

# Optimal Segmentation of Random Processes

Marc Lavielle

**Abstract**—Segmentation of a nonstationary process consists in assuming piecewise stationarity and in detecting the instants of change. We consider here that all the data is available in a same time and perform a global segmentation instead of a sequential procedure. We build a change process and define arbitrarily its prior distribution. That allows us to propose the MAP estimate as well as some minimum contrast estimate as a solution. One of the interests of the method is its ability to give the best solution, according to the resolution level required by the user, that is, to the prior distribution chosen. The method can address a wide class of parametric and nonparametric models. Simulations and applications to real data are proposed.

**Index Terms**—Detection of changes, MAP estimate, minimum contrast estimate, parametric and nonparametric distributions, segmentation.

## I. INTRODUCTION

LET  $X = (X_i, i \geq 1)$  be a nonstationary real process. We assume that  $X$  is *piecewise stationary*. Then, there exist instants  $(t_k, k \geq 0)$  such that  $(X_{t_k+1}, \dots, X_{t_{k+1}})$  is stationary for all  $k \in \mathbb{N}$ . The problem consists of detecting the changes in the distribution of  $X$ , that is, in recovering the family  $(t_k)$  when a trajectory of  $X$  is observed.

First, we shall assume that the distribution of the process  $X$  depends on a parameter  $\theta$ . Thus, the problem consists now of detecting the changes of  $\theta$ . The changes can affect the mean and the covariance structure of the process, the transition probabilities in a Markov random chain, etc.

When the detection delay  $\tau$  (which is the time between the change and its detection) needs to be well controlled, a sequential detection is performed, which means to decide at time  $t + \tau$  if a change has occurred at time  $t$ . Most of the test statistics used by the detection algorithms are built from the likelihood ratio or the Kullback distance [1]–[4]. The goal of these procedures is to minimize the probabilities of false alarms and omissions as well as the delay  $\tau$ . Here, we assume that all the data is available. The detection is off line. Thus, the criteria of good recovery are only related to the detection errors. Instead of a sequential procedure that does not use the information provided by the future, we shall perform a global segmentation of the process by detecting all the changes at the same time.

To do this, we shall introduce a new random process  $R$  that takes the value 1 at the change instants and is zero between

two changes. Detecting the instants of changes consists of recovering the change process  $R$ . Without any additional prior information, this change process is defined as a sequence of independent Bernoulli variables. When a particular realization  $\underline{x}$  of the vector  $\underline{X} = (X_1, \dots, X_n)$  is observed, the maximum *a posteriori* (MAP) estimate of the vector  $\underline{R} = (R_1, \dots, R_n)$  is computed by maximizing the *a posteriori* distribution  $\Pr(\underline{R} = \underline{r} / \underline{X} = \underline{x})$ .

More generally, the instants of change are estimated by minimizing a penalized contrast function of the form

$$U_{\underline{x}}(\underline{r}) = V_{\underline{x}}(\underline{r}) + \beta \sum_{i=1}^n r_i \quad (1)$$

where the first term  $V_{\underline{x}}(\underline{r})$  measures the fidelity to the observation  $\underline{x}$ , whereas the second term is related to the number of changes. Equation (1) defines the MAP estimate when  $-V_{\underline{x}}(\underline{r})$  is the log-likelihood of  $\underline{x}$  for a sequence of changes  $\underline{r}$ .

The parameter  $\beta$  is a tradeoff between these two criteria and controls the probabilities of detection errors. Small changes in the distribution of the process are detected with a small value of  $\beta$ . A bigger value of  $\beta$  allows us to detect only the more important changes.

When the contrast function (or energy function)  $U_{\underline{x}}$  to be minimized can be decomposed into a sum of local potentials, a simulated annealing procedure can be used to reach the optimal configuration  $\hat{\underline{r}}$  [5].

Numerical experiments with parametric models are proposed in Section III. First, the algorithm is used for detecting jumps in the mean of independent Gaussian variables. We see with this example that the resolution level of the segmentation directly depends on the prior distribution of  $R$ , that is, on the value of the parameter  $\beta$ . A method is proposed for choosing this parameter  $\beta$ . In a second example, the changes affect both the mean and the variance, and we show with a simulation that the original changes can be well recovered. Finally, the algorithm is used for detecting changes in the parameters of an AR process. An application to an electroencephalogram (EEG) is presented.

We show in Section IV that this algorithm can also be used for detecting changes in a nonparametric distribution. We consider a sequence of random variables that marginal distribution is piecewise constant. A new sequence of random variables with a parametric distribution is built from the empirical distribution of the original variables, and the changes are now detected in this sequence.

A simulation shows the very good behavior of the method. An application to real data is proposed as well; the heart rate of a newborn baby is segmented for identifying heavy and light

Manuscript received February 8, 1994; revised July 14, 1997. The associate editor coordinating the review of this paper and approving it for publication was Prof. Douglas Williams.

The author is with the Université René Descartes, Paris, France and the Université Paris-Sud, Orsay, France (e-mail: lavielle@math-info.univ-paris5.fr).

Publisher Item Identifier S 1053-587X(98)03255-3.

sleep periods. In this example, the true instants of change are known and well recovered by the algorithm.

In a same way, changes in the spectrum of a nonparametric process are detected. The new statistic that is used for the segmentation is built up from the empirical spectral distribution. Then, we look for changes in the mean of this statistic. That allows us to detect changes in some given bands of frequency. Such a method is shown to be very useful in some applications such as EEG analysis.

The optimization procedures are presented in Section V. The global minimum of the contrast function is computed with a simulated annealing procedure. The iterative conditional mode (ICM) algorithm is much faster and leads to a local minimum.

## II. THE MODEL

Let  $X = (X_i, i \geq 1)$  be a nonstationary  $d$ -dimensional real process. To assume that  $X$  is piecewise stationary means that there exist instants  $(t_k, k \geq 0)$  such that  $(X_{t_k+1}, \dots, X_{t_{k+1}})$  is stationary for all  $k \in \mathbb{N}$ . Here, we set  $t_0 = 0$ .

We consider here that a complete trajectory  $\underline{x} = (x_1 \dots x_n)$  is observed, and we shall perform a global detection of changes by using all the trajectory at the same time.

The problem of detecting the changes can also be seen as a global segmentation of  $\underline{X} = (X_1 \dots X_n)$  in stationary pieces. In fact, it is equivalent to looking for the optimal family of instants of change  $(t_k, k \geq 0)$  or for the optimal segmentation of the process.

