OPTIMAL SEGMENTATION OF NON-STATIONARY RANDOM PROCESSES

Marc LAVIELLE

Laboratoire de Statistiques, Bât 425, Université Paris-Sud, 91405 Orsay.
U.F.R. Mathématique, Univ. Paris V, 45 Rue des Saints-Pères 75006 Paris.

RÉSUMÉ

Segmenter un processus non stationnaire consiste à supposer qu'il est stationnaire par morceaux et à détecter les instants de rupture. On construit donc un processus de ruptures que l'on munit d'une distribution à priori. Cela nous permet de proposer comme solution l'estimateur du MAP. Un des intérêt majeur de cette méthode est sa capacité de fournir, par le choix d'un à priori judicieux, la meilleure solution possible, en accord avec le niveau de résolution choisi par l'utilisateur. Cet algorithme peut être utilisé dans de nombreux modèles paramétriques et non paramétriques. Il peut également être utilisé pour le lissage par fonctions spline lorsque le nombre et la position des nœuds sont inconnus. Des simulations et des applications sur des données réelles sont proposées.

ABSTRACT

Segmentation of a non-stationary process consists in assuming piecewise stationarity and in detecting the instants of change. Thus, we build a change process and define arbitrarily a prior distribution. That allows us to propose the MAP 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 the prior distribution chosen. The method can adress a wide class of parametric and non-parametric models. Furthermore, it can be used for smoothing a series with spline functions when the number and the position of the knots are unknown. Simulations and applications to real data are proposed.

1 Modelisation

Let $X = \{X_i\}_{i\geq 0}$ be a non-stationary real process. We assume that X is piecewise stationary. Then, there exist instants $\{t_k\}_{k\geq 0}$ such that $(X_{t_k+1},\ldots,X_{t_{k+1}})$ is stationary for all $k \in \mathbb{N}$. The problem consists in detecting the changes in the distribution of X, that is, in recovering the family $\{t_k\}$ when a trajectory $X_1 \ldots X_n$ is observed.

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

When the detection delay (the time beetween the change and its detection) needs to be well controled, a sequential detection is performed. That means to decide at time $t+\tau$ if a change has occured 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], [5].

The goal of these procedures is to minimize the probabilities of false alarms and omissions as well as the delay τ. Of course, simultaneous optimization of all these criteria is not possible: the smaller the delay is the bigger the number of false alarms is.

We shall assume here that all the data is available and that there is no time restriction. 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, let $R = \{R_i\}_{i\geq 0}$ be the random process defined by:

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

Then, R takes the value 1 at the change instants and is zero beetween two changes. Detecting the instants of changes consists in recovering the change process R. When a particular realization x of the process X is observed, we can think in the MAP (Maximum a Posteriori) estimate, maximizing the a posteriori distribution $P(R/X = x, \theta)$ with respect to R and θ. When this distribution possesses a density $g(r/x, \theta)$ with respect to a given measure, the MAP estimate is given by:

$$(\hat{r}, \hat{\theta}) = \text{Argmax } g(r/x, \theta) \quad (2)$$

$$= \text{Argmax } h(x/r, \theta)\pi(r). \quad (3)$$

where $h(x/r, \theta)$ is the density of the conditionnal distribution $P(X/R = r, \theta)$ and $\pi(r)$ is the prior probability to have the configuration r.

Thus, for a given configuration of changes r, $\hat{\theta}$ is the maximum likelihood estimate of θ in each stationnary piece and will be denoted $\hat{\theta}(r)$.

The first term $h(x/r, \theta)$ depends directly on the model while $\pi(r)$ must be arbitrarly defined. With no additional information, we could think in defining R as a sequence of independent Bernoulli variables:

$$\pi(r) = Z(\alpha)e^{-\alpha \sum_{i=1}^{n} r_i} \quad (4)$$

where $Z(\alpha)$ is a constant to make π a probability measure.

Writing now $h(x/r, \theta) = c \exp -l_x(\theta, r)$, we have to minimize the *energy function*

$$U_x(r) = l_x(\hat{\theta}(r), r) + \alpha \sum_{i=1}^{n} r_i. \quad (5)$$

The first term $l_x(\hat{\theta}(r), r)$ is related to the fit to the observation x while the second term is related to the number of changes. In fact, parameter α controls the probabilities of detection errors. The smaller α is, the bigger the prior probability of a change is and the fewer the omissions are. On the other hand, the bigger α is and the fewer the false alarms are.

