# An Introduction to process segmentation

by

Franck Picard


**Research Report No. 6**
**March 2007**

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
http://genome.jouy.inra.fr/ssb/

# An introduction to process segmentation

**Franck Picard**

Laboratoire Statistique et Génome
UMR CNRS 8071 - INRA 1152 - Univ. d'Evry, France

March 26, 2007

Many application fields in Statistics provide signals that are modelled through time series which are not stationary. Since the interpretation of such signals is complex, one aim is often to isolate zones in which the signal can be considered as stationary. In this context, the signal can be studied with parametric models for which the parameters are supposed to be affected by abrupt changes at unknown instants. The purpose of the statistical study is then to detect changes in these parameters. Quality control or monitoring has been one of the earliest applications of change detection. In this context, a production process is observed and must be controlled; the quick identification of disorders may be crucial for safety or quality control reasons. Dedicated statistical methods are based on the observation of sequential data, for which the detection of the change has to be done with the past observations as the only available information. The reader is referred to Basseville and Nikiforov (1993) for a complete review of *on-line* detection of abrupt changes, which does not constitute the purpose of our work. We are focused instead on the case where the analyst studies one global signal. In this case the change detection is done *off-line*, and the problem shifts to the global segmentation of the process.

## The multiple change-point problem

In the global segmentation context we aim at delimiting segments for which the characteristics of the signal are homogeneous within segments and different from one segment to another. We note $\{y_t\}_{t=1,\dots,n}$ the observed data which are modelled by a random process $\{Y_t\}_{t=1,\dots,n}$ that is supposed to be drawn from a probability distribution $f(\cdot)$ that depends on a parameter $\theta$.

*email:* picard@genopole.cnrs.fr

Then we assume that this parameter is affected by $K-1$ abrupt changes at unknown instants noted $t_1 < \ldots < t_{K-1}$, with the convention $t_0 = 1$ and $t_K = n$. The model is formulated as follows:

$$\forall t \in I_k, \quad Y_t \sim f(\theta_k),$$

with $I_k = \{t \in ]t_{k-1}, t_k]\}$ being the interval of size $n_k$ for which the parameter $\theta$ is constant and equals $\theta_k$. Many parameters can be affected by abrupt changes, the simplest ones being the mean and the covariance of the process, but changes can also affect the spectral distribution, or transition probabilities of Markov chains for instance.

From a statistical point of view, the problem of global segmentation gives rise to three main issues: (1) the determination of the parameter(s) affected by the change(s), (2) the estimation of the breakpoint instants, and the estimation of the parameters within segments, (3) the determination of the number of segments. The problem of determining which characteristics of the signal are affected by the changes may require a precise knowledge of the phenomenon under study. In the following, we will restrict the study to the case of changes in the mean only or in the mean and the variance of an independent Gaussian process. This model is detailed in section 1.

**Estimating the breakpoint coordinates**

Once the model has been specified, the problem is to estimate the location of the breakpoints and the parameters within segments. We will focus on two classical methods for this purpose: the maximum likelihood method and the least-squares method. For this estimation step, the number of segments has to be fixed. In the global segmentation setting, the estimation of the breakpoints can be viewed as a partitioning problem, where the purpose is to find the best partition of the data into $K$ segments. Since the number of possible partitions of the data into $K$ segments is $C_{n-1}^{K-1}$, the exploration of all possible partitions would be of order $\mathcal{O}(n^K)$. This computational problem explains why many segmentation methods only consider the detection of one change, compared to the multiple change-point problem. In section 2 we will explain how dynamic programming provides a solution to this problem of partitioning, and how the CART algorithm proposed by Breiman et al. (1984) can be used for the detection of multiple changes in the mean for large samples.

### Model selection

Once an estimation procedure is available for a fixed number of segments, the question of choosing this number remains. In practice this number is unknown and should be estimated. This problem can be viewed as a model selection issue. To date the number of segments is estimated with a penalized criterion:

$$crit(K) = J_K - \beta_n pen(K). \tag{1}$$

The first term $J_K$ measures the quality of fit of the model to the data. It can be the log-likelihood at its maximum noted $\log \widehat{\mathcal{L}}_K$, or minus the sum of squares of the model for instance. The second term is an increasing function of the number of segments, and is used to penalize the selection of an overly high-dimensional model. The term $\beta_n$ is a positive constant. This criterion establishes a trade-off between a good quality of fit and a reasonable number of segments. The definition of an appropriate penalty function and constant has focused much attention. In Section 3 we detail existing methods for model selection procedures in the multiple change-point context.

