HAL open science

Segmentation in the mean of heteroscedastic data via resampling or cross-validation

Alain Celisse, Sylvain Arlot

To cite this version:

Alain Celisse, Sylvain Arlot. Segmentation in the mean of heteroscedastic data via resampling or cross-validation. Workshop Change-Point Detection Methods and Applications, Sep 2008, Paris, France. hal-02816806

HAL Id: hal-02816806

https://hal.inrae.fr/hal-02816806v1

Submitted on 6 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

Segmentation of the mean of heteroscedastic data via cross-validation

Sylvain Arlot and Alain Celisse

April 8, 2009

Abstract

This paper tackles the problem of detecting abrupt changes in the mean of a heteroscedastic signal by model selection, without knowledge on the variations of the noise. A new family of change-point detection procedures is proposed, showing that cross-validation methods can be successful in the heteroscedastic framework, whereas most existing procedures are not robust to heteroscedasticity. The robustness to heteroscedasticity of the proposed procedures is supported by an extensive simulation study, together with recent theoretical results. An application to Comparative Genomic Hybridization (CGH) data is provided, showing that robustness to heteroscedasticity can indeed be required for their analysis.

1 Introduction

The problem tackled in the paper is the detection of abrupt changes in the mean of a signal without assuming its variance is constant. Model selection and cross-validation techniques are used for building change-point detection procedures that significantly improve on existing procedures when the variance of the signal is not constant. Before detailing the approach and the main contributions of the paper, let us motivate the problem and briefly recall some related works in the change-point detection literature.

1.1 Change-point detection

The change-point detection problem, also called one-dimensional segmentation, deals with a stochastic process the distribution of which abruptly changes at some unknown instants. The purpose is to recover the location of these changes and their number. This problem is motivated by a wide range of applications, such as voice recognition, financial time-series analysis [29] and Comparative Genomic Hybridization (CGH) data analysis [35]. A large literature exists about change-point detection in many frameworks [see 12, 17, for a complete bibliography].

The first papers on change-point detection were devoted to the search for the location of a unique change-point, also named breakpoint [see 34, for instance]. Looking for multiple change-points is a harder task and has been studied later. For instance, Yao [49] used the BIC criterion for detecting multiple change-points in a Gaussian signal, and Miao and Zhao [33] proposed an approach relying on rank statistics.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

The setting of the paper is the following. The values $Y_1, \ldots, Y_n \in \mathbb{R}$ of a noisy signal at points $t_1, \ldots, t_n$ are observed, with

$$Y_i = s(t_i) + \sigma(t_i)\epsilon_i \ , \qquad \mathbb{E}\left[\epsilon_i\right] = 0 \quad \text{and} \quad \mathrm{Var}(\epsilon_i) = 1 \ . \tag{1}$$

The function s is called the *regression function* and is assumed to be piecewise-constant, or at least well approximated by piecewise constant functions, that is, s is smooth everywhere except at a few breakpoints. The noise terms $\epsilon_1, \ldots, \epsilon_n$ are assumed to be independent and identically distributed. No assumption is made on $\sigma : [0,1] \mapsto [0,\infty)$. Note that all data $(t_i, Y_i)_{1 \leq i \leq n}$ are observed before detecting the change-points, a setting which is called *off-line*.

As pointed out by Lavielle [28], multiple change-point detection procedures generally tackle one among the following three problems:

1. Detecting changes in the mean s assuming the standard-deviation σ is constant,
2. Detecting changes in the standard-deviation σ assuming the mean s is constant,
3. Detecting changes in the whole distribution of Y, with no distinction between changes in the mean s, changes in the standard-deviation σ and changes in the distribution of ϵ.

In applications such as CGH data analysis, changes in the mean s have an important biological meaning, since they correspond to the limits of amplified or deleted areas of chromosomes. However in the CGH setting, the standard-deviation σ is not always constant, as assumed in problem 1. See Section 6 for more details on CGH data, for which heteroscedasticity—that is, variations of σ— correspond to experimental artefacts or biological nuisance that should be removed.

Therefore, CGH data analysis requires to solve a fourth problem, which is the purpose of the present article:

4. Detecting changes in the mean s with no constraint on the standard-deviation $\sigma : [0,1] \mapsto [0,\infty)$.

Compared to problem 1, the difference is the presence of an additional nuisance parameter σ making problem 4 harder. Up to the best of our knowledge, no change-point detection procedure has ever been proposed for solving problem 4 with *no prior information on* σ.

1.2 Model selection

Model selection is a successful approach for multiple change-point detection, as shown by Lavielle [28] and by Lebarbier [30] for instance. Indeed, a set of change-points—called a segmentation—is naturally associated with the set of piecewise-constant functions that may only jump at these change-points. Given a set of functions (called a model), estimation can be performed by minimizing the least-squares criterion (or other criteria, see Section 3). Therefore, detecting changes in the mean of a signal, that is the choice of a segmentation, amounts to select such a model.

More precisely, given a collection of models $\{S_m\}_{m \in \mathcal{M}_n}$ and the associated collection of least-squares estimators $\{\widehat{s}_m\}_{m \in \mathcal{M}_n}$, the purpose of model selection is to provide a model index $\widehat{m}$ such that $\widehat{s}_{\widehat{m}}$ reaches the "best performance" among all estimators $\{\widehat{s}_m\}_{m \in \mathcal{M}_n}$.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

Model selection can target two different goals. On the one hand, a model selection procedure is *efficient* when its quadratic risk is smaller than the smallest quadratic risk of the estimators $\{\widehat{s}_m\}_{m\in\mathcal{M}_n}$, up to a constant factor $C_n \geq 1$. Such a property is called an *oracle inequality* when it holds for every finite sample size. The procedure is said to be *asymptotic efficient* when the previous property holds with $C_n \to 1$ as n tends to infinity. Asymptotic efficiency is the goal of AIC [2, 3] and Mallows' C_p [32], among many others.

