Using penalized contrasts for the change-point problem

Marc Lavielle

To cite this version:

Marc Lavielle. Using penalized contrasts for the change-point problem. RR-5339, INRIA. 2004, pp.18. <inria-00070662>

HAL Id: inria-00070662

https://hal.inria.fr/inria-00070662

Submitted on 19 May 2006

Using penalized contrasts for the change-point problem

Marc Lavielle

N° 5339

Octobre 2004

Thème COG

Rapport de recherche

N 0249-6399 ISRN INRIA/RR--5339--FR+ENG

Using penalized contrasts for the change-point problem

Marc Lavielle *

Thème COG — Systèmes cognitifs
Projet Select

Rapport de recherche n° 5339 — Octobre 2004 — 18 pages

Abstract: A methodology for model selection based on a penalized contrast is developed. This methodology is applied to the change-point problem, for estimating the number of change points and their location. We aim to complete previous asymptotic results by constructing algorithms that can be used in diverse practical situations. First, we propose an adaptive choice of the penalty function for automatically estimating the dimension of the model, that is, the number of change points. In a Bayesian framework, we define the posterior distribution of the change-point sequence as a function of the penalized contrast. MCMC procedures are available for sampling this posterior distribution. The parameters of this distribution are estimated with a stochastic version of EM algorithm (SAEM). An application to EEG analysis and some Monte-Carlo experiments illustrate these algorithms.

Key-words: Penalized contrast, Model selection, Change-point problem, SAEM algorithm.

This work was supported by ECOS Nord V00M03

* Université René Descartes and Université Paris-Sud, France. e-mail: Marc.Lavielle@math.u-psud.fr

Unité de recherche INRIA Futurs
Parc Club Orsay Université, ZAC des Vignes,

Utilisation de contrastes pénalisés pour le problème de détection de ruptures

Résumé : Nous proposons une méthode de sélection de modèles, basée sur l'utilisation de contrastes pénalisés. Cette méthodologie est appliquée au problème de détection de ruptures, afin d'estimer le nombre de ruptures et leurs positions. Notre objectif est de compléter des résultats théoriques obtenus précédemment par des procédures algorithmiques applicables dans différentes situations pratiques. Dans un premier temps, nous proposons de déterminer de façon adaptative la pénalisation afin d'estimer autaomatiquement la dimension du modèle, c'est-'a-dire le nombre de ruptures ici. Dans un cadre bayésien, on définit la distribution a posteriori des instants de rupture comme une fonction du contraste pénalisé. Une procédure MCMC permet alors d'échantilloner cette distribution. Les paramètres sont estimés en utilisant une version stochastique de l'algorithme EM (SAEM). Une application à l'analyse d'enregistrements EEG et des Monte-Carlo illustrent numériquement ces algorithmes.

Mots-clés : Contraste pénalisé, Sélection de modèle, Détection de ruptures, Algorithme SAEM

1 Introduction

Detection of abrupt changes in the characteristics of some physical system is one of the important practical problems arising in signal processing (speech processing, geophysics, EEG, EMG and ECG analysis, etc., see [1] and [4] for several examples of application).

In a probabilistic framework, we consider a sequence of random variables $Y_1, \ldots, Y_n$, that take values in $\mathbb{R}^p$. We assume that some characteristics of the Y_i's changes abruptly at some unknown instants $\tau_1^\star < \tau_2^\star < \ldots < \tau_{K^\star-1}^\star$. Here, $K^\star$ (resp. $K^\star - 1$) is the unknown number of segments (resp. change points). The changes can affect the marginal distribution of the Y_i's (the mean, the variance, or some quantiles for example), or the joint distribution of the sequence (the spectral distribution for example).

Among the previously proposed methods for detecting multiple changes, we mention sequential methods (see [1] and the references therein) and local methods (see [6]). We shall adopt here a global approach, where all the change points are simultaneously detected by minimizing a penalized contrast $J(\boldsymbol{\tau}, \boldsymbol{y}) + \beta \text{pen}(\boldsymbol{\tau})$ (see [3, 8, 10, 11]). Here, $J(\boldsymbol{\tau}, \boldsymbol{y})$ measures the fit of $\boldsymbol{\tau}$ with $\boldsymbol{y}$. Its role is to locate the change points as accurately as possible. The penalty term $\text{pen}(\boldsymbol{\tau})$ only depends on the dimension $K(\boldsymbol{\tau})$ of the model $\boldsymbol{\tau}$ and increases with $K(\boldsymbol{\tau})$. Thus, it is used for determining the number of change points. The penalization parameter β adjusts the trade-off between the minimization of $J(\boldsymbol{\tau}, \boldsymbol{y})$ (obtained with a high dimension of $\boldsymbol{\tau}$), and the minimization of $\text{pen}(\boldsymbol{\tau})$ (obtained with a small dimension of $\boldsymbol{\tau}$).

Asymptotic results concerning penalized least-squares estimates have been obtained in theoretical general contexts in [8, 10], extending the previous results of Yao [11]. We shall show that this kind of contrast can also be useful in practice. The main problem is the choice of a good penalty function and a good coefficient β. In the Gaussian case, Yao [11] suggests the Schwarz criterion. A complete discussion of the most popular criteria (AIC, Mallow's C_p, BIC), and many other references can be found in [2]. In a more general context, we can use a contrast other than the least-squares criterion, since the variables are not necessarily Gaussian and independent. Nevertheless, we propose an adaptive procedure for automatically choosing the penalty parameter β in Section 2.

In a Bayesian framework, we the conditional distribution of the change-point sequence $p(\boldsymbol{\tau}|\boldsymbol{y}) \propto \exp\{-\alpha(J(\boldsymbol{\tau}) + \beta \text{pen}(\boldsymbol{\tau}))\}$. Obviously, the mode of this distribution is the minimum penalized contrast estimate previously defined. A MCMC (Monte Karlo Markov Chain) procedure provides a way to sample and examine this posterior distribution, instead of only computing its mode. Furthermore, the artificial introduction of a "temperature" parameter allows us to concentrate this posterior distribution around the models $\boldsymbol{\tau}$ of highest probability. For the change-point problem, the so-called SAEM (Stochastic Approximation of Expectation-Maximization, [5]) algorithm provides an estimate of the parameters α and β. In the particular case of detecting jumps in the mean of a sequence of Gaussian variables, it was shown in [9] that this algorithm converges to the maximum likelihood estimate of α and β.