### The multiple change-point problem in the Bayesian setting

The last section will be devoted to a different approach which has been used to study multiple change-point problems, the Bayesian approach. In this context, the number of breakpoints and their location are viewed as random variables. The objective is to estimate their *posterior* distribution with MCMC methods. In this section we will compare two parametrizations which have been proposed by Green (1995) and Lavielle and Lebarbier (2001). Our objective is to explain the main differences between the two approaches and to draw analogies with the frequentist setting, when possible.

#### 1. Detection of changes in the mean of a Gaussian process

In this section we consider that the data are independent and drawn from a Gaussian distribution, such as

$$\forall t \in \{1, \ldots, n\}, \;\; Y_t \sim \mathcal{N}(\mu(t), \sigma(t)^2).$$

Then we assume that the mean and the variance of the process are affected by $K-1$ abrupt changes at unknown instants noted $t_1 < \ldots < t_{K-1}$. This model will be denoted $\mathcal{M}_1$, in contrast to model $\mathcal{M}_2$ where the only parameter

affected by the changes is the mean, with a constant variance $\sigma^2$. Then we have:

$$\forall t \in I_k \;\; Y_t \sim \begin{cases} \mathcal{N}(\mu_k, \sigma_k^2) & \text{model } \mathcal{M}_1, \\ \\ \mathcal{N}(\mu_k, \sigma^2) & \text{model } \mathcal{M}_2, \end{cases}$$

Since the data are independent, the log-likelihood of the model can be written as a sum of local log-likelihoods calculated on each individual segment, that is:

$$\log \mathcal{L}_K = \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} \log \left\{ \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(y_t - \mu_k)^2}{2\sigma^2} \right\}.$$

This additivity property will be central for the downstream estimation procedures that are based on maximum likelihood.

## 2. Estimation procedures when the number of segments is fixed

### 2.1 *The maximum likelihood method*

If the breakpoints are known, the estimators of the mean and the variance are the classical maximum likelihood estimators:

$$\begin{aligned} \hat{\mu}_k &= \frac{1}{n_k} \sum_{t=t_{k-1}+1}^{t_k} y_t, \\ \hat{\sigma}_k^2 &= \frac{1}{n_k} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \hat{\mu}_k)^2 \;\text{ for }\; \mathcal{M}_1, \\ \hat{\sigma}^2 &= \frac{1}{n} \sum_{k} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \hat{\mu}_k)^2 \;\text{ for }\; \mathcal{M}_2. \end{aligned}$$

For a model with $K$ segments the log-likelihood at its maximum is:

$$\begin{aligned} \log \widehat{\mathcal{L}}_K &= -\frac{n}{2} \log 2\pi - \frac{1}{2} \sum_k n_k \log \hat{\sigma}_k^2 \;\text{ for }\; \mathcal{M}_1, \\ \log \widehat{\mathcal{L}}_K &= -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}^2 \;\text{ for }\; \mathcal{M}_2. \end{aligned}$$

Nevertheless, the position of the breakpoints is unknown and should be estimated. This problem can be formulated as a partitioning problem whose aim is to find the best partition of the grid $\{1, \ldots, n\}$ into $K$ segments. If

we note $\mathcal{P}_K$ the set of all possible partitions of the grid $\{1,\dots,n\}$ into $K$ segments, the breakpoints are estimated as follows:

$$\hat{T}_K = \{\hat{t}_1,\dots,\hat{t}_{K-1}\} = \underset{T_K\in\mathcal{P}_K}{\text{Argmax}} \left\{\log \widehat{\mathcal{L}}_K(T_K)\right\}.$$

Dynamic programming is an efficient recursive approach that can be used to reduce the computational time of the exhaustive search.

### 2.2 *Dynamic programming and the shortest path problem to estimate the breakpoint instants*

Dynamic programming has been introduced by Bellman and Dreyfus (1962) and Auger and Lawrence (1989) were the first to use it in the context of global segmentation. It is a recursive approach based on the Bellman optimality principle (Bellman and Dreyfus (1962)). Let's consider model $\mathcal{M}_2$ with a constant variance. The quantity to be optimized is then:

$$J_K = \sum_{k=1}^{K} \sum_{t=t_{k-1}+1}^{t_k} (y_t - \hat{\mu}_k)^2.$$

The mean squares criterion is Broken down into a sum of minimum mean squares criteria. This break-down allows us to draw an analogy to the shortest path problem. Criterion $J_K$ can be seen as the total length of a path connecting point 1 to point $n$. The problem is then to find the shortest path connecting point 1 to point $n$ with $K-1$ steps, the steps being the breakpoint instants $t_1,\dots,t_{K-1}$.