We say that the solution is optimal according to some criteria of good recovery. For a global detection, the criteria of good recovery are only related to the detection errors and define a resolution level for the segmentation.

We shall introduce a random vector  $\underline{R} = (R_1, \dots, R_n)$  that is defined by

$$R_i = \begin{cases} 1, & \text{if there exists } k \text{ such that } i = t_k \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Then,  $\underline{R}$  takes the value 1 at the change instants and is zero between two changes. Of course, detecting the instants of changes is equivalent to recovering the vector  $\underline{R}$ .

Conventionally, we shall set  $R_n = 1$  such that the number of changes  $S_r = \sum_{i=1}^n R_i$  is the same as the number of segments. Let  $\pi(\underline{r}) = \Pr(\underline{R} = \underline{r})$  be the prior probability to have the configuration  $\underline{r}$ . With no additional information, we could think of defining  $\underline{R}$  as a sequence of independent Bernoulli variables

$$\pi(\underline{r}) = \lambda^{S_r} (1 - \lambda)^{n - S_r} \quad (3)$$

where  $\lambda$  is a real parameter between 0 and 1.

When a particular realization  $\underline{x}$  of the process  $X$  is observed, it is natural to look for the most likely configuration of changes for this trajectory  $\underline{x}$ . In other words, we shall look for the most likely value of  $\underline{R}$ , according to the observations and a prior information. The MAP estimate is obtained by maximizing the conditional probability  $\Pr(\underline{R} = \underline{r} / \underline{X} = \underline{x})$ .

In the case of a parametric model, the distribution of the process  $X$  depends on a parameter whose value remains constant in each stationary piece. For a given configuration

$\underline{r}$  of  $K$  segments, let  $\underline{\theta} = (\theta_1, \dots, \theta_K)$  be the sequence of parameters such that  $\theta_k$  is the parameter in the  $k$ th segment. Define  $h(\cdot / \underline{r}; \underline{\theta})$  as the density function of the distribution of  $\underline{X}$  conditionally to  $\underline{R} = \underline{r}$ .

Then,  $\underline{\theta}$  can be estimated simultaneously with  $\underline{R}$  by maximizing the *a posteriori* distribution of  $\underline{R}$

$$[\hat{\underline{r}}, \hat{\underline{\theta}}(\hat{\underline{r}})] = \arg \max_{(\underline{r}, \underline{\theta})} \Pr(\underline{R} = \underline{r} / \underline{X} = \underline{x}; \underline{\theta}) \quad (4)$$

$$= \arg \max_{(\underline{r}, \underline{\theta})} h(\underline{x} / \underline{r}; \underline{\theta}) \pi(\underline{r}) \quad (5)$$

and  $\hat{\underline{r}}$  is obtained as

$$\hat{\underline{r}} = \arg \max_{\underline{r} \in \{0; 1\}^n} h[\underline{x} / \underline{r}; \hat{\underline{\theta}}(\underline{r})] \pi(\underline{r}). \quad (6)$$

*Remark:* In a Bayesian framework,  $\hat{\underline{\theta}}(\underline{r})$  is a Bayesian estimate of  $\underline{\theta}$ . Let  $f$  be the density of the prior distribution of  $\underline{\theta}$ . Then

$$[\hat{\underline{r}}, \hat{\underline{\theta}}(\hat{\underline{r}})] = \arg \max_{(\underline{r}, \underline{\theta})} h(\underline{x} / \underline{r}; \underline{\theta}) \pi(\underline{r}) f(\underline{\theta}). \quad (7)$$

In most of the examples proposed below, the maximum likelihood estimate of  $\underline{\theta}$  is used. For a given configuration of changes  $\underline{r}$ , the maximum likelihood estimate of  $\theta_k$  is computed in segment  $k$  as

$$\hat{\theta}_k(\underline{r}) = \arg \max_{\theta \in \Theta} l(x_{t_{k-1}+1}, \dots, x_{t_k}; \theta) \quad (8)$$

where  $l$  is the log-likelihood of  $(x_{t_{k-1}+1}, \dots, x_{t_k})$ . Now, let

$$l_{\underline{x}}(\underline{r}, \underline{\theta}) = \sum_{k=1}^{S_r} l(x_{t_{k-1}+1}, \dots, x_{t_k}; \theta_k). \quad (9)$$

Assuming that the different segments are independent, we have that  $\hat{\underline{\theta}}(\underline{r}) = [\hat{\theta}_k(\underline{r})]$  and that  $\hat{\underline{r}}$  is obtained as

$$\hat{\underline{r}} = \arg \min_{\underline{r} \in \{0; 1\}^n} \left\{ -l_{\underline{x}}[\underline{r}; \hat{\underline{\theta}}(\underline{r})] + \alpha \sum_{i=1}^n r_i \right\} \quad (10)$$

where  $\alpha = \log(1/\lambda - 1)$ .

More generally, any contrast function  $V$  can be used for estimating  $\theta_k$  instead of maximizing the log-likelihood function  $l$

$$\hat{\theta}_k(\underline{r}) = \arg \min_{\theta \in \Theta} V(x_{t_{k-1}+1}, \dots, x_{t_k}; \theta). \quad (11)$$

Thus,  $\underline{r}$  is estimated by minimizing the penalized contrast function  $U_{\underline{x}}$  defined by

$$U_{\underline{x}}(\underline{r}) = V_{\underline{x}}[\underline{r}; \hat{\underline{\theta}}(\underline{r})] + \beta \sum_{i=1}^n r_i \quad (12)$$

where

$$V_{\underline{x}}(\underline{r}, \underline{\theta}) = \sum_{k=1}^{S_r} V(x_{t_{k-1}+1}, \dots, x_{t_k}; \theta_k) \quad (13)$$

and where  $\beta$  is a tuning parameter that must be fixed to a positive value.