We apply these algorithms to an EEG recording with abrupt changes in its spectrum. Here, the contrast function we use is constructed from the empirical spectral distribution function.

The last section is devoted to some numerical experiments. The two proposed approaches estimate the number of changes in the mean and the variance well. On the other hand, AIC (Akaike Information Criteria) and BIC (Bayesian Information Criteria) strongly overestimate the number of change-points.

The Matlab programs are available at http://www.math.u-psud.fr/~lavielle/programs.

2 A penalized contrast estimate for the change-point problem

2.1 The contrast function

In most situations, the characteristic of the Y_i's that changes abruptly is a parameter $\theta \in \Theta$, that remains constant between two changes. We will strongly use this assumption to define our contrast function $J(\boldsymbol{\tau}, \boldsymbol{y})$.

Let K be some integer and let $\boldsymbol{\tau} = (\tau_1, \tau_2, \ldots, \tau_{K-1})$ be a sequence of integers satisfying $0 < \tau_1 < \tau_2 < \ldots < \tau_{K-1} < n$. For any $1 \leq k \leq K$, let $U(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}; \theta)$ be a contrast function useful for estimating the unknown true value of the parameter in the segment k. In other words, the minimum contrast estimate $\hat{\theta}(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k})$, computed on segment k of $\boldsymbol{\tau}$, is defined as a solution of the following minimization problem:

$$U(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}; \hat{\theta}(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k})) \leq U(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}; \theta) \quad , \ \forall \theta \in \Theta, \tag{1}$$

For any $1 \leq k \leq K$, let G be

$$G(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}) = U(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}; \hat{\theta}(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k})). \tag{2}$$

Then, define the contrast function $J(\boldsymbol{\tau}, \boldsymbol{y})$ as

$$J(\boldsymbol{\tau}, \boldsymbol{y}) = \frac{1}{n} \sum_{k=1}^{K} G(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}) \tag{3}$$

where $\tau_0 = 0$ and $\tau_K = n$.

Several examples of contrast functions are given in the sections devoted to numerical experiments.

When the true number $K^\star$ of segments is known, the sequence $\hat{\boldsymbol{\tau}}_n$ of change-point instants that minimizes this kind of contrast has the property (see [8, 10]) that, under extremely general conditions, for any $1 \leq k \leq K^\star - 1$,

$$\mathrm{P}(|\hat{\tau}_{n,k} - \tau_k^\star| > \delta) \to 0 \quad \text{when } \delta \to \infty \text{ and } n \to \infty. \tag{4}$$

In particular, this result holds for weakly and strongly dependent process.

As an example, consider the following model

$$Y_i = \mu_i + \sigma_i \varepsilon_i, \quad 1 \leq i \leq n \tag{5}$$

where (ε_i) is a sequence zero-mean random variables with unit variance.

In the case of changes in the mean, for example, we assume that (μ_i) is a piecewise constant sequence and (σ_i) is a constant sequence. In otherwords, there exist some instants $\tau_1^\star < \tau_2^\star < \ldots < \tau_{K^\star-1}^\star$ such that, for any $1 \leq k \leq K^\star$, $\mu_{\tau_{k-1}^\star+1} = \mu_{\tau_{k-1}^\star+2} = \ldots = \mu_{\tau_k^\star}$. A Gaussian log-likelihod can be used to define the contrast function, even if (ε_i) is not a Gaussian sequence. Let

$$U(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}; \mu) = \sum_{i=\tau_{k-1}+1}^{\tau_k} (Y_i - \mu)^2. \tag{6}$$

Then,

$$G(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}) = \sum_{i=\tau_{k-1}+1}^{\tau_k} (Y_i - \overline{Y}_{\tau_{k-1}+1:\tau_k})^2 \tag{7}$$

where $\overline{Y}_{\tau_{k-1}+1:\tau_k}$ is the empirical mean of $(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k})$.

On the other hand, changes in the variance means that (μ_i) is a constant sequence and (σ_i) is a piecewise constant sequence. As before, a Gaussian log-likelihod can be used to define the contrast function, even if (ε_i) is not a Gaussian sequence. Let $\mu = \mu_1 = \ldots = \mu_{\tau_k^\star}$ and

$$U(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}; \sigma^2) = (\tau_k - \tau_{k-1}) \log(\sigma^2) + \frac{1}{\sigma^2} \sum_{i=\tau_{k-1}+1}^{\tau_k} (Y_i - \mu)^2. \tag{8}$$

Then,

$$G(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}) = (\tau_k - \tau_{k-1}) \log(\hat{\sigma}^2_{\tau_{k-1}+1:\tau_k}) \tag{9}$$

where

$$\hat{\sigma}^2_{\tau_{k-1}+1:\tau_k} = \frac{1}{\tau_k - \tau_{k-1}} \sum_{i=\tau_{k-1}+1}^{\tau_k} (Y_i - \overline{Y})^2$$

is the empirical variance of $(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k})$, and $\overline{Y}$ is the empirical mean of $Y_1, \ldots Y_n$.