Denoting $J_k(i,j)$ the cost (length) of the path connecting point $i$ and $j$ in $k$ steps, the algorithm is as follows:

$$\begin{aligned} \forall 0 \le i \le j,\ \ J_1(i,j) &= \sum_{t=i+1}^{j} (y_t - \bar{Y}_{ij})^2, \\ \forall 1 \le k \le K-1,\ \ J_{k+1}(1,j) &= \min_{1\le h\le j} \left\{J_k(1,h) + J_1(h+1,j)\right\}. \end{aligned}$$

In this context, the Bellman optimality principle is formulated as follows: "subpaths of optimal paths are themselves optimal". This global minimization property is crucial since it ensures the optimized criterion to be at its global maximum (compared with other estimation algorithms such as the EM algorithm that only ensures a local maximum). Moreover, this algorithm reduces the computational burden of the exhaustive search from $\mathcal{O}(n^K)$ to $\mathcal{O}(n^2)$ for a given $K$. This approach has been used by many authors and the reader is referred to Auger and Lawrence (1989), Braun et al. (2000) and Hawkins (2001) for instance.

2.3 *A CART-based approach for the multiple change-point problem*

Even if dynamic programming considerably reduces the computational time of the exhaustive search, it cannot be used for overly large samples. If the data to be partitioned are DNA sequences for instance, the storage of a cost matrix that is $n \times n$ with $n \sim 10^9$ is difficult. For this reason, Gey and Lebarbier (2002) recently proposed combining dynamic programming with a CART-based approach for the estimation of the breakpoints, when the size of the data is large. The role of the CART-based method for segmentation is to restrict the collection of visited partitions $\mathcal{P}_K$ to the relevant ones. This leads to a fast algorithm of order $\mathcal{O}(n \log n)$.

The CART algorithm is computed in two steps (Breiman et al. (1984)). The first one is called the growing procedure and consists in the recursive construction of a collection of partitions using data-dependent dyadic splitting. The computational schema of the first step is as follows:

- Compute the change-point $\hat{t}_c$ such as $\hat{t}_c = \underset{j}{\text{Argmin}} \{J_1(1, j) + J_1(j + 1, n)\}$.
  The objective of this step is to find the first best partition of $\{1, \ldots, n\}$ into 2 segments.

- Apply the same procedure on the new defined segments, and so on until the number of points within each resulting segment is smaller than a given threshold.

Other sequential methods have been proposed for the change-point estimation problem, see Ghorbanzdeh (1995), Picard (1985) and Chong (2001) for instance. Nevertheless, those methods aim at finding the relevant breakpoints directly, leading to sequential tests that require the definition of many tuning parameters. The use of a CART-based method is different. The first step (growing procedure) provides a collection of segmentations and the only parameter to be tuned is the minimum size for a segment to be split. In a second step (the pruning step), a relevant segmentation is chosen with a model selection procedure.

Once this segmentation has been chosen, it appears that some breakpoints can be irrelevant. This is due to the sequential nature of the CART algorithm that does not guarantee the finding of the global optimum. In order to circumvent this difficulty, Gey and Lebarbier (2002) propose combining the CART algorithm with a partial exhaustive search. The general idea is to consider that the breakpoints that have been proposed by CART

(at the end of the growing and pruning procedures) constitute candidates that can be removed if they correspond to false alarms. This is done by dynamic programming, which performs a partial exhaustive search on the proposed breakpoints to free the results from the hierarchic nature of the CART candidates. This leads to a hybrid algorithm that has been shown to be efficient (see Gey and Lebarbier (2002)).

2.4 *Statistical properties of the breakpoint estimators*

Once the position of the breakpoints has been estimated, a classical question is the statistical properties of the resulting estimators. Nevertheless, since the breakpoint parameters are discrete, the likelihood is not continuous with respect to those parameters. This particularity hampers the use of classical techniques to show their consistency for instance. Many articles have considered this problem, see Yao and Au (1989), Siegmund (1988), Lavielle (1999), Braun et al. (2000) for instance. Yao and Au (1989) have shown that in the case of a jump in the mean of an independent Gaussian process, the breakpoint estimators were consistent, and Braun et al. (2000) later show the consistency in the case of processes whose variance depends on the mean. Lavielle and Moulines (2000), Lavielle (1999) further extended those results to the case of time series and dependent processes, showing that the rate of convergence of $\hat{t}_k$ does not depend on the covariance structure of the process. In the case of a jump in the mean Yao and Au (1989) provide a theorem concerning the limiting distribution of the breakpoint estimators.