On the other hand, assuming that s belongs to one of the models $\{S_m\}_{m\in\mathcal{M}_n}$, a procedure is *model consistent* when it chooses the smallest model containing s asymptotically with probability one. Model consistency is the goal of BIC [39] for instance. See also the article by Yang [46] about the distinction between efficiency and model consistency.

In the present paper as in [30], the quality of a multiple change-point detection procedure is assessed by the quadratic risk; hence, *a change in the mean hidden by the noise should not be detected*. This choice is motivated by applications where *the signal-to-noise ratio may be small*, so that exactly recovering every true change-point is hopeless. Therefore, *efficient* model selection procedures will be used in order to detect the change-points.

Without prior information on the locations of the change-points, the natural collection of models for change-point detection depends on the sample size n. Indeed, there exist $\binom{n-1}{D-1}$ different partitions of the n design points into D intervals, each partition corresponding to a set of $(D-1)$ change-points. Since D can take any value between 1 and n, 2^{n-1} models can be considered. Therefore, model selection procedures used for multiple change-point detection have to satisfy *non-asymptotic* oracle inequalities: the collection of models cannot be assumed to be fixed with the sample size n tending to infinity. (See Section 2.3 for a precise definition of the collection $\{S_m\}_{m\in\mathcal{M}_n}$ used for change-point detection.)

Most model selection results consider "polynomial" collections of models $\{S_m\}_{m\in\mathcal{M}_n}$, that is $\mathrm{Card}(\mathcal{M}_n) \leq Cn^\alpha$ for some constants $C, \alpha \geq 0$. For polynomial collections, procedures like AIC or Mallows' C_p are proved to satisfy oracle inequalities in various frameworks [9, 15, 10, 16], assuming that data are *homoscedastic*, that is, $\sigma(t_i)$ does not depend on t_i.

However as shown in [6], Mallows' C_p is suboptimal when data are *heteroscedastic*, that is the variance is non-constant. Therefore, other procedures must be used. For instance, resampling penalization is optimal with heteroscedastic data [5]. Another approach has been explored by Gendre [25], which consists in simultaneously estimating the mean and the variance, using a particular polynomial collection of models.

However in change-point detection, the collection of models is "exponential", that is $\mathrm{Card}(\mathcal{M}_n)$ is of order $\exp(\alpha n)$ for some $\alpha > 0$. For such large collections, especially larger than polynomial, the above penalization procedures fail. Indeed, Birgé and Massart [16] proved that the minimal amount of penalization required for a procedure to satisfy an oracle inequality is of the form

$$\mathrm{pen}(m) = c_1 \frac{\sigma^2 D_m}{n} + c_2 \frac{\sigma^2 D_m}{n} \log\left(\frac{n}{D_m}\right) \quad , \tag{2}$$

where c_1 and c_2 are positive constants and σ^2 is the variance of the noise, assumed to be constant. Lebarbier [30] proposed $c_1 = 5$ and $c_2 = 2$ for optimizing the penalty (2) in the context of change-point detection. Penalties similar to (2) have been introduced independently by other authors [38, 1, 11, 45] and are shown to provide satisfactory results.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

Nevertheless, all these results assume that data are homoscedastic. Actually, the model selection problem with heteroscedastic data and an exponential collection of models has never been considered in the literature, up to the best of our knowledge.

Furthermore, penalties of the form (2) are very close to be proportional to D_m, at least for small values of D_m. Therefore, the results of [6] lead to conjecture that the penalty (2) is suboptimal for model selection over an exponential collection of models, when data are heteroscedastic. The suggest of this paper is to use cross-validation methods instead.

1.3 Cross-validation

Cross-validation (CV) methods allow to estimate (almost) unbiasedly the quadratic risk of any estimator, such as $\widehat{s}_m$ (see Section 3.2 about the heuristics underlying CV). Classical examples of CV methods are the leave-one-out [Loo, 27, 43] and V-fold cross-validation [VFCV, 23, 24]. More references on cross-validation can be found in [7, 19] for instance.

CV can be used for model selection, by choosing the model S_m for which the CV estimate of the risk of $\widehat{s}_m$ is minimal. The properties of CV for model selection with a polynomial collection of models and homoscedastic data have been widely studied. In short, CV is known to adapt to a wide range of statistical settings, from density estimation [42, 20] to regression [44, 48] and classification [26, 47]. In particular, Loo is asymptotically equivalent to AIC or Mallows' C_p in several frameworks where they are asymptotically optimal, and other CV methods have similar performances, provided the size of the training sample is close enough to the sample size [see for instance 31, 40, 22]. In addition, CV methods are robust to heteroscedasticity of data [5, 7], as well as several other resampling methods [6]. Therefore, CV is a natural alternative to penalization procedures assuming homoscedasticity.

Nevertheless, nearly nothing is known about CV for model selection with an exponential collection of models, such as in the change-point detection setting. The literature on model selection and CV [14, 40, 16, 21] only suggests that minimizing directly the Loo estimate of the risk over 2^{n-1} models would lead to overfitting.

In this paper, a remark made by Birgé and Massart [16] about penalization procedure is used for solving this issue in the context of change-point detection. Model selection is perfomed in two steps: First, choose a segmentation given the number of change-points; second, choose the number of change-points. CV methods can be used at each step, leading to Procedure 6 (Section 5). The paper shows that such an approach is indeed successful for detecting changes in the mean of a heteroscedastic signal.

1.4 Contributions of the paper

The main purpose of the present work is to design a CV-based model selection procedure (Procedure 6) that can be used for detecting multiple changes in the mean of a heteroscedastic signal. Such a procedure experimentally adapts to heteroscedasticity when the collection of models is exponential, which has never been obtained before. In particular, Procedure 6 is a reliable alternative to Birgé and Massart's penalization procedure [15] when data can be heteroscedastic.