If the changes affect both the mean and the variance, a contrast function based on a Gaussian log-likelihood is

$$G(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k}) = (\tau_k - \tau_{k-1}) \log(\hat{\sigma}^2_{\tau_{k-1}+1:\tau_k}) \tag{10}$$

where

$$\hat{\sigma}^2_{\tau_{k-1}+1:\tau_k} = \frac{1}{\tau_k - \tau_{k-1}} \sum_{i=\tau_{k-1}+1}^{\tau_k} (Y_i - \overline{Y}_{\tau_{k-1}+1:\tau_k})^2$$

2.2 Penalty function for the change-point problem

When the number of change points is unknown, it can be estimated by minimizing a penalized version of $J(\boldsymbol{\tau}, \boldsymbol{y})$. For any sequence of change-point instants $\boldsymbol{\tau}$, let $\mathrm{pen}(\boldsymbol{\tau})$ be a function of $\boldsymbol{\tau}$ that increases with the number $K(\boldsymbol{\tau})$ of segments of $\boldsymbol{\tau}$. Then, let $\hat{\boldsymbol{\tau}}_n$ be the sequence of change-point instants that minimizes

$$H(\boldsymbol{\tau}) = J(\boldsymbol{\tau}, \boldsymbol{y}) + \beta \mathrm{pen}(\boldsymbol{\tau}). \tag{11}$$

If β is a function of n that goes to 0 at an appropriate rate as n goes to infinity, the estimated number of segments $K(\hat{\boldsymbol{\tau}}_n)$ converges in probability to $K^\star$ and (4) still holds (see [8, 10] for more details).

Given a real observed signal with a fixed, finite length n, asymptotic results are not very useful for selecting the penalty term $\beta\mathrm{pen}(\boldsymbol{\tau})$. Various authors suggest different penalty functions, according to the model they consider. For example, the Schwarz criterion is used by Braun *et al.* [3] for detecting changes in a DNA sequence.

Consider first the penalty function $\mathrm{pen}(\boldsymbol{\tau})$. By definition, $\mathrm{pen}(\boldsymbol{\tau})$ should increase with the number of segments $K(\boldsymbol{\tau})$. Following the most popular information criteria such as AIC and the Schwarz criteria, we suggest to use in practice the simplest penalty function $\mathrm{pen}(\boldsymbol{\tau}) = K(\boldsymbol{\tau})$.

Remark: We can defend this specific choice of the penalty function with theoretical considerations. Indeed, precise results have been recently obtained by Birgé and Massart [2] in the following model:

$$Y_i = s^\star(i) + \sigma\varepsilon_i, \quad 1 \leq i \leq n \tag{12}$$

where $s^\star(i) = \sum_{k=1}^{K^\star} m_k \mathbf{1}_{\{\tau^\star_{k-1}+1 \leq i \leq \tau^\star_k\}}$ is a piecewise constant function. The sequence (ε_i) is a sequence of Gaussian white noise, with variance 1. A penalized least-squares estimate is obtained by minimizing

$$H(\boldsymbol{\tau}, \boldsymbol{y}) = \frac{1}{n} \sum_{k=1}^{K(\boldsymbol{\tau})} \sum_{i=\tau_{k-1}+1}^{\tau_k} (Y_i - \overline{Y}_k)^2 + \beta\mathrm{pen}(\boldsymbol{\tau}), \tag{13}$$

In a non asymptotic context, Birgé and Massart [2] have shown that a penalty function of the form

$$\mathrm{pen}(\boldsymbol{\tau}) = K(\boldsymbol{\tau}) \left(1 + c \log \frac{n}{K(\boldsymbol{\tau})}\right) \quad , \quad \beta = \frac{2\sigma^2}{n} \tag{14}$$

is optimal for minimizing $\mathbb{E}\left(\|\hat{s}_{\boldsymbol{\tau}} - s^\star\|^2\right)$, where $\hat{s}_{\boldsymbol{\tau}}(i) = \sum_{k=1}^{K(\boldsymbol{\tau})} \overline{Y}_k \mathbf{1}_{\{\tau_{k-1}+1 \leq i \leq \tau_k\}}$ is the estimated sequence of means. Based on some numerical experiments, the authors suggest to use $c = 2.5$. Note that when the number $K^\star$ of segments is small compared to the length n of the series, this optimal penalty function is an almost linear function of K. Furthermore, Yao [11] has proved the consistency of the Schwarz criterion for this model, here meaning $\mathrm{pen}(\boldsymbol{\tau}) = K(\boldsymbol{\tau})$ and $\beta = 2\sigma^2(\log n)/n$.

2.3 An adaptive choice of the penalization parameter

For a given contrast function J and a given penalty function pen, the problem now reduces to the choice of the parameter β.

Let K_{MAX} be an upper bound on the dimension of $\boldsymbol{\tau}$. For any $1 \leq K \leq K_{MAX}$, let $\mathcal{T}_K$ be the set of all the models of dimension K:

$$\mathcal{T}_K = \{\boldsymbol{\tau} = (\tau_0, \ldots, \tau_K) \in \mathbb{N}^{K+1} \ , \ \tau_0 = 0 < \tau_1 < \tau_2 < \ldots \tau_{K-1} < \tau_K = n\} \tag{15}$$

By definition the best model $\hat{\boldsymbol{\tau}}_K$ of dimension K minimizes the contrast function J:

$$\hat{\boldsymbol{\tau}}_K = \arg \min_{\boldsymbol{\tau} \in \mathcal{T}_K} J(\boldsymbol{\tau}, \boldsymbol{y}), \tag{16}$$

Note that the sequence $(\hat{\boldsymbol{\tau}}_K, 1 \leq K \leq K_{MAX})$ can easily be computed. Indeed, let $\mathcal{G}$ be the upper triangular matrix of dimension $n \times n$ such that the element (i,j), for $j \geq i$ is $\mathcal{G}_{i,j} = G(Y_i, Y_{i+1}, \ldots Y_j)$, where $G(Y_i, \ldots Y_j)$ is the contrast computed with $(Y_i, Y_{i+1}, \ldots Y_j)$. Thus, for any $1 \leq K \leq K_{MAX}$, we have to find a path $\tau_0 = 0 < \tau_1 < \tau_2 < \ldots, < \tau_{K-1} < \tau_K = n$ that minimizes the total cost

$$J(\boldsymbol{\tau}, \boldsymbol{y}) = \frac{1}{n} \sum_{k=1}^{K} \mathcal{G}_{\tau_{k-1}, \tau_k}. \tag{17}$$

A dynamic programming algorithm can recursively compute the optimal paths $(\hat{\boldsymbol{\tau}}_K, 1 \leq K_{MAX})$, see [7]. This algorithm requires $\mathcal{O}(n^2)$ operations (size of the matrix $\mathcal{G}$).