For a process of length n, a change can occur at the $n-1$ first instants. Thus, R takes its values in a 2^{n-1}-dimensional space. Conventionally, we shall set $r_n = 1$ such that the number of changes $N_r = \sum r_i$ is the same as the number of segments. Let r^* be the configuration that minimizes $U_x(r)$. Computation of the 2^{n-1} values of $U_x(r)$ is not generally tractable. Nevertheless, a simulated annealing procedure is very efficient to reach the solution r^* [3].

2 Examples

2.1 Changes in a polynomial trend

We consider the following process:

$$X_k = f_l(k) + \varepsilon_k \quad (6)$$

for any $t_{l-1} < k \leq t_l$ and where f_l is a polynomial function of degree d and ε an additive noise. If ε is a Gaussian white noise of variance σ_ε^2, it is easy to show that r^* is computed by minimizing

$$U_x(r) = \sum_{l=1}^{N_r} \sum_{k=t_{l-1}+1}^{t_l} (X_k - \hat{f}_l(k))^2 + \beta N_r \quad (7)$$

where $\beta = 2\alpha\sigma_\varepsilon^2$ and $\hat{f}_l$ is the estimate of f_l.

Changes in the mean of a process: We consider here polynoms of degree 0 such that $f_l(k) = \mu_l$ for any $t_{l-1} < k \leq t_l$. Then, r^* is computed by minimizing

$$U_x(r) = \sum_{l=1}^{N_r} \sum_{k=t_{l-1}+1}^{t_l} (X_k - \hat{\mu}_l)^2 + \beta N_r \quad (8)$$

Here, $\hat{\mu}_l$ is the empirical mean of X on the lth segment. The parameter β will define the resolution level. We propose in Figure 1 different segmentations of a same process, obtained with different values of β. If we want details, that is to detect small changes of μ, we must choose a small value for β. On the other hand, only the more important jumps of the mean are detected with a bigger value of β.

Now, if we want to detect changes in both the mean and the variance of X, we must minimize the function

$$U_x(r) = \sum_{l=1}^{N_r} n_l \text{Log}\hat{\sigma}_l^2 + \beta N_r \quad (9)$$

where $\hat{\sigma}_l^2$ is the estimated variance of X on the lth segment and n_l its length.

Smoothing with spline functions: We shall minimize the function $U_x(r)$ under constraints, asking f and its $d-1$ first derivatives to be continuous. In this case, our problem is equivalent to that of smoothing a series X with spline functions, but when the place and the number of the knots are unknown.

We applied this algorithm to real data. The trajectory of a point of a spine on a particular axis during a flexion (in practice, we dispose of three curves, corresponding to the three axis) is displayed. This trajectory has been recorded by a camera and is noisy. For its analysis, a smoothed version of this trajectory is required. Furthermore, it can be very important to detect changes in the curvature of this trajectory, that is, in the second derivative of the function f. The segmentation algorithm has been used with polynomial splines of degree 2. The smoothed curve is displayed in Figure 2 with the original serie.

2.2 Changes in the transitions of a Markov Random Chain

We consider here a Markov Random Chain X that takes its values in a set S. Then, the MAP estimate of r is obtained by minimizing

$$U_x(r) = \sum_{l=1}^{N_r} \sum_{(i,j)\in S\times S} n_l(ij)\mathrm{Log}\frac{n_l(ij)}{n_l(i)} + \alpha N_r \quad (10)$$

where $n_l(i)$ is the number of times that X takes the value i in the lth segment and $n_l(ij)$ the number of times that X passes from i to j in the lth segment.

2.3 Changes in a non-parametric distribution

We consider now a sequence X of independent random variables such that the distribution of X is piecewise constant. 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 $z_0 < z_1 < \ldots < z_M$. For each X_k, we define a new variable Y_k that takes the value m when $z_{m-1} < X_k \leq z_m$.

The distribution of Y can be seen as the projected distribution of X. Assuming that the changes that affect the distribution of X affect the projected distribution, we shall recover r by maximizing the posterior distribution $P(R/Y = y)$. Thus, it can be shown that the solution is obtained by minimizing

$$U_x(r) = \sum_{l=1}^{N_r} \sum_{m=1}^{M} n_l(m)\mathrm{Log}\frac{n_l(m)}{n_l} + \alpha N_r \qquad (11)$$

where $n_l(m)$ is the number of times that Y takes the value m in the lth segment and n_l the length of lth segment.