Another major difficulty tackled in this paper is the computational cost of resampling methods when selecting among 2^n models. Even when the number $(D-1)$ of change-

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

points is given, exploring the $\binom{n-1}{D-1}$ partitions of $[0,1]$ into D intervals and performing a resampling algorithm for each partition is not feasible when n is large and $D > 0$. An implementation of Procedure 6 with a tractable computational complexity is proposed in the paper, using closed-form formulas for Leave-p-out (Lpo) estimators of the risk, dynamic programming, and V-fold cross-validation.

The paper also points out that least-squares estimators are not reliable for change-point detection when the number of breakpoints is given, although they are widely used to this purpose in the literature. Indeed, experimental and theoretical results detailed in Section 3.1 show that least-squares estimators suffer from *local overfitting* when the variance of the signal is varying over the sequence of observations. On the contrary, minimizers of the Lpo estimator of the risk do not suffer from this drawback, which emphasizes the interest of using cross-validation methods in the context of change-point detection.

The paper is organized as follows. The statistical framework is described in Section 2. First, the problem of selecting the "best" segmentation given the number of change-points is tackled in Section 3. Theoretical results and an extensive simulation study show that the usual minimization of the least-squares criterion can be misleading when data are heteroscedastic, whereas cross-validation-based procedures provide satisfactory results in the same framework.

Then, the problem of choosing the number of breakpoints from data is addressed in Section 4. As supported by an extensive simulation study, V-fold cross-validation (VFCV) leads to a computationally feasible and statistically efficient model selection procedure when data are heteroscedastic, contrary to procedures implicitly assuming homoscedasticity.

The resampling methods of Sections 3 and 4 are combined in Section 5, leading to a family of resampling-based procedures for detecting changes in the mean of a heteroscedastic signal. A wide simulation study shows they perform well with both homoscedastic and heteroscedastic data, significantly improving the performance of procedures which implicitly assume homoscedasticity.

Finally, Section 6 illustrates on a real data set the promising behaviour of the proposed procedures for analyzing CGH microarray data, compared to procedures previously used in this setting.

2 Statistical framework

In this section, the statistical framework of change-point detection via model selection is introduced, as well as some notation.

2.1 Regression on a fixed design

Let $\mathcal{S}^*$ denote the set of measurable functions $[0,1] \mapsto \mathbb{R}$. Let $t_1 < \cdots < t_n \in [0,1]$ be some deterministic design points, $s \in \mathcal{S}^*$ and $\sigma : [0,1] \mapsto [0,\infty)$ be some functions and define

$$\forall i \in \{1,\ldots,n\}\,, \qquad Y_i = s(t_i) + \sigma(t_i)\epsilon_i \ , \tag{3}$$

where $\epsilon_1,\ldots,\epsilon_n$ are independent and identically distributed random variables with $\mathbb{E}\left[\epsilon_i\right] = 0$ and $\mathbb{E}\left[\epsilon_i^2\right] = 1$.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

As explained in Section 1.1, the goal is to find from $(t_i, Y_i)_{1\leq i\leq n}$ a piecewise-constant function $f \in \mathcal{S}^*$ close to s in terms of the quadratic loss

$$\|s - f\|_n^2 := \frac{1}{n}\sum_{i=1}^{n} (f(t_i) - s(t_i))^2 \quad .$$

2.2 Least-squares estimator

A classical estimator of s is the *least-squares estimator*, defined as follows. For every $f \in \mathcal{S}^*$, the least-squares criterion at f is defined by

$$P_n\gamma(f) := \frac{1}{n}\sum_{i=1}^{n} (Y_i - f(t_i))^2 \quad .$$

The notation $P_n\gamma(f)$ means that the function $(t, Y) \mapsto \gamma(f; (t, Y)) := (Y - f(t))^2$ is integrated with respect to the empirical distribution $P_n := n^{-1}\sum_{i=1}^{n} \delta_{(t_i,Y_i)}$. $P_n\gamma(f)$ is also called the *empirical risk* of f.

Then, given a set $S \subset \mathcal{S}^*$ of functions $[0, 1] \mapsto \mathbb{R}$ (called a *model*), the least-squares estimator on model S is

$$\mathrm{ERM}(S; P_n) := \arg\min_{f\in S} \{P_n\gamma(f)\} \quad .$$

The notation $\mathrm{ERM}(S; P_n)$ stresses that the least-squares estimator is the output of the empirical risk minimization algorithm over S, which takes a model S and a data sample as inputs. When a collection of models $\{S_m\}_{m\in\mathcal{M}_n}$ is given, $\widehat{s}_m(P_n)$ or $\widehat{s}_m$ are shortcuts for $\mathrm{ERM}(S_m; P_n)$.

2.3 Collection of models

Since the goal is to detect jumps of s, every model considered in this article is the set of piecewise constant functions with respect to some partition of $[0, 1]$.

For every $K \in \{1, \ldots, n-1\}$ and every sequence of integers $\alpha_0 = 1 < \alpha_1 < \alpha_2 < \cdots < \alpha_K \leq n$ (the breakpoints), $(I_\lambda)_{\lambda\in\Lambda_{(\alpha_1,\ldots\alpha_K)}}$ denotes the partition

$$[t_{\alpha_0}; t_{\alpha_1}),\ \ldots\ ,\ [t_{\alpha_{K-1}}; t_{\alpha_K}),\ [t_{\alpha_K}; 1]$$

of $[0, 1]$ into $(K+1)$ intervals. Then, the model $S_{(\alpha_1,\ldots\alpha_K)}$ is defined as the set of piecewise constant functions that can only jump at $t = t_{\alpha_j}$ for some $j \in \{1, \ldots, K\}$.