Then, let

$$\begin{aligned} J_K &= J(\hat{\boldsymbol{\tau}}_K, \boldsymbol{y}), & (18) \\ p_K &= \text{pen}(\boldsymbol{\tau}) \ , \quad \forall \boldsymbol{\tau} \in \mathcal{T}_K & (19) \end{aligned}$$

(as mentioned above, we suggest to use $p_K = K$).

Thus, for any penalization parameter $\beta > 0$, the solution $\hat{\boldsymbol{\tau}}(\beta)$ minimizes the penalized contrast:

$$\begin{aligned} \hat{\boldsymbol{\tau}}(\beta) &= \arg \min_{\boldsymbol{\tau}} (J(\boldsymbol{\tau}, \boldsymbol{y}) + \beta \text{pen}(\boldsymbol{\tau})) & (20) \\ &= \hat{\boldsymbol{\tau}}_{\hat{K}(\beta)} & (21) \end{aligned}$$

where

$$\hat{K}(\beta) = \arg \min_{K \geq 1} \{J_K + \beta p_K\}. \tag{22}$$

The way the solution $\hat{K}(\beta)$ varies with the penalization parameter β is given in the following proposition:

Proposition 2.1 *There exists a sequence $K_1 = 1 < K_2 < \ldots$, and a sequence $\beta_0 = \infty > \beta_1 > \ldots$, with*

$$\beta_i = \frac{J_{K_i} - J_{K_{i+1}}}{p_{K_{i+1}} - p_{K_i}} \ , \quad i \geq 1 \tag{23}$$

such that $\hat{K}(\beta) = K_i$, $\forall \beta \in (\beta_i, \beta_{i-1})$.

The subset $\{(p_{K_i}, J_{K_i}), i \geq 1\}$ is the convex hull of the set $\{(p_K, J_K), K \geq 1\}$.

proof: For any $K \geq 1$, let $\hat{K}(\beta) = K$. Then

$$\begin{aligned} J_K + \beta p_K &< \min_{L>K} (J_L + \beta p_L) \qquad (24) \\ J_K + \beta p_K &< \min_{L<K} (J_L + \beta p_L) \qquad (25) \end{aligned}$$

Thus, β must satisfy

$$\max_{L>K} \frac{J_K - J_L}{p_L - p_K} < \beta < \min_{L<K} \frac{J_L - J_K}{p_K - p_L} \qquad (26)$$

■

The estimated sequence $\hat{\boldsymbol{\tau}}$ should not strongly depend on the choice of the penalization coefficient β. In other words, a small change of β should not lead to a radically different solution $\hat{\boldsymbol{\tau}}$. This stability of the solution with respect to the choice of β will be ensured if we only retain the largest intervals $([\beta_i, \beta_{i-1}], i \geq 1)$.

In summary, we propose the following procedure:

1. For $K = 1, 2, \ldots, K_{MAX}$, compute $\hat{\boldsymbol{\tau}}_K$, $J_K = J(\hat{\boldsymbol{\tau}}_K, \boldsymbol{y})$ and $p_K = \text{pen}(\hat{\boldsymbol{\tau}}_K)$,
2. compute the sequences (K_i) and (β_i), and the lengths (l_i) of the intervals $([\beta_i, \beta_{i-1}])$,
3. retain the greatest value(s) of K_i such that $l_i >> l_j$, for $j > i$.

Remark 1: Choosing the largest interval usually under-estimates the number of changes. Indeed, this interval usually corresponds to a very small number of change points and we only detect the most drastic changes with such a penalty. This explains why we should better look for the highest dimension K_i such that $l_i >> l_j$, for any $j > i$, to recover the smallest details.

Remark 2: A classical and natural graphical method for selecting the dimension K can be summarized as follows: *i)* examine how the contrast J_K decreases when K (that is, p_K) increases ; *ii)* select the dimension K for which J_K ceases to decrease significatively. In other words, this heuristic approach looks for maximum curvature in the plot (p_K , J_K). Proposition 2.1 states that the second derivative of this curve is directly related to the length of the intervals $([\beta_i, \beta_{i-1}], i \geq 1)$. Indeed, if we represent the points (p_K , J_K), for $1 \leq K \leq K_{MAX}$, β_i is the slope between the points (p_{K_i}, J_{K_i}) and $(p_{K_{i+1}}, J_{K_{i+1}})$. Thus, to look for where J_K ceases to decrease means to look for a break in the slope of this curve. Now, the variation of the slope at the point (p_K, J_K) is precisely the length l_i of the interval $[\beta_i, \beta_{i-1}]$.

2.4 An automatic procedure for estimating K

For a practical purpose, it can be useful to perform automatically the step 3 of our procedure. We propose the following algorithm for estimating the dimension K:

1. For any $1 \leq K \leq K_{MAX}$, let
$$\tilde{J}_K = \frac{J_{K_{MAX}} - J_K}{J_{K_{MAX}} - J_1}(K_{MAX} - 1) + 1 \tag{27}$$
The new sequence $(\tilde{J}_K)$ is normalized such that $\tilde{J}_1 = K_{MAX}$ and $\tilde{J}_{K_{MAX}} = 1$. This sequence decreases with an average slope equal to -1.

2. For any $2 \leq K \leq K_{MAX} - 1$, let $D_K = \tilde{J}_{K-1} - 2\tilde{J}_K + \tilde{J}_{K+1}$ and $D_1 = \infty$. Then, the Minimum Penalized Contrast (MPC) estimate of K is
$$\hat{K}_{MPC} = \max\left\{1 \leq K \leq K_{MAX} - 1 \quad \text{such that } D_K > S\right\} \tag{28}$$
$\hat{K}_{MPC}$ is defined as the greatest value of K such that the second derivative of J is greater than a given threshold S. If no second derivative is greater than S, we consider that there are no changes and $\hat{K}_{MPC} = 1$.