As for the confidence set of the change-point estimators, many strategies have been formulated for the single change-point problem. Siegmund (1988) and Worsley (1986) propose methods based on the likelihood ratio statistic, and Cobb (1978) provides an approximation of the conditional distribution of the maximum likelihood estimator of the change-point given the adjacent observations. In the multiple change-point context, current approaches use tests based on a change in the parameter of the distribution (see Avery and Henderson (1999) for a nonparametric approach in the case of Bernoulli sequence, Venter and Steel (1996) for maximum likelihood approaches in the Gaussian case). Those approaches focus on the change in the parameter with which the data are modelled, and not on the existence of a change-point $t_k$.

An interesting question would be to assess a simultaneous confidence region of the breakpoint estimators $\hat{t}_1, \ldots, \hat{t}_{K-1}$. To our knowledge no confidence set has yet been proposed for the sequence of the change-point estimators in the case of multiple breakpoints.

## 3. Model selection procedures to estimate the number of segments

Once the model has been specified and the location of the breakpoints can be estimated with an appropriate method, the problem is to determine the number of segments into which the data should be partitioned. In practice this number is unknown and can be estimated with a penalized criterion defined in Equation .

To date, two approaches have been considered to define the penalty term. The first one considers that there exists a true number of breakpoints $K^\star$ that should be estimated, and a true underlying model from which the data have been generated. In this context, Yao and Au (1989) showed that the Bayesian Information Criterion (BIC) provides a consistent estimator of the number of breakpoints. This criterion uses $J_K = \log \widehat{\mathcal{L}}_K$ and $pen(K) = 2K$ for the number of parameters to be estimated ($K$ means, 1 variance and $K-1$ breakpoints), and $\beta_n = 0.5 \times \log(n)$ for the penalty constant. This result is extended to the case of a dependent process, and Lavielle (1999) shows that if constant $\beta_n$ goes to 0 at an appropriate rate depending on the covariance structure of the process, the estimated number of change points converges to the true number.

Since practical use of penalized criteria is done in a non asymptotic context, another approach for model selection has been provided by Birg and Massart (2001). This model selection procedure has been applied to process segmentation by Lebarbier (2005) and Lavielle (2005), who propose two strategies that lead to different penalty functions and constants.

### 3.1 *Motivation of model selection*

In the context of model selection, we have $n$ independent random variables $\{Y_t\}_{t=1,\dots,n}$ whose distribution $s$ is unknown and has to be recovered. In the case of process segmentation, this function $s$ is recovered using a collection of piecewise constant functions. For this purpose, Lebarbier (2005) defines model $\mathcal{S}_m$ that is the subset of piecewise constant functions on partition $m = \{I_k\}_{k=1,\dots,K_m}$ of dimension $K_m$:

$$\mathcal{S}_m = \left\{ u = \sum_{k=1}^{K_m} u_k \mathbb{1}_{I_k}, (u_k)_{k=1,\dots,K_m} \in \mathbb{R}^{K_m} \right\}.$$

Classical estimation procedures consider that distribution $s$ belongs to

$\mathcal{S}_m$. Nevertheless, since $s$ is unknown, it is unlikely that it belongs to any model. The approach developed by Birg and Massart (2001) considers that model $\mathcal{S}_m$ only constitutes an approximation of $s$. Since $s$ is unknown, it is approximated by $\bar{s}_m$ that belongs to model $\mathcal{S}_m$. Nevertheless, $\bar{s}_m$ itself is unknown and is estimated by $\hat{s}_m$. Then the quality of an estimator $\hat{s}_m$ is assessed with a quadratic risk, $\mathbb{E}\|s - \hat{s}_m\|^2$, and the chosen estimator should minimize this risk. The quadratic risk of $\hat{s}_m$ can be broken down such that:

$$\mathbb{E}\|s - \hat{s}_m\|^2 = \mathbb{E}\|s - \bar{s}_m\|^2 + \mathbb{E}\|\bar{s}_m - \hat{s}_m\|^2.$$

The first term $\mathbb{E}\|s - \bar{s}_m\|^2$ measures the distance of the unknown $s$ to the approximator $\bar{s}_m$ in $\mathcal{S}_m$. This is a bias term that is small if the approximation is good. The second term $\mathbb{E}\|\bar{s}_m - \hat{s}_m\|^2$ measures the quality of the estimation of $\bar{s}_m$ by $\hat{s}_m$. This quantity should be small to prevent estimation errors. The purpose of model selection is then to establish a trade-off between a model that is close to the unknown distribution and which provides a good approximation of the unknown distribution, but that is not too big to prevent from estimation errors. This is called the bias/variance trade-off.