For every $K \in \{1, \ldots, n-1\}$, let $\widetilde{\mathcal{M}}_n(K+1)$ denote the set of such sequences $(\alpha_1, \ldots \alpha_K)$ of length K, so that $\{S_m\}_{m\in\widetilde{\mathcal{M}}_n(K+1)}$ is the collection of models of piecewise constant functions with K breakpoints. When $K = 0$, $\widetilde{\mathcal{M}}_n(1) := \{\emptyset\}$ and the model $S_\emptyset$ is the linear space of constant functions on $[0, 1]$. Remark that for every K and $m \in \widetilde{\mathcal{M}}_n(K+1)$, S_m is a vector space of dimension $D_m = K+1$. In the rest of the paper, the relationship between the number of breakpoints K and the dimension $D = K+1$ of the model $S_{(\alpha_1,\ldots\alpha_K)}$ is used repeatedly; in particular, estimating of the number of breakpoints (Section 4) is equivalent to choosing the dimension of a model. In addition, since a model S_m is uniquely defined by m, the index m is also called a model.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

The classical collection of models for change-point detection can now be defined as $\{S_m\}_{m\in\widetilde{\mathcal{M}}_n}$, where $\widetilde{\mathcal{M}}_n = \bigcup_{D\in\mathcal{D}_n}\widetilde{\mathcal{M}}_n(D)$ and $\mathcal{D}_n = \{1,\dots,n\}$. This collection has a cardinality 2^{n-1}.

In this paper, a slightly smaller collection of models is considered, that is, all $m \in \widetilde{\mathcal{M}}_n$ such that each element of the partition $(I_\lambda)_{\lambda\in\Lambda_m}$ contains at least two design points $(t_j)_{1\le j\le n}$. Indeed, when nothing is known about the noise-level $\sigma(\cdot)$, one cannot hope to distinguish two consecutive change-points from a local variation of σ. For every $D \in \{1,\dots,n\}$, let $\mathcal{M}_n(D)$ denote the set of $m \in \widetilde{\mathcal{M}}_n(D)$ satisfying this property. Then, the collection of models used in this paper is defined as $\{S_m\}_{m\in\mathcal{M}_n}$ where $\mathcal{M}_n = \bigcup_{D\in\mathcal{D}_n}\mathcal{M}_n(D)$ and $\mathcal{D}_n \subset \{1,\dots,n/2\}$. Finally, in all the experiments of the paper, $\mathcal{D}_n = \{1,\dots,4n/10\}$ for reasons detailed in Section 4.2, in particular Remark 3.

2.4 Model selection

Among $\{S_m\}_{m\in\mathcal{M}_n}$, the best model is defined as the minimizer of the *quadratic loss* $\|s-\widehat{s}_m\|_n^2$ over $m \in \mathcal{M}_n$ and called the *oracle* $m^\star$. Since the oracle depends on s, one can only expect to select $\widehat{m}(P_n)$ from the data such that the quadratic loss of $\widehat{s}_{\widehat{m}}$ is close to that of the oracle with high probability, that is,

$$\|s-\widehat{s}_{\widehat{m}}\|_n^2 \le C \inf_{m\in\mathcal{M}_n}\left\{\|s-\widehat{s}_m\|_n^2\right\} + R_n \tag{4}$$

where C is close to 1 and R_n is a small remainder term (typically of order n^{-1}). Inequality (4) is called an *oracle inequality*.

3 Localization of the breakpoints

A usual strategy for multiple change-point detection [28, 30] is to dissociate the search for the best segmentation given the number of breakpoints from the choice of the number of breakpoints.

In this section, the number $K = D-1$ of breakpoints is fixed and the goal is to localize them. In other words, the goal is to select a model among $\{S_m\}_{m\in\mathcal{M}_n(D)}$.

3.1 Empirical risk minimization's failure with heteroscedastic data

As explained by many authors such as Lavielle [28], minimizing the least-squares criterion over $\{\widehat{s}_m\}_{m\in\mathcal{M}(D)}$ is a classical way of estimating the best segmentation with $(D-1)$ change-points. This leads to the following procedure:

Procedure 1.

$$\widehat{m}_{\mathrm{ERM}}(D) := \arg\min_{m\in\mathcal{M}_n(D)}\{P_n\gamma(\widehat{s}_m)\} = \mathrm{ERM}\left(\widetilde{S}_D; P_n\right) \quad ,$$
$$\text{where} \qquad \widetilde{S}_D := \cup_{m\in\mathcal{M}_n(D)} S_m$$

is the set of piecewise constant functions with exactly $(D-1)$ change-points, chosen among $t_2,\dots,t_n$ (see Section 2.3).

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript

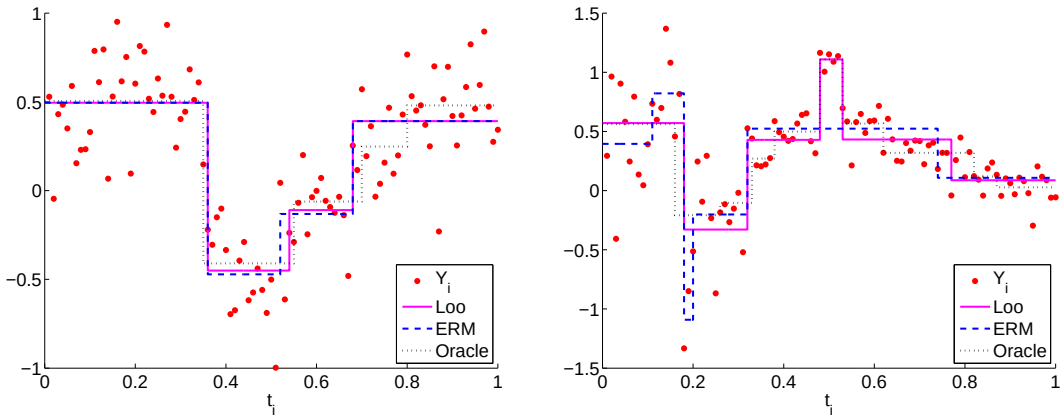


Figure 1: Comparison of $\widehat{s}_{m^\star(D)}$ (dotted black line), $\widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)}$ (dashed blue line) and $\widehat{s}_{\widehat{m}_{\mathrm{Loo}}(D)}$ (plain magenta line, see Section 3.2.2), D being the "optimal" dimension (see Figure 3). Data are generated as described in Section 3.3.1 with $n = 100$ data points. Left: homoscedastic data (s_2, σ_c), $D = 4$. Right: heteroscedastic data $(s_3, \sigma_{pc,3})$, $D = 6$.