The first term  $V_{\underline{x}}[\underline{r}; \hat{\underline{\theta}}(\underline{r})]$  is related to the fit to the observation  $\underline{x}$ , whereas the second term is related to the number of

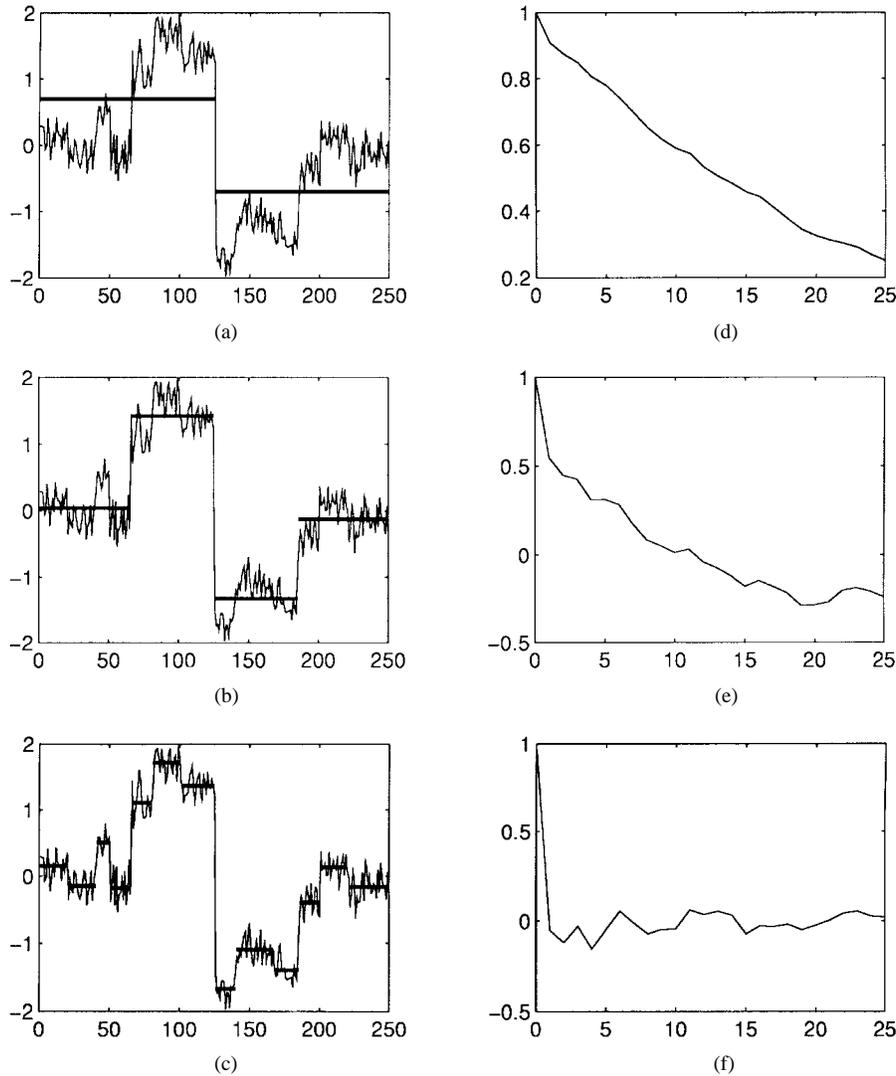


Fig. 1. Detection of changes in the mean of a Gaussian process obtained with (a)  $\beta = 10$ , (b)  $\beta = 5$ , and (c)  $\beta = 1$ . The autocorrelations of the residuals are displayed in (d)–(f).

changes. The parameter  $\beta$  controls the probabilities of detection errors; the smaller  $\beta$  is, the bigger the prior probability of a change, and the fewer the omissions are. On the other hand, the bigger  $\beta$  is, the fewer the false alarms. Thus,  $\beta$  is a parameter that controls the resolution level of the segmentation; small changes will be detected for small values of  $\beta$ .

### III. EXAMPLES WITH PARAMETRIC MODELS

#### A. Changes in the Mean of a Process

We consider the following process:

$$X_i = \mu_k + \varepsilon_i, \quad t_{k-1} + 1 \leq i \leq t_k \quad (14)$$

where  $\varepsilon$  is an additive noise. If  $\varepsilon$  is a Gaussian white noise with variance  $\sigma_\varepsilon^2$ , it is easy to show that  $\hat{\mu}$  is computed by minimizing

$$U_{\underline{x}}(\underline{\mu}) = \sum_{k=1}^{S_r} \sum_{i=t_{k-1}+1}^{t_k} [X_i - \hat{\mu}_k(\underline{x})]^2 + \beta S_r \quad (15)$$

where  $\beta = 2\alpha\sigma_\varepsilon^2$ . Here,  $\hat{\mu}_k(\underline{x})$  is the least-square estimate of  $\mu_k$  on the  $k$ th segment of configuration  $\underline{x}$

$$\hat{\mu}_k(\underline{x}) = \frac{1}{t_k - t_{k-1}} \sum_{i=t_{k-1}+1}^{t_k} X_i. \quad (16)$$

When  $\varepsilon$  is not a Gaussian white noise, it is important to remark that this function  $U_{\underline{x}}$  can be used as a contrast function. We just have to choose the value of  $\beta$ .

We simply consider the case  $\beta > 0$  since the method does not have any sense for a prior probability  $\lambda \geq 0.5$ . In this case,  $\beta$  is negative, and the optimal segmentation is obtained for  $\hat{\mu} = (1, 1 \dots 1)$ , that is, by detecting a change at each instant.

We propose in Fig. 1 different segmentations of the same process, obtained with different positive values of  $\beta$ . It is clear in this example that the parameter  $\beta$  defines the resolution level. If we want details, that is, to detect small changes of  $\mu$ , we must choose a small value for  $\beta$ . On the other hand, only the more important jumps of the mean are detected with a bigger value of  $\beta$ .