Unfortunately, the probability distribution of the statistics $\max_K D_K$ cannot be obtained in a closed form and the threshold S cannot be computed as a quantile of this distribution. Nevertheless, many different numerical experiments led us to propose $S = 0.75$. Indeed, we have noticed that smallest values of S usually over-estimate the number of segments, while larger values under-estimate this number.

3 A Bayesian approach

The minimization of a contrast of the form $J(\boldsymbol{\tau}, \boldsymbol{y}) + \beta \text{pen}(\boldsymbol{\tau})$, for a given value of β, produces only one solution $\hat{\boldsymbol{\tau}}_{\hat{K}(\beta)}$, but the description of the "energy landscape" J is to construct. Indeed, it can be interesting to see how J varies when we slightly modify $\hat{\boldsymbol{\tau}}_{\hat{K}(\beta)}$. In the case of change points, what happens if we move a change point, or if we add or remove a segment?

An easy way to describe this energy landscape J consists in constructing the following posterior distribution:
$$p(\boldsymbol{\tau}|\boldsymbol{y}; \alpha, \beta) = D(\boldsymbol{y}; \alpha, \beta) e^{-\alpha(J(\boldsymbol{\tau}, \boldsymbol{y}) + \beta \text{pen}(\boldsymbol{\tau}))} \tag{29}$$
where $D(\boldsymbol{y}; \alpha, \beta)$ is a normalizing constant, and where $\alpha > 0$. Thus, the mode of this posterior distribution is the minimum contrast estimate of $\boldsymbol{\tau}$. This posterior distribution depends on two unknown parameters α and β that should be estimated.

Lavielle and Lebarbier consider in [9] the problem of detecting changes in the mean of a sequence of random variables. They use an MCMC procedure for estimating the posterior distribution $p(\boldsymbol{\tau}|\boldsymbol{y}; \alpha, \beta)$. They also use the SAEM algorithm proposed by Delyon *et al.* [5], for computing the maximum likelihood estimate of (α, β).

We propose to adopt the same approach in a more general context. We present here this methodology without giving any more details. The description of the MCMC procedure and the SAEM algorithm can be found in [9].

1. Estimate (α, β) using SAEM.
2. For different values of T, with $0 < T \leq 1$,
   - use the MCMC algorithm to sample the conditional distribution $p(\boldsymbol{\tau}|\boldsymbol{y}; \hat{\alpha}/T, \hat{\beta})$,
     (a) estimate and plot the marginal conditional probabilities P("There is a change-point at t_i"$|\boldsymbol{y}; \hat{\alpha}/T, \hat{\beta})$, for $1 \leq i \leq n-1$, where t_i is the ith instant of observation.
     (b) estimate and plot the conditional probability $\mathrm{P}(K(\boldsymbol{\tau}) = k|\boldsymbol{y}; \hat{\alpha}/T, \hat{\beta})$
   - compute the Maximum a Posteriori (MAP) estimate of $\boldsymbol{\tau}$ by minimizing $J(\boldsymbol{\tau}, \boldsymbol{y}) + \hat{\beta}\mathrm{pen}(\boldsymbol{\tau})$.

The "tuning" parameter T is usually called "temperature". This parameter controls how the distribution p is concentrated around its mode. It should be chosen small enough to neglect the models $\boldsymbol{\tau}$ having a low posterior probability and to increase the probability of the most likely models. Here, the MCMC algorithm creates an homogeneous Markov Chain since the temperature parameter remains constants. Maximization of the conditional distribution could be achieved using a simulated annealing procedure. In this case, the temperature is not constant but decreases slowly to zero. Simulated annealing is very slow and the dynamic programing algorithm described above should be preferred for computing the mode of this distribution.

We will denote $\hat{\boldsymbol{\tau}}_{MAP}$ the MAP estimate of $\boldsymbol{\tau}$ and $\hat{K}_{MAP} = K(\hat{\boldsymbol{\tau}}_{MAP})$ the MAP estimate of K. Here,

$$\hat{K}_{MAP} = \arg\min_{K}(J_K + \hat{\beta} p_K). \tag{30}$$

4 Application to EEG segmentation

It is well known that EEG recordings present abrupt changes in its spectrum. Indeed, epileptogenic transients can produce changes in the following frequency bands: δ (1.5-3.5 Hz), θ (3.5-7.5 Hz), α (7.5-12.5 Hz), and β (12.5-19.5 Hz). A EEG recording is dispayed Figure 1. We clearly see several changes in the spectral characteristics of the series. In particular, a very strong α-activity appears between 3.5s and 4.5s. We shall see that our procedure is very efficient for detecting automatically this kind of changes.

For any k and any $u \in [0, \pi]$, let

$$I_k(u) = \frac{1}{2\pi n_k} \left| \sum_{j=\tau_{k-1}+1}^{\tau_k} Y_j e^{iju} \right|^2 \tag{31}$$

be the periodogram of the sequence (Y_j) computed in segment k at frequency u. Assume that the energy of the process in some frequency bands $[\lambda_j, \mu_j)$, $1 \leq j \leq J$, of $[0, \pi]$ changes