Remark 1. Dynamic programming [13] leads to an efficient implementation of Procedure 1 with computational complexity $\mathcal{O}\left(n^2\right)$.

Among models corresponding to segmentations with $(D-1)$ change-points, the oracle model can be defined as

$$m^\star(D) := \arg \min_{m \in \mathcal{M}_n(D)} \left\{ \|s - \widehat{s}_m\|_n^2 \right\} \quad .$$

Figure 1 illustrates how far $\widehat{m}_{\mathrm{ERM}}(D)$ typically is from $m^\star(D)$ according to variations of the standard-deviation σ. On the one hand, when data are homoscedastic, empirical risk minimization yields a segmentation close to the oracle (Figure 1, left). On the other hand, when data are heteroscedastic, empirical risk minimization introduces artificial breakpoints in areas where the noise-level is above average, and misses breakpoints in areas where the noise-level is below average (Figure 1, right). In other words, when data are heteroscedastic, *empirical risk minimization over $\widetilde{S}_D$ locally overfits in high-noise areas*, and locally underfits in low-noise areas.

The failure of empirical risk minimization with heteroscedastic data observed on Figure 1 is general [21, Chapter 7] and can be explained by Lemma 1 below. Indeed, the criteria $P_n\gamma(\widehat{s}_m)$ and $\|s - \widehat{s}_m\|_n^2$, respectively minimized by $\widehat{m}_{\mathrm{ERM}}(D)$ and $m^\star(D)$ over $\mathcal{M}_n(D)$, are close to their respective expectations, as proved by the concentration inequalities of [7, Proposition 9] for instance. Lemma 1 enables to compare these expectations.

Lemma 1. *Let $m \in \mathcal{M}_n$ and define $s_m := \arg\min_{f \in S_m} \|s - f\|_n^2$. Then,*

$$\mathbb{E}\left[P_n\gamma\left(\widehat{s}_m\right)\right] = \|s - s_m\|_n^2 - V(m) + \frac{1}{n}\sum_{i=1}^n \sigma(t_i)^2 \tag{5}$$

$$\mathbb{E}\left[\|s - \widehat{s}_m\|_n^2\right] = \|s - s_m\|_n^2 + V(m) \tag{6}$$

where

$$V(m) := \frac{\sum_{\lambda \in \Lambda_m} \left(\sigma_\lambda^r\right)^2}{n} \text{ and } \forall \lambda \in \Lambda_m, \quad \left(\sigma_\lambda^r\right)^2 := \frac{\sum_{i=1}^n \sigma(t_i)^2 \mathbb{1}_{t_i \in I_\lambda}}{\mathrm{Card}\left(\{k \mid t_k \in I_\lambda\}\right)} \quad . \tag{7}$$

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

Lemma 1 is proved in [21]. As it is well-known in the model selection literature, the expectation of the quadratic loss (6) is the sum of two terms: $\|s - s_m\|_n^2$ is the bias of model S_m, and $V(m)$ is a variance term, measuring the difficulty of estimating the D_m parameters of model S_m. Up to the term $n^{-1}\sum_{i=1}^n \sigma(t_i)^2$ which does not depend on m, the empirical risk underestimates the quadratic risk (that is, the expectation of the quadratic loss), as shown by (5), because of the sign in front of $V(m)$.

Nevertheless, when data are homoscedastic, that is when $\forall i$, $\sigma(t_i) = \overline{\sigma}$, $V(m) = D_m \overline{\sigma}^2 n^{-1}$ is the same for all $m \in \mathcal{M}_n(D)$. Therefore, (5) and (6) show that for every $D \geq 1$, when data are homoscedastic

$$\arg\min_{m \in \mathcal{M}_n(D)} \left\{ \mathbb{E}\left[P_n \gamma\left(\widehat{s}_m\right) \right] \right\} = \arg\min_{m \in \mathcal{M}_n(D)} \left\{ \mathbb{E}\left[\|s - \widehat{s}_m\|_n^2 \right] \right\} \ .$$

Hence, $\widehat{m}_{\mathrm{ERM}}(D)$ and $m^\star(D)$ tend to be close to one another, as on the left of Figure 1.

On the contrary, when data are heteroscedastic, the variance term $V(m)$ can be quite different among models $m \in \mathcal{M}_n(D)$, even though they have the same dimension D. Indeed, $V(m)$ increases when a breakpoint is moved from an area where σ is small to an area where σ is large. Therefore, the empirical risk minimization algorithm rather puts breakpoints in noisy areas in order to minimize $-V(m)$ in (5). This is illustrated in the right panel of Figure 1, where the oracle segmentation $m^\star(D)$ has more breakpoints in areas where σ is small.

3.2 Cross-validation

Cross-validation (CV) methods are natural candidates for fixing the failure of empirical risk minimization when data are heteroscedastic, since CV methods are naturally adaptive to heteroscedasticity (see Section 1.3). The purpose of this section is to properly define how CV can be used for selecting $\widehat{m} \in \mathcal{M}_n(D)$ (Procedure 2), and to recall theoretical results showing why this procedure adapts to heteroscedasticity (Proposition 1).