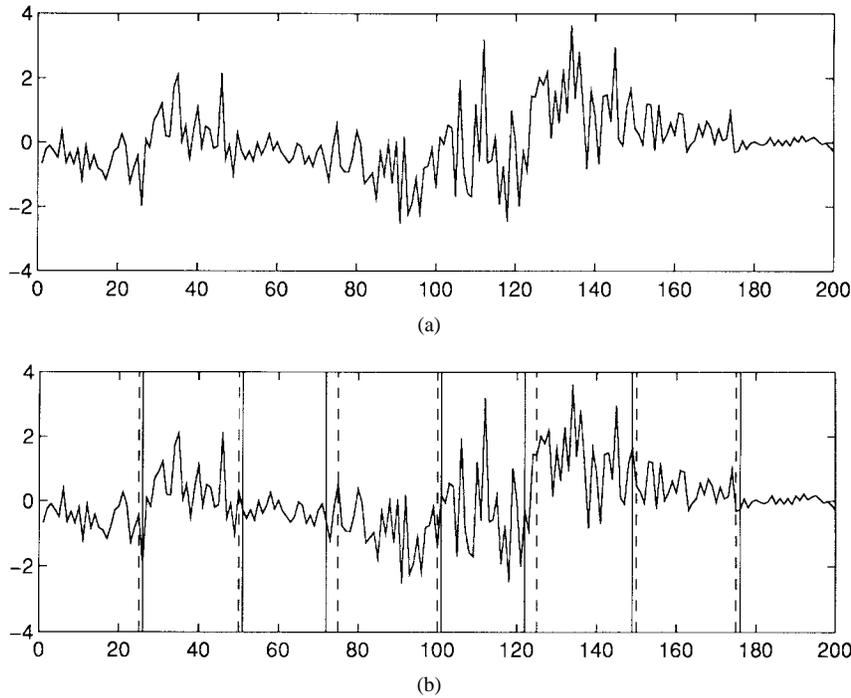


Fig. 2. Detection of changes in the mean and the variance of a Gaussian process. (a) Observed series. (b) Segmentation. The original change points are the dotted line, and the estimated change points are the solid line.

A main problem consists of the choice of  $\beta$ . In the model described above, the noise is assumed to be noncorrelated. Thus, a natural choice for  $\beta$  would be the greatest value for which the residuals are considered to be noncorrelated. We also display in Fig. 1 the estimated autocorrelation function of the residuals obtained with the three segmentations proposed in this example. We see that the hypothesis of noncorrelated residuals is accepted with  $\beta = 1$  [Fig. 2(f)] but rejected with greater values [Fig. 2(d) and (e)].

In many applications, the whiteness of the noise is a strong hypothesis that does not have any physical sense. Indeed, the resolution level will be selected empirically, depending on the relevance of some details. The eye of the specialist is generally the best way to select the best value of  $\beta$ . The only job of the algorithm is to compute the optimal segmentation according to the resolution level chosen by the user.

### B. Changes in the Mean and the Variance of a Process

We shall assume now that the changes affect also the variance  $\sigma_\varepsilon^2$  of the noise. That means that we have to detect changes in both the mean and the variance of  $X$ . Using the log-likelihood of a Gaussian process  $X$ , we must minimize the contrast function

$$U_{\underline{x}}(\underline{r}) = \sum_{k=1}^{S_r} n_k(\underline{r}) \log \hat{\sigma}_k^2(\underline{r}) + \beta S_r \quad (17)$$

where  $\beta = 2\alpha$ . Here,  $\hat{\sigma}_k^2(\underline{r})$  is the estimated variance of  $X$  on the  $k$ th segment, and  $n_k(\underline{r})$  is its length

$$\hat{\sigma}_k^2(\underline{r}) = \frac{1}{n_k(\underline{r})} \sum_{i=t_{k-1}+1}^{t_k} [X_i - \hat{\mu}_k(\underline{r})]^2 \quad (18)$$

$$n_k(\underline{r}) = t_k - t_{k-1}. \quad (19)$$

A Gaussian process  $X$  was simulated with changes in the parameters [Fig. 2(a)]. In this example, the original changes are perfectly detected for  $8 \leq \beta \leq 14$  [Fig. 2(b)].

### C. Changes in the Spectrum of an AR Process

We consider the  $AR(p)$  process

$$X_i = \sum_{j=1}^p a_{kj} X_{i-j} + \varepsilon_i, \quad t_{k-1} + p + 1 \leq i \leq t_k \quad (20)$$

where  $\varepsilon$  is the innovation process.

We can use the sum of the residual squares to estimate the coefficients  $(a_{kj})$ . In other words,  $[\hat{a}_{k1}(\underline{r}), \dots, \hat{a}_{kp}(\underline{r})]$  are the empirical estimates of the coefficients  $(a_{k1}, \dots, a_{kp})$  on the  $k$ th segment of configuration  $\underline{r}$  obtained by minimizing the contrast function  $V$  defined by

$$\begin{aligned} V(x_{t_{k-1}+1}, \dots, x_{t_k}; a_{k1}, \dots, a_{kp}) \\ = \sum_{i=t_{k-1}+p+1}^{t_k} \left( X_i - \sum_{j=1}^p a_{kj} X_{i-j} \right)^2. \end{aligned} \quad (21)$$

Then,  $\hat{\underline{r}}$  is computed by minimizing

$$U_{\underline{x}}(\underline{r}) = \sum_{k=1}^{S_r} \sum_{i=t_{k-1}+p+1}^{t_k} \left( X_i - \sum_{j=1}^p \hat{a}_{kj}(\underline{r}) X_{i-j} \right)^2 + \beta S_r. \quad (22)$$

We present here an application to EEG analysis. The most popular methods use an  $AR$  model to describe locally the EEG [6]–[8]. In fact, the basic assumption that underlies this type of processing is that of piecewise second-order stationarity. In

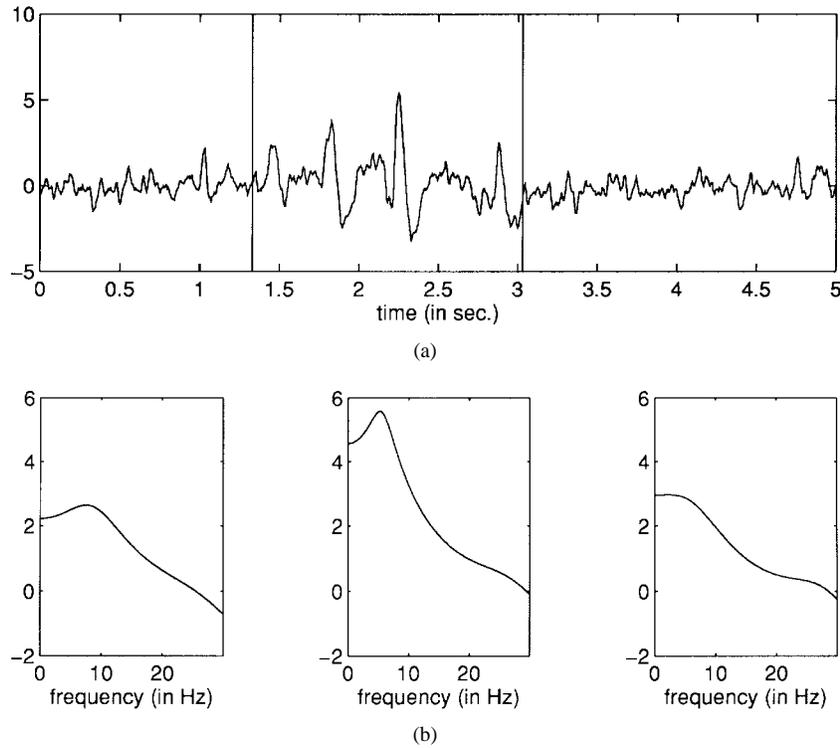


Fig. 3. Detection of changes in the spectrum of a EEG with an  $AR(7)$  modeling. (a) Observed series and the estimated change points. (b) Three estimated log-spectra, corresponding to the three segments.

other words, a spectral analysis of the EEG allows us to detect changes in the electrical activity of the brain.