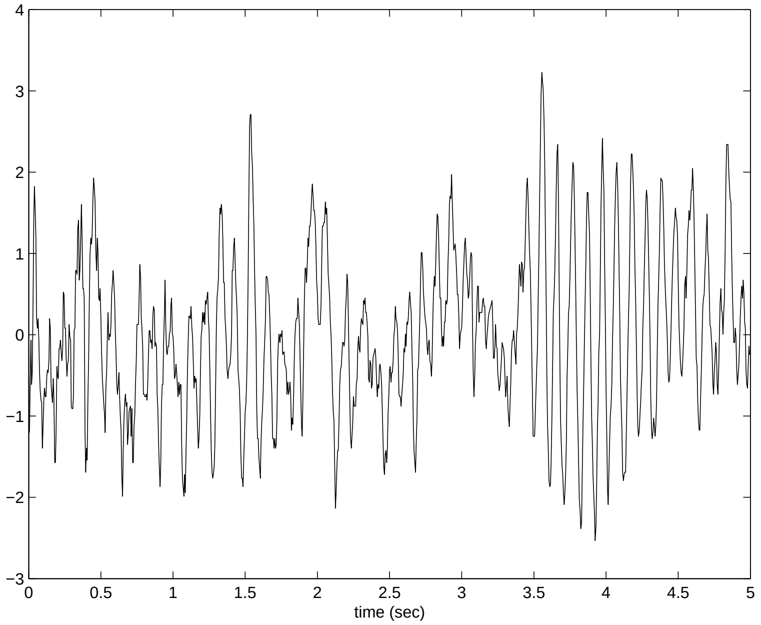

Figure 1: An example of EEG recording

abruptly at some unknown instant. Then, let

$$F_{kj} = \int_{\lambda_j}^{\mu_j} I_k(u)du \tag{32}$$

be the energy of $(Y_{\tau_{k-1}+1}, \ldots, Y_{\tau_k})$ in the frequency band $[\lambda_j, \mu_j)$. We suggest in [10] to use the following contrast for detecting the changes:

$$J_n(\boldsymbol{\tau}, \boldsymbol{y}) = -\frac{1}{n} \sum_{k=1}^{K^\star} \left(n_k \sum_{j=1}^{J} F_{kj}^2 \right) . \tag{33}$$

Figure 2 represents the points (K, J_K), $1 \leq K \leq K_{MAX} = 25$ obtained from the EEG recording in Figure 1, with the contrast function proposed in (33). We can easily see two breaks in the slope at $K = 5$ and $K = 3$.

The lengths (l_i) and the second derivatives (D_{K_i}) are displayed in Table 1. We clearly see that these two dimensions, and especially $K = 5$, are the only ones to be considered. Indeed, for $0.60 < \beta < 1.62$, $\hat{K}(\beta) = 5$. This interval is significantly larger than any of the following ones (for $\hat{K} \geq 7$). The second important interval is $[1.62, 4.22]$, related to $\hat{K} = 3$. The corresponding change points are displayed in Figure 3. The brain activity (α-activity)

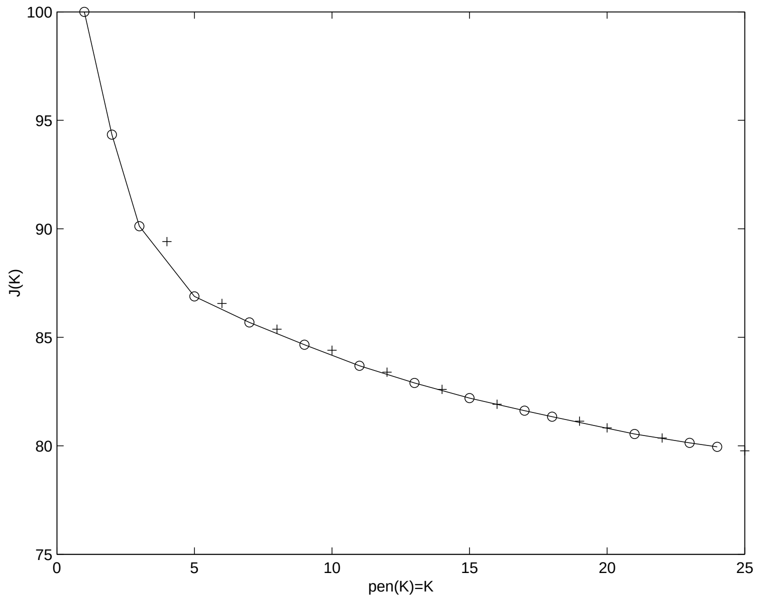

Figure 2: The points (p_K, J_K) for $1 \leq K \leq K_{MAX} = 25$. The points of the convex hull are represented with a circle and the other ones with a plus

between 3.5s and 4.5s is always well detected. On the other hand, there may be a little doubt concerning the activity detected around 1.5s, detected only in the first solution. The procedure described in Section 2.4 automatically detects the five segments with the threshold $S = 0.75$. Indeed, we obtain from (28) that $\hat{K}_{MPC} = 5$ since $D_5 > S$ and $D_K < S$ for any $K > 5$.

K_i	β_i	β_{i-1}	l_i	D_{K_i}
2	4.22	5.66	1.44	2.28
3	1.62	4.22	2.60	5.36
5	0.60	1.62	1.02	1.13
7	0.51	0.60	0.09	0.17
9	0.48	0.51	0.03	0.16
11	0.40	0.48	0.08	0.013
13	0.35	0.40	0.05	0.03
15	0.29	0.35	0.06	0.06

Table 1: Intervals of the penalization parameter and the second derivatives

The SAEM algorithm has been used with the EEG recording proposed above. We obtained $\hat{\alpha} = 5.35$ and $\hat{\beta} = 1.34$. Then, it is interesting to remark that $\hat{K}_{MAP} = \hat{K}_{MPC} = 5$.