3.2.1 Heuristics

The cross-validation heuristics [4, 43] relies on a data splitting idea: For each candidate algorithm—say $\mathrm{ERM}(S_m; \cdot)$ for some $m \in \mathcal{M}_n(D)$—, part of the data—called training set—is used for training the algorithm. The remaining part—called validation set—is used for estimating the risk of the algorithm. This simple strategy is called *validation* or *hold-out*. One can also split data several times and average the estimated values of the risk over the splits. Such a strategy is called *cross-validation* (CV). CV with general repeated splits of data has been introduced by Geisser [23, 24].

In the fixed-design setting, $(t_i, Y_i)_{1 \leq i \leq n}$ are not identically distributed so that CV estimates a quantity slightly different from the usual prediction error. Let T be uniformly distributed over $\{t_1, \ldots, t_n\}$ and $Y = s(T) + \sigma(T)\epsilon$, where ϵ is independent from $\epsilon_1, \ldots, \epsilon_n$

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

with the same distribution. Then, the CV estimator of the risk of $\widehat{s}(P_n)$ estimates

$$\mathbb{E}_{(T,Y)}\left[(\widehat{s}(T)-Y)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\epsilon}\left[(s(t_i)+\sigma(t_i)\epsilon_i-\widehat{s}(t_i))^2\right]$$
$$= \|s-\widehat{s}\|_n^2 + \frac{1}{n}\sum_{i=1}^{n}\sigma(t_i)^2 \ .$$

Hence, minimizing the CV estimator of $\mathbb{E}_{(T,Y)}\left[(\widehat{s}_m(T)-Y)^2\right]$ over m amounts to minimize $\|s-\widehat{s}_m\|_n^2$, up to estimation errors.

Even though the use of CV in a fixed-design setting is not usual, theoretical results detailed in Section 3.2.4 below show that CV actually leads to a good estimator of the quadratic risk $\|s-\widehat{s}_m\|_n^2$. This fact is confirmed by all the experimental results of the paper.

3.2.2 Definition

Let us now formally define how CV is used for selecting some $m \in \mathcal{M}_n(D)$ from data. A (statistical) algorithm $\mathcal{A}$ is defined as any measurable function $P_n \mapsto \mathcal{A}(P_n) \in \mathcal{S}^*$. For any $t_i \in [0,1]$, $\mathcal{A}(t_i;P_n)$ denotes the value of $\mathcal{A}(P_n)$ at point t_i.

For any $I^{(t)} \subset \{1,\ldots,n\}$, define $I^{(v)} := \{1,\ldots,n\}\setminus I^{(t)}$,

$$P_n^{(t)} := \frac{1}{\mathrm{Card}(I^{(t)})}\sum_{i\in I^{(t)}}\delta_{(t_i,Y_i)} \quad \text{and} \quad P_n^{(v)} := \frac{1}{\mathrm{Card}(I^{(v)})}\sum_{i\in I^{(v)}}\delta_{(t_i,Y_i)} \ .$$

Then, the *hold-out estimator of the risk* of any algorithm $\mathcal{A}$ is defined as

$$\widehat{R}_{\mathrm{ho}}(\mathcal{A},P_n,I^{(t)}) := P_n^{(v)}\gamma\left(\mathcal{A}\left(P_n^{(t)}\right)\right) = \frac{1}{\mathrm{Card}(I^{(v)})}\sum_{i\in I^{(v)}}\left(\mathcal{A}(t_i;P_n^{(t)})-Y_i\right)^2 \ .$$

The *cross-validation estimators of the risk* of $\mathcal{A}$ are then defined as the average of $\widehat{R}_{\mathrm{ho}}(\mathcal{A},P_n,I_j^{(t)})$ over $j=1,\ldots,B$ where $I_1^{(t)},\ldots,I_B^{(t)}$ are chosen in a predetermined way [24]. Leave-one-out, leave-p-out and V-fold cross-validation are among the most classical examples of CV procedures. They differ one another by the choice of $I_1^{(t)},\ldots,I_B^{(t)}$.

- *Leave-one-out* (Loo), often called *ordinary CV* [4, 43], consists in training with the whole sample except one point, used for testing, and repeating this for each data point: $I_j^{(t)} = \{1,\ldots,n\}\setminus\{j\}$ for $j=1,\ldots,n$. The Loo estimator of the risk of $\mathcal{A}$ is defined by
$$\widehat{R}_{\mathrm{Loo}}(\mathcal{A},P_n) := \frac{1}{n}\sum_{j=1}^{n}\left[\left(Y_j-\mathcal{A}\left(t_j;P_n^{(-j)}\right)\right)^2\right] \ ,$$
where $P_n^{(-j)} = (n-1)^{-1}\sum_{i,\,i\neq j}\delta_{(t_i,Y_i)}$.
- *Leave-p-out* (Lpo_p, with any $p\in\{1,\ldots,n-1\}$) generalizes Loo. Let $\mathcal{E}_p$ denote the collection of all possible subsets of $\{1,\ldots,n\}$ with cardinality $n-p$. Then, Lpo

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

consists in considering every $I^{(t)} \in \mathcal{E}_p$ as training set indices:

$$\widehat{R}_{\mathrm{Lpo}_p}(\mathcal{A}, P_n) := \binom{n}{p}^{-1} \sum_{I^{(t)} \in \mathcal{E}_p} \left[\frac{1}{p} \sum_{j \in I^{(v)}} \left[\left(Y_j - \mathcal{A}\left(t_j; P_n^{(t)}\right) \right)^2 \right] \right] \quad . \tag{8}$$

- *V-fold cross-validation* (VFCV) is a computationally efficient alternative to Lpo and Loo. The idea is to first partition the data into V blocks, to use all the data but one block as a training sample, and to repeat the process V times. In other words, VFCV is a blockwise Loo, so that its computational complexity is V times that of $\mathcal{A}$. Formally, let $B_1, \ldots, B_V$ be a partition of $\{1, \ldots, n\}$ and $P_n^{(\overline{B_k})} := (n - \mathrm{Card}(B_k))^{-1} \sum_{i \notin B_k} \delta_{(t_i, Y_i)}$ for every $k \in \{1, \ldots, V\}$. The VFCV estimator of the risk of $\mathcal{A}$ is defined by

$$\widehat{R}_{\mathrm{VF}_V}(\mathcal{A}, P_n) := \frac{1}{V} \sum_{k=1}^{V} \left[\frac{1}{\mathrm{Card}(B_k)} \sum_{j \in B_k} \left[\left(Y_j - \mathcal{A}\left(t_j; P_n^{(\overline{B_k})}\right) \right)^2 \right] \right] \quad . \tag{9}$$

  The interested reader will find theoretical and experimental results on VFCV and the best way to use it in [7, 21] and references therein, in particular [18].