A simulation is shown in Figure 3-a. Independent Gaussian variables were simulated in the first and third segments while a uniform distribution was used in the second segment. In this examples, the changes were well detected by the algorithm and a value of α beetween 8 and 12. The changes are not significative enough to be detected with a value of α greater than 12 while a value smaller than 8 produces false alarms.

Finally, Figure 3-b represents the heart-rate of a new born baby. We want to identify heavy and light sleep periods from this series. Instead of a parametric modelisation (that gives poor results here), we look for changes in a non-parametric distribution. The changes detected by the algorithm with $200 \leq \alpha \leq 400$ are the vertical full lines. In this example, external measurements (such as that of the eye-lids' movements) allow us to know the real instant of changes, indicated with dashed-lines.

3 The optimization procedure

The simulated annealing algorithm is an iterative procedure that defines a non-homogeneous Markov Chain $\{r(k)\}_{k\geq 0}$ that converges to the optimal solution r^* with probability one:

- Choose an initial configuration $r(0)$.
- At iteration k, choose a new configuration $\tilde{r}$ as a modification of $r(k-1)$. Let $\Delta U = U_x(\tilde{r}) - U_x(r(k-1))$.

  Set $r(k) = \tilde{r}$ with probability one if $\Delta U < 0$ and with probability $e^{\frac{-\Delta U}{T(k)}}$ elsewhere. Here $\{T(k)\}$ is a decreasing sequence called temperature.

When the total energy is a sum of local potentials, ΔU is easy to compute for local modifications. In our segmentation algorithm, the modifications consist in adding a new change, in removing one, and in translating one.

When we are looking for changes in a polynomial trend, with some restrictions on the continuity of the derivatives, this is no more true. In fact, a local modification of the configuration, such as a new change, will modify the complete set of spline functions to satisfy the constraints. Nevertheless, a modified version of the algorithm can be used in such a case [3]. We shall define a new energy function at iteration k:

$$\tilde{U}_x(r,k) = U_x(r) + \lambda(k)\sum_{j=0}^{d}\sum_{l=1}^{N_r-1}(f_l^{(j)}(t_l) - f_{l+1}^{(j)}(t_l))^2$$

where $f_l^{(j)}(t_l)$ is the jth derivative of f at the lth knot t_l.

Here, $\{\lambda(k)\}$ is a sequence that tends to infinity to ensure that the constraints are well satisfied for k large enough.

References

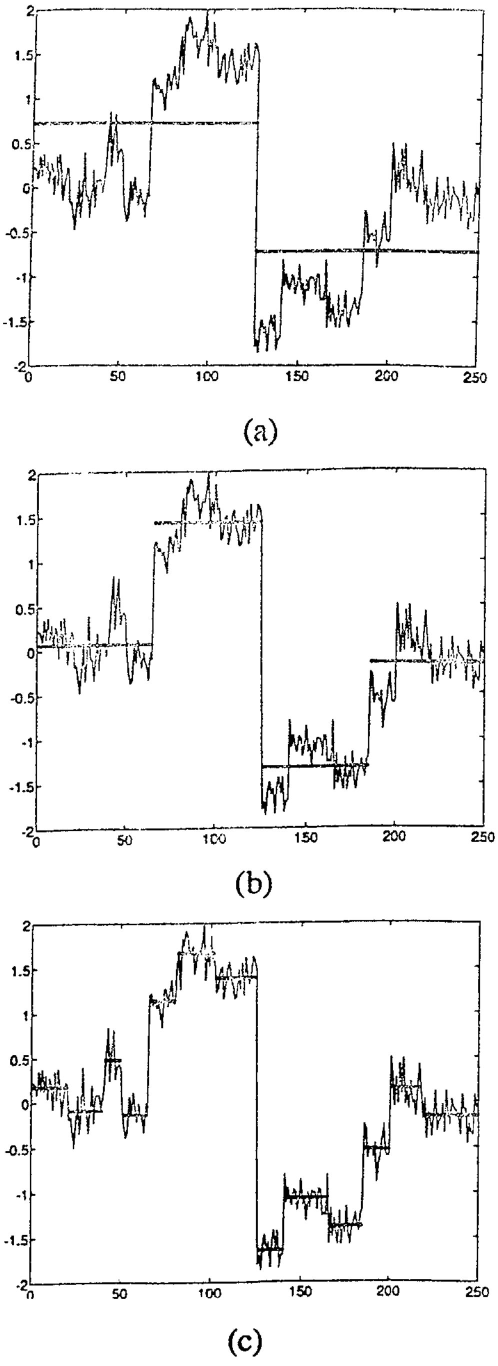
[1] Benveniste A., Metivier M., Priouret P. (1987) Algorithmes adaptatifs et approximations stochastiques (Masson).