An  $AR(7)$  was used to detect changes in the EEG displayed in Fig. 3(a). The method successfully detected epileptic spikes (1.3–3.0 s) in the background activity. The estimated coefficients were used to compute the estimated spectrum in each segment [Fig. 3(b)]. In this case, the paroxysical activity is clearly characterized as an activity in the range of frequencies 3.5–7.5 Hz.

#### IV. CHANGES IN A NONPARAMETRIC DISTRIBUTION

##### A. Changes in the Marginal Distribution

We consider now a process  $X$  such that the marginal distribution of the  $X_i$ 's is piecewise constant. The instants  $(t_k, k \geq 0)$  are such that  $X_{t_{k-1}+1}, \dots, X_{t_k}$  have the same marginal distribution for all  $k \in \mathbb{N}$ . Assuming that this distribution possesses a density with respect to a given measure, we want to detect changes in the density function.

Assuming that  $X$  is a sequence of independent random variables, we shall build a new statistic from the empirical distribution of  $X$ .

Let  $(z_m, 0 \leq m \leq M)$  be a sequence of real numbers such that  $z_0 < z_1 < \dots < z_M$ . For each  $X_i$ , we define a new variable  $Y_i$  that takes the value  $m$  when  $z_{m-1} < X_i \leq z_m$ .

The distribution of  $Y$  can be seen as the projected distribution of  $X$ . Using the fact that the changes that affect the distribution of  $X$  also affect the projected distribution, we shall recover  $\underline{R}$  by maximizing the posterior distribution  $\Pr(\underline{R} = \underline{r} / \underline{Y} = \underline{y})$ .

Let  $h_k$  be the probability density function of  $X$  in the  $k$ th segment. For any  $t_{k-1} + 1 \leq i \leq t_k$ , let

$$p_{km} = \Pr(Y_i = m) = \int_{z_{m-1}}^{z_m} h_k(x) dx. \quad (23)$$

Thus

$$\Pr(\underline{Y} = \underline{y}) = \prod_{k=1}^{S_r} \prod_{m=1}^M p_{km}^{n_{km}(\underline{r})} \quad (24)$$

where  $n_{km}(\underline{r})$  is the number of times that  $Y$  takes the value  $m$  in the  $k$ th segment of configuration  $\underline{r}$ .

Since  $p_{km}$  is estimated by

$$\hat{p}_{km}(\underline{r}) = \frac{n_{km}(\underline{r})}{n_k(\underline{r})} \quad (25)$$

where  $n_k(\underline{r}) = \sum_{m=1}^M n_{km}(\underline{r})$  is the length of the  $k$ th segment, the solution  $\hat{\underline{r}}$  is obtained by minimizing

$$U_{\underline{r}}(\underline{r}) = - \sum_{k=1}^{S_r} \sum_{m=1}^M n_{km}(\underline{r}) \log \frac{n_{km}(\underline{r})}{n_k(\underline{r})} + \beta S_r \quad (26)$$

where  $\beta = \alpha$ .

A simulation is shown in Fig. 4. Independent Gaussian variables were simulated in the first and third segments, whereas a uniform distribution was used in the second segment [Fig. 4(a)]. In these examples, the changes were well detected by the algorithm for a number of classes  $M = 20$  and  $8 \leq \beta \leq 12$  [Fig. 4(b)]. The changes are not significant enough to be detected with a value of  $\beta$  greater than 12, whereas a value smaller than 8 produces false alarms.

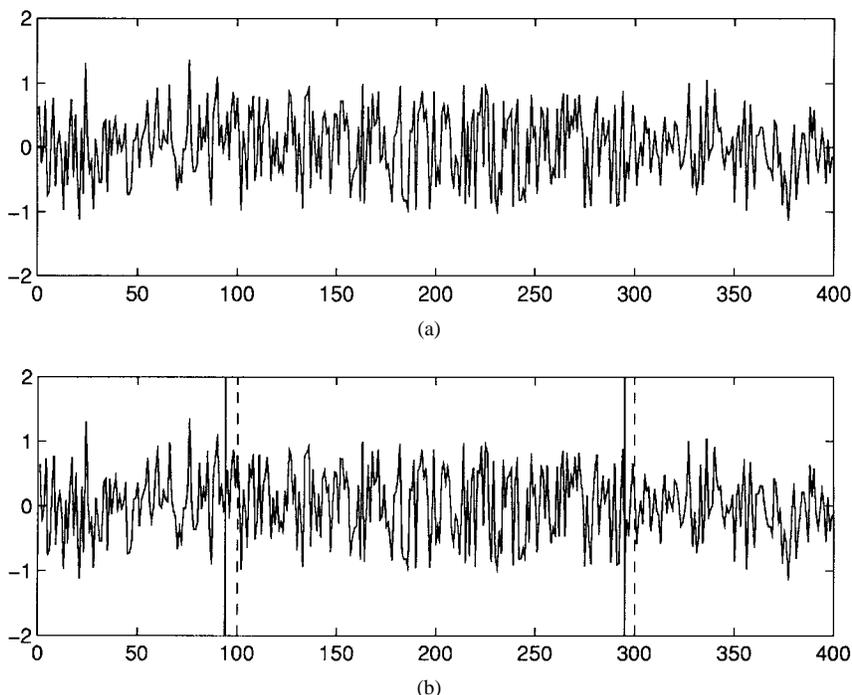


Fig. 4. Detection of changes in the marginal distribution of a sequence of independent random variables. (a) Observed series. The variables have a Gaussian distribution between 0 and 100 and between 300 and 400, and the distribution is uniform between 100 and 300. (b) Segmentation. The original change-points are shown as a dotted line, and the estimated change points are shown as a solid line.