Given the Loo estimator of the risk of each algorithm $\mathcal{A}$ among $\{\mathrm{ERM}(S_m; \cdot)\}_{m \in \mathcal{M}_n(D)}$, the segmentation with $(D-1)$ breakpoints chosen by Loo is defined as follows.

Procedure 2.

$$\widehat{m}_{\mathrm{Loo}}(D) := \arg \min_{m \in \mathcal{M}_n(D)} \left\{ \widehat{R}_{\mathrm{Loo}}\left(\mathrm{ERM}\left(S_m; \cdot\right), P_n\right) \right\} \quad .$$

The segmentations chosen by Lpo and VFCV are defined similarly and denoted respectively by $\widehat{m}_{\mathrm{Lpo}_p}(D)$ and by $\widehat{m}_{\mathrm{VF}_V}(D)$.

As illustrated by Figure 1, when data are heteroscedastic, $\widehat{m}_{\mathrm{Loo}}(D)$ is often closer to the oracle segmentation $m^\star(D)$ than $\widehat{m}_{\mathrm{ERM}}(D)$. This improvement will be explained by theoretical results in Section 3.2.4 below.

3.2.3 Computational tractability

The computational complexity of $\mathrm{ERM}(S_m; P_n)$ is $\mathcal{O}(n)$ since for every $\lambda \in \Lambda_m$, the value of $\widehat{s}_m(P_n)$ on I_λ is equal to the mean of $\{Y_i\}_{t_i \in I_\lambda}$. Therefore, a naive implementation of Lpo_p has a computational complexity $\mathcal{O}\left(n\binom{n}{p}\right)$, which can be intractable for large n in the context of model selection, even when $p = 1$. In such cases, only VFCV with a small V would work straightforwardly, since its computational complexity is $\mathcal{O}(nV)$.

Nevertheless, closed-form formulas for the Lpo estimator of the risk have been derived in the density estimation [20, 19] and regression [21] frameworks. Some of these closed-form formulas apply to regressograms $\widehat{s}_m$ with $m \in \mathcal{M}_n$. The following theorem gives a closed-form expression for $\widehat{R}_{\mathrm{Lpo}_p}(m) := \widehat{R}_{\mathrm{Lpo}_p}(\mathrm{ERM}(S_m; \cdot), P_n)$ which can be computed with $\mathcal{O}(n)$ elementary operations.

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

Theorem 1 (Corollary 3.3.2 in [21]). *Let $m \in \mathcal{M}_n$, S_m and $\widehat{s}_m = \mathrm{ERM}(S_m;\cdot)$ be defined as in Section 2. For every $(t_1,Y_1),\ldots,(t_n,Y_n) \in \mathbb{R}^2$ and $\lambda \in \Lambda_m$, define*

$$S_{\lambda,1} := \sum_{j=1}^{n} Y_j \mathbb{1}_{\{t_j \in I_\lambda\}} \quad \text{and} \quad S_{\lambda,2} := \sum_{j=1}^{n} Y_j^2 \mathbb{1}_{\{t_j \in I_\lambda\}} \ .$$

Then, for every $p \in \{1,\ldots,n-1\}$, the Lpo_p estimator of the risk of $\widehat{s}_m$ defined by (8) is given by

$$\widehat{R}_{\mathrm{Lpo}_p}(m) = \sum_{\lambda \in \Lambda_m} \frac{1}{pN_\lambda} \left[\left\{ (A_\lambda - B_\lambda) S_{\lambda,2} + B_\lambda S_{\lambda,1}^2 \right\} \mathbb{1}_{\{n_\lambda \geq 2\}} + \{+\infty\} \mathbb{1}_{\{n_\lambda = 1\}} \right] \ ,$$

where for every $\lambda \in \Lambda_m$,

$$n_\lambda := \mathrm{Card}\left(\{i \mid t_i \in I_\lambda\}\right) \qquad N_\lambda := 1 - \mathbb{1}_{\{p \geq n_\lambda\}} \binom{n-n_\lambda}{p-n_\lambda} \Big/ \binom{n}{p}$$

$$A_\lambda := V_\lambda(0)\left(1 - \frac{1}{n_\lambda}\right) - \frac{V_\lambda(1)}{n_\lambda} + V_\lambda(-1)$$

$$B_\lambda := V_\lambda(1)\frac{2 - \mathbb{1}_{n_\lambda \geq 3}}{n_\lambda(n_\lambda - 1)} + \frac{V_\lambda(0)}{n_\lambda - 1}\left[\left(1 + \frac{1}{n_\lambda}\right)\mathbb{1}_{n_\lambda \geq 3} - 2\right] - \frac{V_\lambda(-1)\mathbb{1}_{n_\lambda \geq 3}}{n_\lambda - 1}$$

$$\text{and} \quad \forall k \in \{-1,0,1\}\,, \quad V_\lambda(k) := \sum_{r=\max\{1,(p-n_\lambda)\}}^{\min\{n_\lambda,(n-p)\}} r^k \frac{\binom{n-p}{r}\binom{p}{n_\lambda - r}}{\binom{n}{n_\lambda}} \ .$$

Remark 2. $V_\lambda(k)$ can also be written as $\mathbb{E}\left[Z^k \mathbb{1}_{Z>0}\right]$ where Z has hypergeometric distribution with parameters $(n, n-p, n_\lambda)$.