[2] Canny J. (1986) A Computational Approach to Edge Detection. *IEEE* vol PAMI-8, N 6, 679-698.

[3] Geman D. (1990) Random Fields and Inverse Problems in Imaging. *Lecture note in Mathematics, Ecole d'été de Probabilité de Saint Flour XVIII, 1988.* vol 1427. Springer Verlag.

[4] Hinckley D.V. (1970) Inference about the change-point in a sequence of random variables. *Biometrika* Vol 57, 1-17.

[5] Lavielle M. (1993), Detection of Changes in the Spectrum of a Multidimensional Process. *IEEE* vol SP-41, N 2.



(a)

(b)

(c)

Fig.1 Detection of changes in the mean of a random process with different levels of resolution: a) $\beta = 10$ b) $\beta = 5$ c) $\beta = 2$

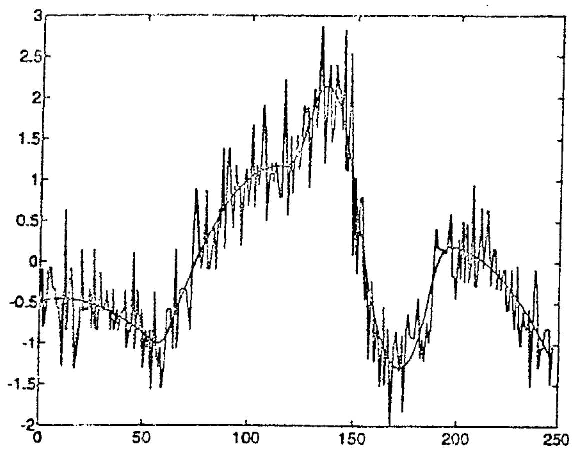
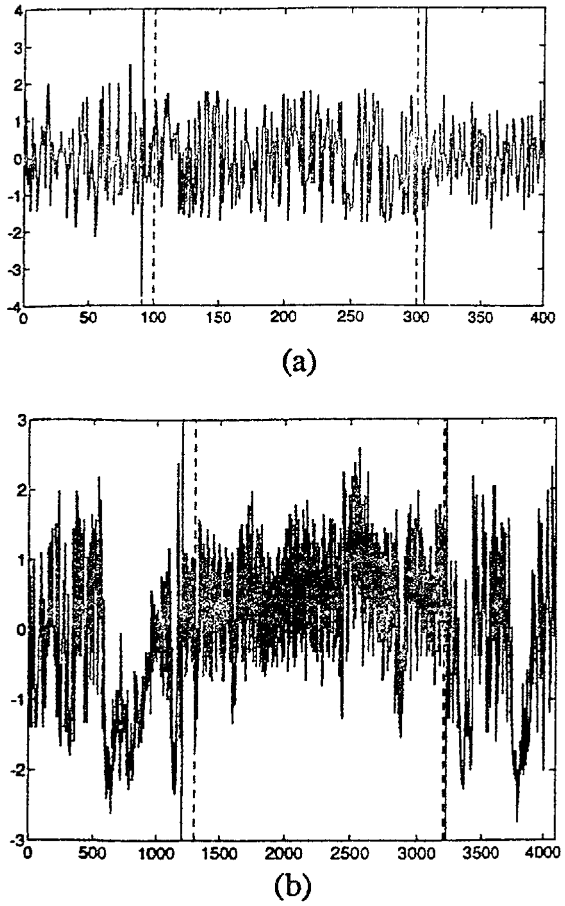


Fig.2 Smoothing with polynomial splines.by detecting changes in the second derivative. The original series and the smoothed curve.



(a)

(b)

Fig.3 Detection of changes in a non-parametric distribution. - - - : the original changes. —— the estimated changes.
a) Simulated data b) real data: the heart-rate of a new-born baby.