An ideal estimator of $s$, noted $\hat{s}_m$ could be defined as the estimator that achieves the best bias/variance trade-off. The objective of model selection is then to construct a criterion that will be used to select a partition $\hat{m}$ which behaves as well as the best estimator, up to some constant. This criterion is composed of two terms, a first term that quantifies the closeness of model $\mathcal{S}_m$ to the data, that increases with the dimension of the model, and a penalty term to control the estimation errors.

In the context of process segmentation, a model is selected through its dimension, *ie* we aim at selecting $\hat{m}$ the partition of dimension $K_m$. This is achieved with a penalty function defined by Lebarbier (2005), such that:

$$\beta_n \times pen(K) = \frac{K_m}{n}\sigma^2 \times \left\{ c_1 \log\left(\frac{n}{K_m}\right) + c_2 \right\}, \tag{2}$$

with $c_1, c_2$ two positive constants to be calibrated and $\sigma^2$ to be estimated. This function increases with the dimension of the model $K_m$, and the $\log(n/K_m)$ term reflects the richness of collection of partitions, since there exists $C_{n-1}^{K_m-1}$ possible partitions of the grid $\{1, \ldots, n\}$ into $K_m$ segments.

The performance of this penalty function has been assessed by simulation studies, and compared to other penalized criteria, such as the Mallows $C_p$ criterion, and the Bayesian Information Criterion (BIC) in a non asymptotic context. The main difference between those criteria is that criteria constructed on asymptotic considerations do not consider the complexity of the different models. Let us recall that the construction of BIC in the context of process segmentation considers that the number of parameters to be estimated is $K_m$ means, $K_m - 1$ breakpoints and 1 variance, whereas this new penalty considers that there exists $C_{n-1}^{K_m-1}$ possible partitions when $K_m$ is fixed. This leads to a penalization that is more stringent, and to the selection of a lower number of segments. Note that the construction of a penalty function is based on different objectives that will explain its behavior. For instance, the use of BIC to select the number of segment is motivated by the finding of the true number and of the true breakpoint coordinates. On the other hand, the penalty given by Lebarbier (2005) aims at minimizing a quadratic risk, and will tend to ignore some irrelevant breakpoints corresponding to small jumps in the mean.

To complete the introduction of model selection for segmentation process, the reader is referred to Lebarbier (2005) for further information concerning penalty 3.1, the calibration of constant $c_1, c_2$ and the estimation of $\sigma^2$. Model selection theory has been applied to a wide range of statistical problems. See Birg and Massart (2001) for a general presentation of model selection theory, Castellan (2000) for the application of model selection to the estimation of histograms, and Gey and Nedelec (2002) for model selection for CART Regression Trees.

### 3.2 *An adaptive method to estimate the number of segments*

In contrast to Lebarbier (2005) who aims at finding a universal penalty for selecting the number of segments, Lavielle (2005) has developed an adaptive method that is heuristically based. The motivation of such method is that the penalties defined for the BIC criterion or by Lebarbier (2005) are adapted to a very particular context. In the first one, the objective is to recover the true configuration, and the second one aims at minimizing a very specific criterion (the quadratic risk of the estimator), but none of these methods holds in the non-Gaussian case or for dependent variables for instance. The aim of Lavielle (2005) is to propose a method that can be used in many different situations, with very few hypotheses.

First of all let us notice that when the number of segments is small regarding the size of the data, penalty 3.1 is linear in the number of segments, and Lavielle (2005) suggests using a penalty in the form:

$$pen(K) = 2K.$$

The new objective is to estimate $\beta$ adaptively to the data. This estimation is done considering the behavior of the quality of fit criterion that is used. If this criterion is the least-squares criterion noted $J_K$, it will decrease as the number of segments increases, and the method consists in the determination of the number of segments for which the criterion ceases to decrease significantly. The proposition considers the slope between points $(K_i, J_{K_i})$ and $(K_{i+1}, J_{K_{i+1}})$. Looking where $J_K$ ceases to decrease significantly means looking for a break in the slope of this curve. An illustration is provided in Figure 1.

[Figure 1 about here.]