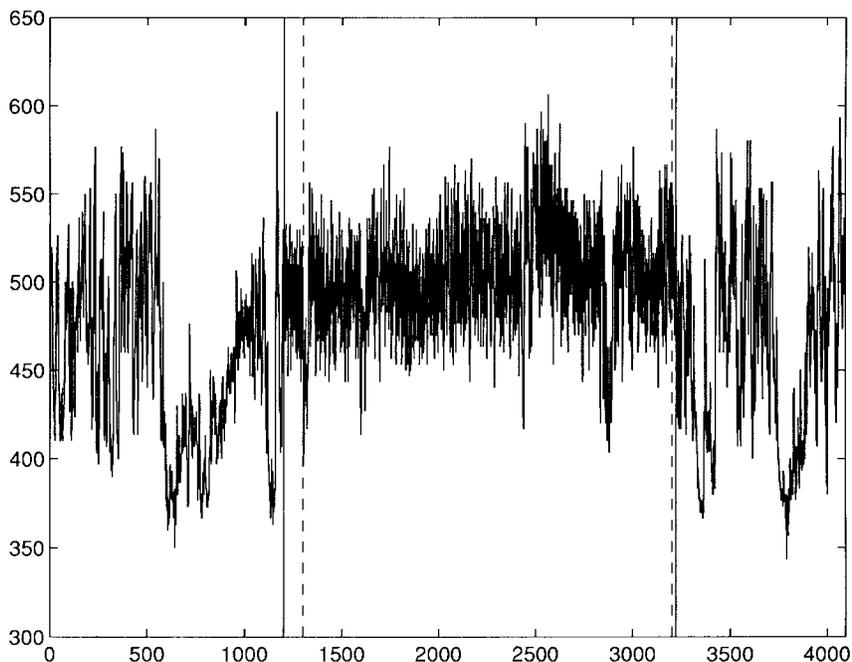


Fig. 5. Detection of changes in the heart rate of a newborn baby. The change points estimated with external measurements are shown as a dotted line, and the change points estimated with the algorithm are shown as a solid line.

Another application of the method with real data is proposed. Fig. 5(a) represents the heart rate of a newborn baby. It can be very useful to identify automatically heavy and light sleep periods from this series. In this example, external measurements (such as that of the movement of the eye-lids) let us know that the heavy sleep period is approximately between data 1300 and data 3200. We can see in Fig. 5(b) that

the changes detected by the algorithm with  $200 \leq \beta \leq 400$  agree with the exact instants of change.

#### B. Changes in the Spectrum

In [4], Lavielle proposes a sequential procedure for detecting changes in the spectrum of a multidimensional process, assuming any kind of nonparameterized distribution. We shall

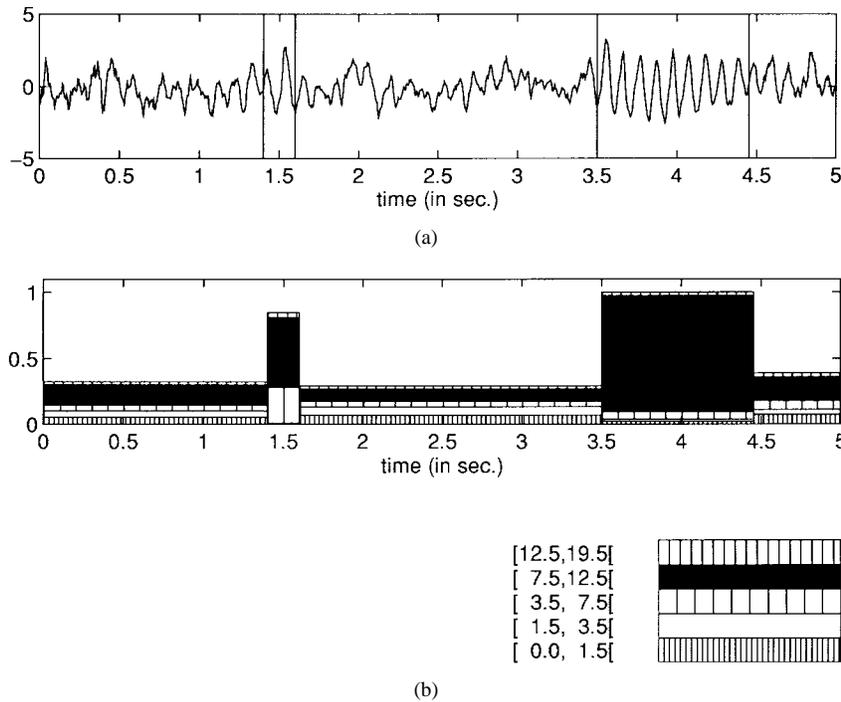


Fig. 6. Detection of changes in the spectrum of a EEG. (a) Observed series and the estimated change-points. (b) Estimated spectral distributions corresponding to the different segments. Here, five well-known bands of frequency were used for the segmentation.

see in this section how to extend this method for a global segmentation of the process.

In the previous section, we used the empirical marginal distribution for detecting changes in the marginal distribution of the process in given classes  $[z_{m-1}, z_m[$ . In a similar way, we shall use now the empirical spectral distribution for detecting changes in the spectrum of the process in given bands of frequency  $[\lambda_{m-1}, \lambda_m[$ .

We shall define a new statistic  $\underline{Z}(\underline{r}) = [Z_{km}(\underline{r}), 1 \leq k \leq S_r, 1 \leq m \leq M]$ , where

$$Z_{km}(\underline{r}) = \int_{\lambda_{m-1}}^{\lambda_m} I_k(\underline{r}, \lambda) d\lambda \quad (27)$$

where  $I_k(\underline{r}, \lambda)$  is the periodogram computed on the  $k$ th segment of the configuration  $\underline{r}$

$$I_k(\underline{r}, \lambda) = \frac{1}{n_k(\underline{r})} \left| \sum_{t=t_{k-1}+1}^{t_k} X_t e^{i\lambda t} \right|^2 \quad (28)$$

and

$$Z_{km}(\underline{r}) = \frac{1}{n_k(\underline{r})} \left[ (\lambda_m - \lambda_{m-1}) \sum_{t=t_{k-1}+1}^{t_k} X_t^2 + 2 \sum_{t=t_{k-1}+1}^{t_k-1} \sum_{s=1}^{t_k-t} X_t X_{t+s} \cdot \left( \frac{\sin \lambda_m s - \sin \lambda_{m-1} s}{s} \right) \right]. \quad (29)$$