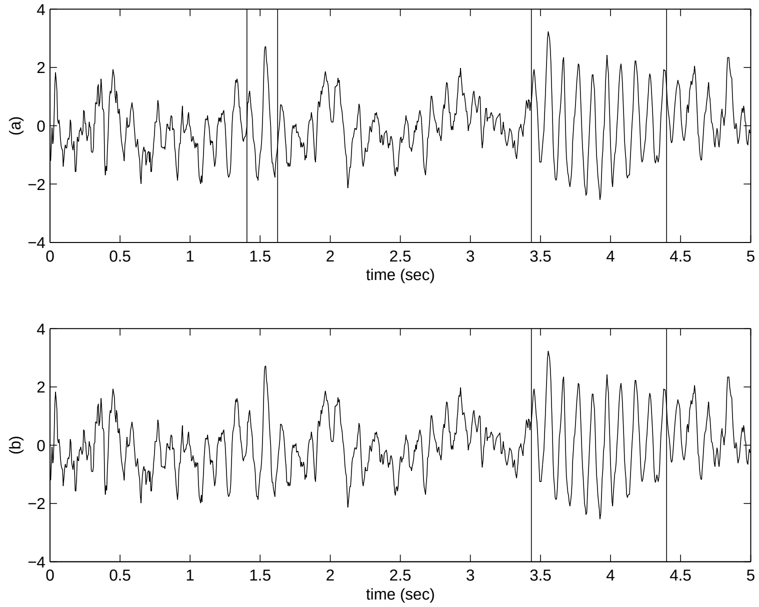

Figure 3: The two segmentations of the EEG recording obtained with $K = 5$ and $K = 3$ segments.

Indeed, this value of $\hat{\beta}$ belongs to the interval [0.60 , 1.62]. For this value of β, the MAP estimate of $\boldsymbol{\tau}$ is the minimum contrast estimate displayed in Figure 3-a, having 5 segments. On the other hand, this value of β is not very far from the next interval, [1.62 , 4.22], associated with three segments (see Figure 3-b).

The marginal posterior distributions P“There is a change at i”$|\boldsymbol{y};\hat{\alpha}/T,\hat{\beta}$ and the posterior distribution of the number of segments $PK(\boldsymbol{\tau}) = k|\boldsymbol{y};\hat{\alpha}/T,\hat{\beta}$ are displayed in Figure 4.

We can see that change-points are not always located exactly at the same instants. For example, the event around 4s is detected with a very high probability, but the end of this event (around 4.4s) fluctuates more than the beginning (around 3.4s). That shows that this instance of α-activity of the brain begins suddenly, but the return to a normal activity is more progressive. This feature cannot be described if we just compute the most likely change-point locations.

With $T = 1$, many change points, with very low probabilities, appear at any instant. This explains why the number of segments is often greater than 5. These events are not significant and should be removed. This is the role of a low temperature, since all the minor events disappear for $T = 0.6$. The two main events (around 1.5s and 4s) now clearly appear. Both are well detected with probability 0.67, while the first one is not detected with

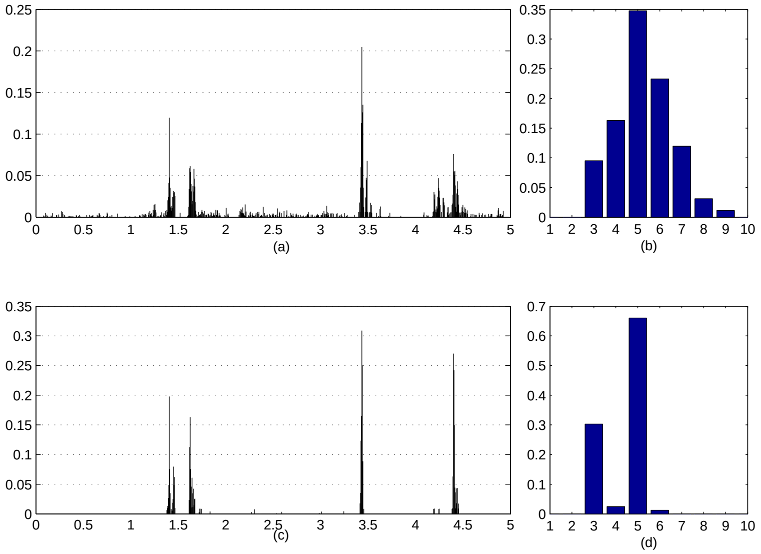

Figure 4: Estimation of the posterior distribution $P(\boldsymbol{\tau}|\boldsymbol{y};\hat{\alpha}/T,\hat{\beta})$ obtained with $T = 1$ in (a) and (b) ; $T = 0.6$ in (c) and (d). The marginal posterior probabilities $P($"There is a change-point at t_i"$|\boldsymbol{y};\hat{\alpha}/T,\hat{\beta})$, for $1 \leq i \leq n-1$ are displayed in (a) and (c) ; $P(K(\boldsymbol{\tau}) = k|\boldsymbol{y};\hat{\alpha}/T,\hat{\beta})$ are displayed in (b) and (d).

probability 0.30. Indeed, although $\hat{\beta}$ belongs to the interval associated to five segments, the probability of a model with only three segments is not negligible.

5 Some simulations

The aim of this section is to compare the Minimum Penalized Contrast estimate $\hat{K}_{MPC}$ proposed in Section 2 with the Maximum a Posteriori estimate $\hat{K}_{MAP}$ proposed in Section 3 and with some well known estimators. The Matlab programs are available at http://www.math.u-psud.fr/~lavielle/programs.

For each of the two models we consider below, the observed time series $\boldsymbol{y}$ has a length $n = 500$ and four change points are present at instants 100, 200, 300 and 400.

For the two parametric models considered below, the AIC and BIC criteria can be computed, assuming a Gaussian distribution.

5.1 Changes in the mean

Following [8] and [9], a least-squares criteria can be used for detecting changes in the mean:

$$J_n(\boldsymbol{\tau},\boldsymbol{y}) = \frac{1}{n}\sum_{k=1}^{K^\star}\sum_{i=\tau_{k-1}+1}^{\tau_k}(Y_i-\overline{Y}_k)^2, \tag{34}$$

where $\overline{Y}_k$ is the empirical mean of $(Y_{\tau_{k-1}+1},\ldots,Y_{\tau_k})$.

Here, the simulated series $\boldsymbol{y}$ are sequences of 500 independent Gaussian random variables of variance $\sigma^2 = 1$. The means in the five segments are $(0, a, 0, 2a, 0)$.