This method is heuristically based and requires the tuning of a parameter to assess the "significance" of the slope break. Nevertheless, it appears to be very flexible and has been shown to be efficient in many situations. Simulation results comparing this adaptive method to the penalty defined by Lebarbier (2005) show that it is more robust to the addition of noise (Picard et al. (2005)).

## 4. Bayesian formulation of the multiple change-point problem

In order to complete this review on segmentation methods, we present another modelling strategy that has been considered for this problem, in the Bayesian framework. See Green (1995), Carlin (1992), Barry and Hartigan (1993), Avery and Henderson (1999), Lavielle and Lebarbier (2001) for instance. Previous sections were dedicated to strategies whose objective is to provide the best segmentation on the data, based on a specific criterion. The objective is different in the Bayesian setting, where the number of segments as well as their position is random. As a consequence, their *posterior* distribution will be used to choose the most appropriate number of segments, and will provide local information regarding the position of the breakpoints.

The model can be specified as follows. Let $\{Y_t\}$ be a real process such that

$$\forall t \in \{1, \ldots, n\}, \;\; Y_t = s(t) + \varepsilon_t,$$

where $\varepsilon_t$ is a sequence of zero-mean random variables. The function $s$ to be recovered is supposed piecewise constant. With the conventional notations:

$$\forall t \in I_k, \;\; s(t) = \mu_k.$$

4.1 *The multiple change-point problem and the reversible jump algorithm*

Two approaches have been considered to model the sequence of change-points. Green (1995) specifies the *prior* model as follows. Suppose that the number of segments $K$ is drawn from a Poisson distribution $\mathcal{P}(\lambda)$. Given $K$, the breakpoint positions $t_1, \ldots, t_{K-1}$ are distributed as the even-numbered order statistics from $2K-1$ points uniformly distributed on $[1, n]$, and the means $\mu_k$ are independently drawn from the gamma density $\Gamma(\alpha, \beta)$.

A Monte Carlo Markov Chain algorithm is required to calculate the *posterior* probabilities of both breakpoint instants and means. Nevertheless, those probabilities depend on the number of segments which may vary. Many authors have solved this problem by fixing $K$ at 1. The development of the reversible jump MCMC sampler has allowed this limitation to be circumvented, and the multiple change-point problem was one of its first applications. The reader is referred to Green (1995) for further details on the application of the Reversible Jump algorithm to the multiple change-point problem.

4.2 *A reparametrization of the multiple change-point problem*

Instead of a parametrization that considers the breakpoint instants $\{t_k\}_k$, Lavielle (1998) and Lavielle and Lebarbier (2001) propose introducing a sequence of constant size $\{R_t\}$, such that:

$$R_t = \begin{cases} 1 & \text{if there exists } k \text{ such that } t = t_k. \\ 0 & \text{otherwise.} \end{cases}$$

The variables are supposed to be independent with *prior* Bernoulli distribution $\mathcal{B}(\lambda)$. Let us concentrate on the differences between the model specified by Green (1995) compared to this formulation.

In the formulation proposed by Lavielle and Lebarbier (2001) the variable of interest is the *presence* of a breakpoint, which is supposed independent from the presence of a breakpoint at close instants. In the framework defined by Green (1995) however the sequence of breakpoint instants $\{t_k\}$ indicates the *position* of the breakpoints, and positions are not independent from each other. As a consequence the *posterior* distribution of the $\{t_k\}$ in the reversible jump context will directly quantify the uncertainty regarding the

breakpoints location. With the reparametrization of the model, this information will be provided by the quantity $\Pr\{\sum_{t=t_a}^{t_b} R_t = k|y;\theta\}$, which is the probability of having exactly $k$ change-points between instants $t_a$ and $t_b$.

Another difference lies in the distribution of the number of segments. In the first case, this number is assumed to follow a Poisson distribution, and the distribution of the breakpoint instants only depends on its current value. In the formulation proposed by Lavielle and Lebarbier (2001), the *prior* distribution of the sequence $\{R_t\}$ defines the *prior* distribution of the number of segments. Since $K_R = \sum_{t=1}^{n-1} R_t + 1$, and $R_t \sim \mathcal{B}(\lambda)$, it follows that $K_R \sim \mathcal{B}(n-1,\lambda)$. The choice of the number of segments will depend on the choice of $\lambda$. More than a strict impact of parameter $\lambda$ on the distribution on the number of segments, the Bernoulli *prior* on $R_t$ specifies the distribution of the distance between two breakpoint instants since

$$\Pr\{R_{t+1} = 0, \ldots, R_{t+\ell-1} = 0, R_{t+\ell} = 1 | R_t = 1\} = \lambda(1-\lambda)^{\ell-1}.$$

In this formulation, the *prior* distribution has a double impact: it specifies the distribution of the number of segments, as well as the distribution of the length of the segments, which implicitly becomes geometric.