The  $Z_{km}(\underline{r})$  are asymptotically Gaussian variables, independent in both time and frequency domains, that converge to the spectral distribution of  $X$  [4], [9], [10]. Indeed, let  $f_k$  be

the spectral density of  $X$  in the  $k$ th stationary segment, and let  $\underline{r}^*$  be the true configuration of changes. Let  $n_k^* = n_k(\underline{r}^*)$  and  $Z_{km}^* = Z_{km}(\underline{r}^*)$ . Then, if  $n_k^* \rightarrow \infty$  when  $n \rightarrow \infty$ , it can be shown that

$$\sqrt{n_k^*} \left[ Z_{km}^* - \int_{\lambda_{m-1}}^{\lambda_m} f_k(\lambda) d\lambda \right] \xrightarrow{n \rightarrow \infty} \mathcal{N} \left[ 0, \int_{\lambda_{m-1}}^{\lambda_m} f_k^2(\lambda) d\lambda \right] \quad (30)$$

and

$$E \left[ \sqrt{n_k^* n_j^*} (Z_{km}^* - E Z_{km}^*) (Z_{jl}^* - E Z_{jl}^*) \right] \xrightarrow{n \rightarrow \infty} 0 \quad \text{if } (k, m) \neq (j, l). \quad (31)$$

We look for changes in the spectral distribution of  $X$ , that is, in  $[\int_{\lambda_{m-1}}^{\lambda_m} f(\lambda), 1 \leq m \leq M]$ . To do that, using (30), we detect changes in the mean of  $Z$ .

If we want to detect changes in the mean of a process, we can remark that it is equivalent to estimating  $\underline{r}$  by minimizing the contrast function defined in (15) or by minimizing the function  $J_{\underline{x}}$  defined by

$$J_{\underline{x}}(\underline{r}) = - \sum_{k=1}^{S_r} n_k(\underline{r}) \hat{\mu}_k(\underline{r})^2 + \beta S_r. \quad (32)$$

Thus, for detecting changes in the spectrum of the process  $X$ , we shall estimate  $\underline{r}$  by minimizing the function  $U_{\underline{x}}$  defined by

$$U_{\underline{x}}(\underline{r}) = - \sum_{k=1}^{S_r} \sum_{m=1}^M n_k(\underline{r}) Z_{km}(\underline{r})^2 + \beta S_r. \quad (33)$$

An application to EEG analysis is presented in Fig. 6. Five well known bands of frequencies (in hertz) were used:  $[0; 1.5[$ ,  $[1.5; 3.5[$ ,  $[3.5; 7.5[$ ,  $[7.5; 12.5[$  and  $[12.5; 19.5[$ .

Each one of these bands of frequency correspond to a particular electrical activity of the brain [8]. In this example, a period of  $\alpha$  activity ([7.5, 12.5]) is well identified by the algorithm around 4 s and a shorter one around 1.5 s. The estimated spectral distribution for each segment is displayed in Fig. 6(b).

## V. THE OPTIMIZATION PROCEDURES

### A. The Simulated Annealing Algorithm

For a process of length  $n$ , a change can occur at the  $n-1$  first instants. Thus,  $\underline{R}$  takes its values in a space that contains  $2^{n-1}$  elements. Let  $\hat{\underline{r}}$  be the configuration that minimizes  $U_{\underline{x}}(\underline{r})$ . An exhaustive research of  $\hat{\underline{r}}$  by computing the  $2^{n-1}$  values of  $U_{\underline{x}}(\underline{r})$  is not generally tractable. Nevertheless, a simulated annealing procedure can be used to reach the solution  $\hat{\underline{r}}$ .

The simulated annealing algorithm is an iterative procedure that defines a nonhomogeneous Markov chain  $\{r(j), j \geq 0\}$  that converges to the optimal solution  $\hat{\underline{r}}$  with probability one; see [5].

- Choose an initial configuration  $\underline{r}(0)$ .
- At iteration  $j$ 
  - choose a new configuration  $\tilde{\underline{r}}$  as a modification of  $\underline{r}(j-1)$ ;
  - let  $\Delta U = U_{\underline{x}}(\tilde{\underline{r}}) - U_{\underline{x}}[\underline{r}(j-1)]$ ;
  - Set  $\underline{r}(j) = \tilde{\underline{r}}$  with probability one if  $\Delta U < 0$  and with probability  $\exp\{-\Delta U/T(j)\}$  elsewhere. [Here,  $T(j)$  is a decreasing sequence called temperature.]
- Stop when no more modifications are accepted.

When the total energy is a sum of local potentials, a local perturbation of the configuration  $\underline{r}(j)$  will affect few terms of this sum, and the energy variation  $\Delta U$  will be easy to compute. In our segmentation algorithm, the modifications consist of adding a new change, in removing one, and in translating one.

### B. The ICM Algorithm

The simulated annealing algorithm is very useful when we need to obtain the global minima of the energy function  $U$  with probability one. The main limitation is the computational effort required by the algorithm. In theory, the temperature  $T$  must decrease very slowly to ensure the convergence to  $\hat{\underline{r}}$ , and a big number of iterations is required; see [5].

Now, if we want to obtain much faster a “good solution” with a “high probability,” the iterative conditional mode (ICM) algorithm can be preferred; see [11], [12]. This algorithm is the deterministic version of the simulated annealing procedure, setting  $T(j) = 0$  for any  $j$ . Then, the ICM leads to a minima of the energy function since only the modifications that produce a decrease of this function are accepted. Nevertheless, most of the local minima can be avoided by introducing a wider family of modifications, that is, by filling the transition matrix of the Markov chain. We could think, for example, of adding, removing, or translating two changes at a same iteration instead of one.

### C. Example

Let's go back to the example proposed in Section III, where the changes affect the mean of the process. Let  $[t_k(j-1), k > 0]$  be the estimated instants of changes at iteration  $j-1$ .

