\tau_k - \tilde{\tau}_k)(\tilde{\tau}_k - \tau_{k-1})}{(\tau_k - \tau_{k-1})}\right)^{1/2} \times \exp\left\{-\frac{1}{2V}(\tilde{m}_k^2(\tilde{\tau}_k - \tau_{k-1}) + \tilde{m}_{k+1}^2(\tau_k - \tilde{\tau}_k) - m_k^2(\tau_k - \tau_{k-1}))\right\} \times \frac{0.4(n-1)}{k}\left(\frac{\tau_k - \tau_{k-1}}{((\tilde{\tau}_k - \tau_{k-1})(\tau_k - \tilde{\tau}_k))^{1/2}}\right).$$

Move 3:

$$A = \frac{h(y|\tilde{m},\tilde{\tau};\theta)}{h(y|m,\tau;\theta)} \times \frac{1-\lambda}{\lambda} \times (2\pi)^{1/2}\left(\frac{(\tau_{k+1} - \tau_{k-1})}{(\tau_{k+1} - \tau_k)(\tau_k - \tau_{k-1})}\right)^{1/2} \times \exp\left\{-\frac{1}{2V}(\tilde{m}_k^2(\tau_{k+1} - \tau_{k-1}) - m_{k+1}^2(\tau_{k+1} - \tau_k) - m_k^2(\tau_k - \tau_{k-1}))\right\} \times \frac{k}{0.4(n-1)}\left(\frac{((\tau_{k+1} - \tau_k)(\tau_k - \tau_{k-1}))^{1/2}}{\tau_{k+1} - \tau_{k-1}}\right).$$

Move 4:

$$A = \frac{h(y|\tilde{m},\tilde{\tau};\theta)}{h(y|m,\tau;\theta)} \times \left(\frac{(\tau_{k+1} - \tilde{\tau}_k)(\tilde{\tau}_k - \tau_{k-1})}{(\tau_{k+1} - \tau_k)(\tau_k - \tau_{k-1})}\right)^{1/2} \times \exp\left\{-\frac{1}{2V}(m_k^2 - m_{k+1}^2)(\tilde{\tau}_k - \tau_k)\right\}.$$

References

[1] M. Basseville, N. Nikiforov, The Detection of Abrupt Changes – Theory and Applications, Information and System Sciences Series, Prentice-Hall, Englewood Cliffs, NJ, 1993.

[2] J. Besag, P.J. Green, D. Higdon, K. Mengersen, Bayesian computation and stochastic systems, Stat. Sci. 10 (1995) 3–66.

[3] R. Biscay, M. Lavielle, A. González, I. Clark, P. Valdés, Maximum a posteriori estimation of change points in the EEG, Int. J. Bio-Med. Comput. 38 (1995) 189–196.

[4] B.E. Brodsky, B.S. Darkhovsky, Nonparametric Methods in Change-Point Problems, Kluwer Academic Publishers, the Netherlands, 1993.

[5] O. Cappé, A. Doucet, M. Lavielle, E. Moulines, Methods for blind maximum-likelihood linear system identification, Signal Processing 73 (1999) 3–25.

[6] G. Casella, C.P. Robert, Rao–Blackwellisation of sampling schemes, Biometrika 83 (1) (1996) 81–84.

[7] M. Csörgö, L. Horváth, Limit Theorems in Change-Point Analysis, Wiley, UK, 1997.

[8] B. Delyon, M. Lavielle, E. Moulines, Convergence of a stochastic approximation version of the EM algorithm, Ann. Stat. 27 (1) (1999) 94–128.

[9] A. Dempster, N. Laird, D. Rubin, Maximum-likelihood from incomplete data via the EM algorithm, J. Roy. Stat. Soc. B 39 (1977) 1–38.

[10] M. Duflo, Algorithmes Stochastiques, SMAI, Springer, Berlin, 1996.

[11] S. Geman, D. Geman, Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images, IEEE Trans. Pattern Anal. Mach. Intell. 6 (1984) 721–741.

[12] P.J. Green, Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, Biometrika 82 (4) (1995) 711–732.

[13] M. Lavielle, A stochastic procedure for parametric and non-parametric estimation in the case of incomplete data, Signal Processing 42 (1995) 3–17.

[14] M. Lavielle, Optimal segmentation of random processes, IEEE Trans. Signal Process. 46 (5) (1998) 1365–1373.
[15] M. Lavielle, E. Moulines, Least squares estimation of an unknown number of shifts in a time series, J. Time Ser. Anal. 21 (1) (2000) 33–59.
[16] M. Metivier, P. Priouret, Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissants, Probab. Theory Related Fields 74 (1987) 403–428.
[17] S.P. Meyn, R.L. Tweedie, Markov Chains and Stochastic Stability, Springer, New York, 1993.
[18] C.R. Rao, in: Linear Statistical Inference and its Applications, Series in Probability and Mathematical Statistics, Wiley, New York, 1965.
[19] C.P. Robert, Méthodes de Monte Carlo par Chaînes de Markov, Statistique mathématique et Probabilité, Economica, 1996.
[20] C.P. Robert, in: Discretization and MCMC Convergence Assessment, Lecture Notes in Statistics, vol. 135, Springer, Berlin, 1998.
[21] J.Y. Tourneret, M. Chabert, Off-line detection and estimation of abrupt changes corrupted by multiplicative colored Gaussian noise, Proceedings of ICASSP'97, Munich, April 1997.
[22] J.Y. Tourneret, M. Coulon, M. Doisy, Least Squares estimation of multiple abrupt changes contamined by multiplicative noise using MCMC, Proceedings of HOS'99, Caesarea, June 1999.