### 4.3 *Recovering the Maximum A Posteriori estimator of the breakpoints sequence*

The main advantage of the formulation proposed by Lavielle and Lebarbier (2001) lies in the computational approach that can be used to recover the *posterior* distribution of the sequence $\{R_t\}$. The authors emphasize the hierarchy of the model, that is:

$$p(R, \mu|y;\theta) = p(R|y;\theta) \times p(\mu|y, R;\theta),$$

with $\theta$ the set of hyperparameters. The first term $p(R|y;\theta)$ is used to recover the sequence of the breakpoint instants, and once this distribution is known, the signal is reconstructed with a Gibbs sampler to calculate $p(\mu|y, R;\theta)$, the hyperparameters of the model being estimated with a stochastic approximation of the EM algorithm, SAEM (Delyon et al. (1999)). Since the size of sequence $\{R_t\}$ is fixed, a Hastings-Metropolis algorithm can be used to sample sequences of 0 and 1 of size $n$. This parametrization prevents the use of a reversible jump algorithm, which is known to converge slowly.

Moreover Lavielle and Lebarbier (2001) show that the posterior distribution of $R$ is in the form

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

where

$$U_\theta(y,R) = \phi\sum_{k=1}^{K_R}\sum_{t_{k-1}+1}^{t_k}(y_t-\bar{y}_k)^2 + \gamma K_R,$$

and where $(\phi,\gamma)$ depend on the hyperparameters of the model. The Maximum A Posteriori (MAP) estimator of $R$ that minimizes $U_\theta(y,R)$ is then a penalized least-squares estimator. An analogy can be drawn with the breakpoint estimators defined in the frequentist context:

$$\{\hat{t}_1,\ldots,\hat{t}_{K-1}\} = \underset{t_1,\ldots,t_{K-1}}{\text{Argmin}}\left\{\frac{1}{n}\sum_{k=1}^{K}\sum_{t=t_{k-1}+1}^{t_k}(y_t-\hat{\mu}_k)^2 - 2\beta K\right\}.$$

The recovery of the MAP estimator of the breakpoint sequence can face local maxima that should be avoided. To do so, Lavielle and Lebarbier (2001) propose a modification of the Hastings-Metropolis algorithm, with the introduction of a temperature parameter $T$ such that:

$$p_T(R|y;\theta) = C_T(y;\theta)\exp\{-\frac{U_\theta(y,R)}{T}\}.$$

The interest in this temperature parameter is that when $T$ tends to 0, $p_T(\cdot|y;\theta)$ converges to the uniform distribution on the set of global maxima of $p(\cdot|y;\theta)$. Simulated annealing algorithms consist in using a sequence of temperatures $T^{(i)}$ that decrease at each iteration. Nevertheless, the use of this sequence would require a very large number of iterations. In practice, Lavielle and Lebarbier (2001) suggest running the Hastings-Metropolis algorithm at a fixed low temperature. The problem is to choose this temperature parameter.

## 5. Conclusion

In this work, we presented a brief review of existing statistical methods concerning the multiple change-point problem. Of course this review is not exhaustive, since the bibliography related to this subject is ample. Our scope was to present and explain the main tools that will be used in the following, such as dynamic programming and model selection, but also to present other existing methods, such as Bayesian methods that constitute an alternative modelling strategy.

## References


Auger, I. and Lawrence, C. (1989). Algorithms for the optimal identification of segments neighborhoods. *Bull. Math. Biol.* **51**, 39–54.

Avery, P. and Henderson, D. (1999). Detecting a changed segment is DNA sequences. *Appl.Statis.* **48**, 489–503.

Barry, D. and Hartigan, J. (1993). A Bayesian analysis for change-point problems. *JASA* **88**, 309–319.

Basseville, N. and Nikiforov, I. (1993). *Detection of abrupt changes. Theory and application*. Prentice Hall Information and system sciences series.

Bellman, R. and Dreyfus, S. (1962). *Applied dynamic programming*. Princeton University Press.

Birg, L. and Massart, P. (2001). Gaussian model selection. *J. European Math. Soc.* **3**, 203–268.

Braun, J. V., Braun, R. and Muller, H. (2000). Multiple change-point fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87**, 301–314.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall.

Carlin, B. (1992). Hierarchical Bayesian analysis of change-point problems. *Appl. Statist.* **41**, 389–405.