If the local modification consists of adding a change at the instant  $t$ , then  $r_t(j-1) = 0$  and  $\tilde{r}_t = 1$ , while  $\tilde{r}_s = r_s(j-1)$  for  $s \neq t$ . We assume that  $t$  belongs to the  $k$ th segment. Let  $\hat{\mu}_1$  (resp.,  $\hat{\mu}_2$ ) be the empirical mean of  $X$  between  $t_{k-1}+1$  and  $t$  (resp.,  $t+1$  and  $t_k$ ). Let  $n_1 = t - t_{k-1}$  and  $n_2 = t_k - t$ . Then

$$\Delta U = \beta - \frac{n_1 n_2}{n_1 + n_2} (\hat{\mu}_2 - \hat{\mu}_1)^2. \quad (34)$$

If the modification consists of removing a change at time  $t_k$ , the energy variation is

$$\Delta U = \frac{n_1 n_2}{n_1 + n_2} (\hat{\mu}_2 - \hat{\mu}_1)^2 - \beta \quad (35)$$

where  $\hat{\mu}_1$  (resp.,  $\hat{\mu}_2$ ) is the empirical mean of  $X$  between  $t_{k-1}+1$  and  $t_k$  (resp.,  $t_k+1$  and  $t_{k+1}$ ) and  $n_1 = t_k - t_{k-1}$  and  $n_2 = t_{k+1} - t_k$ .

This provides a simple condition to decide if  $\hat{r}_t = 0$  or  $\hat{r}_t = 1$ , where  $\hat{\underline{r}}$  is a minimum (local or global) of  $U_{\underline{x}}$ . Indeed, by using (34) and (35), we set  $\hat{r}_t = 1$  only if  $n_1 n_2 (n_1 + n_2)^{-1} (\hat{\mu}_2 - \hat{\mu}_1)^2 > \beta$ . A change is present at  $t$  only if the empirical means before and after  $t$  are significantly different.

One interesting aspect of this algorithm is that the decision rule only depends on the empirical mean of the process (similar calculus can easily be done when the modification consists in moving a change).

We can do the same kind of calculus with the example proposed in Section IV-A, where the changes affect the marginal distribution of the process. Using the same notations as above, let  $n_{1m}$  (resp.,  $n_{2m}$ ) be the number of times that  $Y$  takes the value  $m$  between  $t_{k-1}+1$  and  $t$  (resp.,  $t+1$  and  $t_k$ ). Let  $n_3 = n_1 + n_2$  and  $n_{3m} = n_{1m} + n_{2m}$ . For any  $j \in \{1; 2; 3\}$ , let

$$l_j = \sum_{m=1}^M n_j(m) \log \frac{n_j(m)}{n_j}. \quad (36)$$

Then, in the case of adding a change at  $t$ , the energy variation is

$$\Delta U = \beta - l_1 - l_2 + l_3. \quad (37)$$

Let  $\underline{n}_1 = (n_{11}, \dots, n_{1M})$  (resp.,  $\underline{n}_2 = (n_{21}, \dots, n_{2M})$ ) be the histogram of  $X$ , that is, the empirical distribution of  $Y$  between  $t_{k-1}+1$  and  $t$  (resp.,  $t+1$  and  $t_k$ ). Then, it is easy to check that

$$d(\underline{n}_1; \underline{n}_2) = l_1 + l_2 - l_3 \quad (38)$$

is a distance between the two histograms, that is, between the two empirical distributions  $d(\underline{n}_1; \underline{n}_2) \geq 0$  and  $d(\underline{n}_1; \underline{n}_2) = 0$ , if and only if  $\underline{n}_1 = \underline{n}_2$ . Furthermore, from (37), we have that  $\hat{r}_t = 1$  only if  $d(\underline{n}_1; \underline{n}_2) > \beta$ ; we decide that there exists a change at time  $t$  if the empirical distributions before and after  $t$  are significantly different.

We must notice here that any other distance between empirical distributions could be used. Nevertheless, experiments have shown that the distance defined above from the likelihood gives better results than any  $L_p$ -norm ( $p = 1, 2, \infty$ ) between histograms.

## REFERENCES

- [1] M. Basseville and N. Nikiforov, *The Detection of Abrupt Changes—Theory and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
- [2] A. Benveniste, M. Metivier, and P. Priouret, *Adaptive Algorithms and Stochastic Approximations*. New York: Springer-Verlag, 1990.
- [3] D. Hinkley, "Inference about the change point in a sequence of random variables," *Biometrika*, vol. 57, pp. 1–17, 1970.
- [4] M. Lavielle, "Detection of changes in the spectrum of a multidimensional process," *IEEE Trans. Signal Processing*, vol. 41, pp. 742–749, 1993.
- [5] D. Geman, "Random fields and inverse problems in imaging," in *Lecture Notes in Mathematics*. New York: Springer-Verlag, 1990.
- [6] C. Bodenstein and H. Praetorius, "Feature extraction from the electroencephalogram by adaptative segmentation," *Proc. IEEE*, vol. 65, pp. 642–652, 1977.
- [7] R. Biscay, M. Lavielle, A. González, I. Clark, and P. Valdés, "Maximum a posteriori estimation of change points in the EEG," *Int. J. Biomed. Comput.*, vol. 38, pp. 189–196, 1995.
- [8] S. H. Lopes, F. H. da Silva, and A. Dijk, "Detection of nonstationarities in the EEG's using the autoregressive model. An application to EEG's of epileptics," in *CEAN Computerized EEG Analysis*, G. Dolce and H. Kunkel, Eds. Stuttgart, Germany: Gustav Fisher Verlag, 1974.
- [9] I. Ibragimov, "On estimation of the spectral function of a stationary Gaussian process," *Theory Prob. Appl.*, vol. 8, pp. 366–400, 1962.
- [10] D. Picard, "Testing and estimating change points in time series," *J. Appl. Prob.*, vol. 17, pp. 841–867, 1985.
- [11] J. Besag, "On the statistical analysis of dirty pictures," *J. R. Stat. Soc. B*, vol. 48, pp. 259–302, 1986.
- [12] M. Lavielle, "Bayesian deconvolution of Bernoulli–Gaussian processes," *Signal Process.*, vol. 33, pp. 67–79, 1993.



**Marc Lavielle** was born in Paris, France, in 1961. He received the Ph.D. degree in mathematics from the University of Paris-Sud, Orsay, France, in 1990. From 1987 to 1991, he was Assistant Professor at the Universidad Central de Venezuela, Caracas. He is currently Maître de Conférences at the University René Descartes, Paris. His research interests include statistical methods for signal processing, stochastic algorithms, and geophysics.