We simulated 100 series with $a = 0.5$ and 100 series with $a = 1$. For each of these series, we computed $\hat{K}_{MPC}$ using the procedure described in Section 2 and (28) with $S = 0.75$. We also computed $\hat{K}_{MAP}$ using (30). The results are summarized in Table 2. Of course, the smallest jump ($a = 0.5$) is not always detected since $\hat{K}_{MPC}$ (resp. $\hat{K}_{MAP}$) detects five segments 65 times (resp. 49 times) and three segments 27 times (resp. 43) times.

Recall that $\hat{\beta}$ is the maximum-likelihood estimate of β for this model. Then, it is interesting to remark that the two estimates $\hat{K}_{MPC}$ and $\hat{K}_{MAP}$ give very similar results. That means that maximum-likelihood estimation of the parameters of the model is a good criteria for a model selection purposes in these particular experiments.

Here, AIC and BIC both strongly overestimate the number of change-points.

	K	1	2	3	4	5	6	7	8	≥ 9
$a = 0.5$	$\hat{K}_{MPC}$	0	0	27	2	65	4	2	0	0
	$\hat{K}_{MAP}$	1	0	43	5	49	2	0	0	0
	$\hat{K}_{BIC}$	0	0	0	1	20	17	21	16	25
	$\hat{K}_{AIC}$	0	0	0	0	0	0	0	0	100
$a = 1$	$\hat{K}_{MPC}$	0	0	0	0	100	0	0	0	0
	$\hat{K}_{MAP}$	0	0	0	0	99	1	0	0	0
	$\hat{K}_{BIC}$	0	0	0	0	20	12	20	15	33
	$\hat{K}_{AIC}$	0	0	0	0	0	0	0	0	100

Table 2: Changes in the mean. Estimated number of segments

5.2 Changes in the variance

For the detection of changes in the variance of a sequence of random variables, the following contrast, based on a Gaussian log-likelihood, can be used :

$$J_n(\boldsymbol{\tau},\boldsymbol{y}) = \frac{1}{n}\sum_{k=1}^{K} n_k \log(\hat{\sigma}_k^2), \tag{35}$$

where $n_k = \tau_{k-1} - \tau_k$ is the length of segment k, $\hat{\sigma}_k^2 = (1/n_k) \sum_{i=\tau_{k-1}+1}^{\tau_k} (Y_i - \overline{Y})^2$ is the empirical variance computed on segment k, and $\overline{Y}$ the empirical mean of $Y_1, \ldots, Y_n$. It was shown in [8] that this estimate possesses very good asymptotic properties (see Section 2).

Here, the simulated series $\boldsymbol{y}$ are sequences of independent zero-mean Gaussian random variables. The variances in the five segments are $(1, 1+a, 1, 1+2a, 1)$. The results obtained with $a = 1$ and $a = 2$ are summarized in Table 3. As before, the number of change-points is overestimated using AIC or BIC.

	K	1	2	3	4	5	6	7	8	≥ 9
$a = 1$	$\hat{K}_{MPC}$	2	0	26	5	54	5	7	1	0
	$\hat{K}_{MAP}$	26	0	28	4	35	6	1	0	0
	$\hat{K}_{BIC}$	1	0	6	1	24	16	13	7	32
	$\hat{K}_{AIC}$	0	0	0	0	0	0	0	4	96
$a = 2$	$\hat{K}_{MPC}$	0	0	5	0	94	1	0	0	0
	$\hat{K}_{MAP}$	2	0	16	0	81	1	0	0	0
	$\hat{K}_{BIC}$	0	0	1	0	45	19	15	5	15
	$\hat{K}_{AIC}$	0	0	0	0	0	0	1	2	97

Table 3: Changes in the variance. Estimated number of segments

Conclusion

We have shown in this paper that a penalized contrast can be a powerful tool for the detection of abrupt change points. Using a model selection approach, an efficient procedure provides an automatic choice of the penalization parameter. On the other hand, an MCMC procedure allows us to estimate the penalization parameter and to sample a conditional distribution based on the penalized contrast function. The two proposed algorithms give very good results for the change-point problem. It should be now interesting to extend this approach to a more general context of model selection.

References

[1] Basseville M. and Nikiforov N., *The Detection of abrupt changes - Theory and applications.* Prentice-Hall: Information and System sciences series, 1993.

[2] Birgé L. and Massart P., "Gaussian model selection," *J. Eur. Math. Soc.*, vol. 3, no. 3, pp. 203–268, 2001.

[3] Braun J., Braun R., and Muller H., "Multiple changepoint fitting via quasilikehood, with application to DNA sequence segmentation," *Biometrika*, vol. 87, no. 2, pp. 301–314, 2000.

[4] Brodsky B. and Darkhovsky B., *Nonparametric methods in change-point problems.* Kluwer Academic Publishers, the Netherlands, 1993.

[5] Delyon B., Lavielle M., and Moulines E., "Convergence of a stochastic approximation version of the EM algorithm," *The Annals of Stat.*, vol. 27, no. 1, pp. 94–128, 1999.

[6] Gijbels I., Hall P., and Kneip A., "On the estimation of jump points in smooth curves," *The Annals of the Institute of Statistical Mathematics*, vol. 51, pp. 231–251, 1999.

[7] Kay S. M., *Fundamentals of statistical signal processing*, vol. 2. Prentice Hall PTR, 1998.

[8] Lavielle M., "Detection of multiple changes in a sequence of dependent variables," *Stoch. Proc. and Appli.*, vol. 83, pp. 79–102, 1999.

[9] Lavielle M. and Lebarbier E., "An application of MCMC methods to the multiple change-points problem," *Signal Processing*, vol. 81, pp. 39–53, 2001.

[10] Lavielle M. and Ludena C., "The multiple change-points problem for the spectral distribution," *Bernoulli*, vol. 6, no. 5, pp. 845–869, 2000.

[11] Yao Y., "Estimating the number of change-points via Schwarz criterion," *Stat. & Probab. Lett.*, vol. 6, pp. 181–189, 1988.

Contents

Unité de recherche INRIA Futurs
Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399