Castellan, G. (2000). Histogram selection with an Akaike type criterion. *C. R. Acad. Sci., Paris, Sr. I, Math.* **330**, 729–732.

Chong, T.-L. (2001). Estimating the locations and number of change-points by the sample splitting method. *Statist. Papers* **42**, 53–79.

Cobb, G. (1978). The problem of the Nile. Conditional solution to a change-point problem. *Biometrika* **65**, 243–251.

Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics* **27**, 94–28.

Gey, S. and Lebarbier, E. (2002). A CART based algorithm for detection of multiple change-points in the mean of large samples. Technical Report 10, Universit Paris Sud.

Gey, S. and Nedelec, E. (2002). Risk Bounds for CART Regression Trees. *MSRI Proceedings on Nonlinear Estimation and Classification* .

Ghorbanzdeh, D. (1995). Un test de dtection de rupture de la moyenne dans un modle gaussien. *Rev. Statist. Appl.* **43**, 67–76.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Hawkins, D. (2001). Fitting multiple change-point models to data. *Computational Statistics and data analysis* **37**, 323–341.

Lavielle, M. (1998). Optimal segmentation of random processes. *IEEE Transactions on signal processing* **46**, 1365–1373.

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stoch. Proc. and Appl.* **83**, 79–102.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing* **85**, 1501–1510.

Lavielle, M. and Lebarbier, E. (2001). An application of MCMC methods for the multiple change-points problem. *Signal Processing* **81**, 39–53.

Lavielle, M. and Moulines, E. (2000). Least squares estimation of an unknown number of shifts in a time series. *Journal of Time series analysis* **21**, 33–59.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing* **85**, 717–736.

Picard, D. (1985). Testing and estimating change-points in time series. *J. Applied Prob.* **17**, 841–867.

Picard, F., Robin, S., Lavielle, M., Vaisse, C. and Daudin, J.-J. (2005). A statistical approach for CGH microarray data analysis. *BMC Bioinformatics* **6**, 27.

Siegmund, D. (1988). Confidence sets in change-point problems. *International Statistical Review* **56**, 31–48.

Venter, J. and Steel, S. (1996). Finding multiple abrupt change-points. *Computational Statistic and data analysis* **22**, 481–504.

Worsley, K. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika* **73**, 91–104.

Yao, Y. and Au, S. (1989). Least square estimation of a step function. *Sankhya* **3**, 370–381.

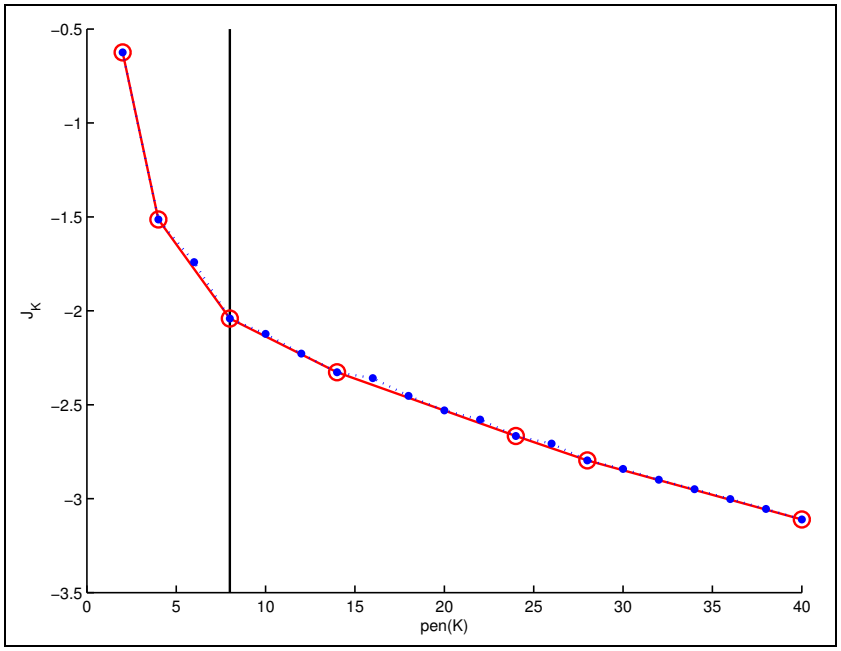


**Figure 1.** Illustration of the model selection procedure proposed by Lavielle (2005). Circles represent the convex hull of contrast $J_K$. The vertical line indicates the number of segments for which the contrast ceases to decrease significantly.