1 Modelisation

Let \( X = \{X_t\}_{t \geq 0} \) be a non-stationary real process. We assume that \( X \) is piecewise stationary. Then, there exist instants \( \{t_k\}_{k \geq 0} \) such that \( (X_{t_{k+1}}, \ldots, X_{t_k}) \) 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 \( \theta \). Thus, the problem consists now in 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, the coefficients of a polynomial trend, etc …

When the detection delay (the time between 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 occurred at time \( t \). Most of the test statistics used by the detection algorithms are built from the likelihood ratio or the Kullback distance \([1, 4, 5]\).

The goal of these procedures is to minimize the probabilities of false alarms and omissions as well as the delay \( \tau \). 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 \( R = \{R_t\}_{t \geq 0} \) be the random process defined by:

\[
R_t = \begin{cases} 
1 & \text{if there exists } k \text{ such that } t = t_k \\
0 & \text{otherwise}
\end{cases}
\] (1)

Then, \( R \) takes the value 1 at the change instants and is zero between two changes. Detecting the instants of changes consists in recovering the change process \( R \). When a particular realization \( z \) 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 $\theta$. When this distribution possesses a density $g(r/x, \theta)$ with respect to a given measure, the MAP estimate is given by:

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

where $h(x/r, \theta)$ is the density of the conditional 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 $r, \theta$ is the maximum likelihood estimate of $\theta$ in each stationary 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 arbitrarily defined. With no additional information, we could think in defining $R$ as a sequence of independent Bernoulli variables:

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

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

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

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

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 $\alpha$ controls the probabilities of detection errors. The smaller $\alpha$ is, the bigger the prior probability of a change is and the fewer the omissions are. On the other hand, the bigger $\alpha$ 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:

$$
X_k = f_t(k) + \varepsilon_k
$$

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

$$
U_x(r) = \sum_{i=1}^{N_r} \sum_{k=t_{i-1}+1}^{t_i} (X_k - \hat{f}_t(k))^2 + \beta N_r
$$

where $\beta = 2\sigma^2$ and $\hat{f}_t$ is the estimate of $f_t$.

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

$$
U_x(r) = \sum_{i=1}^{N_r} \sum_{k=t_{i-1}+1}^{t_i} (X_k - \hat{\mu}_t)^2 + \beta N_r
$$

Here, $\hat{\mu}_t$ is the empirical mean of $X$ on the $l$th segment. The parameter $\beta$ will define the resolution level. We propose in Figure 1 different segmentations of a same process, obtained with different values of $\beta$. 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 $\beta$.

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

$$
U_x(r) = \sum_{i=1}^{N_r} n_i \text{Log} \hat{\sigma}_t^2 + \beta N_r
$$

where $\hat{\sigma}_t^2$ is the estimated variance of $X$ on the $l$th segment and $n_i$ 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
\[ U_x(r) = \sum_{i=1}^{N_x} \sum_{(i,j) \in S \times S} n_i(ij) \log \frac{n_i(ij)}{n_i(i)} + \alpha N_x \tag{10} \]
where $n_i(i)$ is the number of times that $X$ takes the value $i$ in the $l$th segment and $n_i(ij)$ the number of times that $X$ passes from $i$ to $j$ in the $l$th 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
\[ U_x(r) = \sum_{i=1}^{N_x} \sum_{m=1}^{M} n_i(m) \log \frac{n_i(m)}{n_i(i)} + \alpha N_x \tag{11} \]
where $n_i(m)$ is the number of times that $Y$ takes the value $m$ in the $l$th segment and $n_i$ the length of the 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 $\alpha$ between 8 and 12. The changes are not significative enough to be detected with a value of $\alpha$ 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 parameter 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 $e^{-\Delta U}$ elsewhere. Here $\{T(k)\}$ is a decreasing sequence called temperature.

When the total energy is a sum of local potentials, $\Delta U$ 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 $k$:
\[ U_x(r,k) = U_x(r) + \lambda(k) \sum_{j=0}^{d} \sum_{i=1}^{N_x} (f^{(j)}(t_i) - f^{(j)}_{i+1}(t_i))^2 \]
where $f^{(j)}(t_i)$ is the $j$th derivative of $f$ at the $i$th knot $t_i$.

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

References

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.

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.