An important practical consequence of Theorem 1 is that for every D and p, $\widehat{m}_{\mathrm{Lpo}_p}(D)$ can be computed with the same computational complexity as $\widehat{m}_{\mathrm{ERM}}(D)$, that is $\mathcal{O}\left(n^2\right)$. Indeed, Theorem 1 shows that $\widehat{R}_{\mathrm{Lpo}_p}(m)$ is a sum over $\lambda \in \Lambda_m$ of terms depending only on $\{Y_i\}_{t_i \in I_\lambda}$, so that dynamic programming [13] can be used for computing the minimizer $\widehat{m}_{\mathrm{Lpo}_p}(D)$ of $\widehat{R}_{\mathrm{Lpo}_p}(m)$ over $m \in \mathcal{M}_n$. Therefore, Lpo *and* Loo *are computationally tractable for change-point detection* when the number of breakpoints is given.

Dynamic programming also applies to $\widehat{m}_{\mathrm{VF}_V}$ with a computational complexity $\mathcal{O}\left(Vn^2\right)$, since each term appearing in $\widehat{R}_{\mathrm{VF}_V}(m)$ is the average over V quantities that must be computed, except when $V = n$ since VFCV then becomes Loo. Since VFCV is mostly an approximation to Loo or Lpo but has a larger computational complexity, $\widehat{m}_{\mathrm{Lpo}_p}$ will be preferred to $\widehat{m}_{\mathrm{VF}_V}(D)$ in the following.

3.2.4 Theoretical guarantees

In order to understand why CV indeed works for change-point detection with a given number of breakpoints, let us recall a straightforward consequence of Theorem 1 which is proved in details in [21, Lemma 7.2.1 and Proposition 7.2.3].

Proposition 1. *Using the notation of Lemma 1, for any $m \in \mathcal{M}_n$,*

$$\mathbb{E}\left[\widehat{R}_{\mathrm{Lpo}_p}(m)\right] \approx \|s - s_m\|_n^2 + \frac{1}{n-p}\sum_{\lambda \in \Lambda_m} (\sigma_\lambda^r)^2 + \frac{1}{n}\sum_{i=1}^{n} \sigma(t_i)^2 \ , \tag{10}$$

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

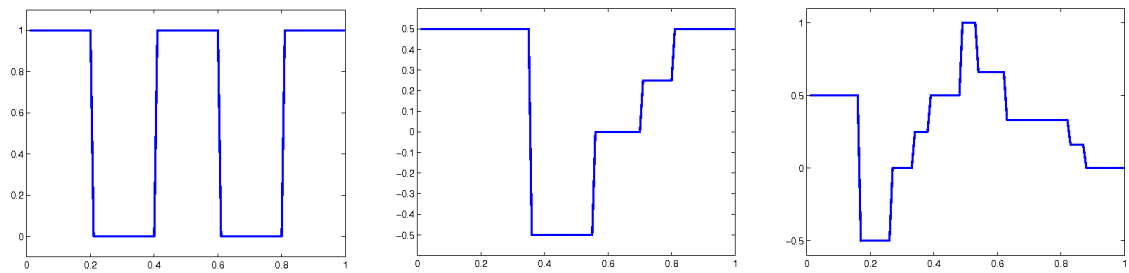


Figure 2: Regression functions s_1, s_2, s_3; s_1 and s_2 are piecewise constant with 4 jumps; s_3 is piecewise constant with 9 jumps.

where the approximation holds as soon as $\min_{\lambda \in \Lambda_m} n_\lambda$ *is large enough (in particular larger than* p*).*

The comparison of (6) and (10) shows that Lpo_p yields an almost unbiased estimator of $\|s - \widehat{s}_m\|_n^2$: The only difference is that the factor $1/n$ in front of the variance term $V(m)$ has been changed into $1/(n-p)$. Therefore, minimizing the Lpo_p estimator of the risk instead of the empirical risk allows to automatically take into account heteroscedasticity of data.

3.3 Simulation study

The goal of this section is to experimentally assess, for several values of p, the performance of Lpo_p for detecting a given number of changes in the mean of a heteroscedastic signal. This performance is also compared with that of empirical risk minimization.

3.3.1 Setting

The setting described in this section is used in all the experiments of the paper.

Data are generated according to (3) with $n = 100$. For every i, $t_i = i/n$ and ϵ_i has a standard Gaussian distribution. The regression function s is chosen among three piecewise constant functions s_1, s_2, s_3 plotted on Figure 2. The model collection described in Section 2.3 is used with $\mathcal{D}_n = \{1, \ldots, 4n/10\}$. The noise-level function $\sigma(\cdot)$ is chosen among the following functions:

1. Homoscedastic noise: $\sigma_c = 0.25\,\mathbb{1}_{[0,1]}$,
2. Heteroscedastic piecewise constant noise: $\sigma_{pc,1} = 0.2\,\mathbb{1}_{[0,1/3]} + 0.05\mathbb{1}_{[1/3,1]}$, $\sigma_{pc,2} = 2\sigma_{pc,1}$ or $\sigma_{pc,3} = 2.5\sigma_{pc,1}$.
3. Heteroscedastic sinusoidal noise: $\sigma_s = 0.5 \sin(t\pi/4)$.

All combinations between the regression functions $(s_i)_{i=1,2,3}$ and the five noise-levels $\sigma.$ have been considered, each time with $N = 10\,000$ independent samples. Results below only report a small part of the entire simulation study but intend to be representative of the main observed behaviour. A more complete report of the results, including other

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

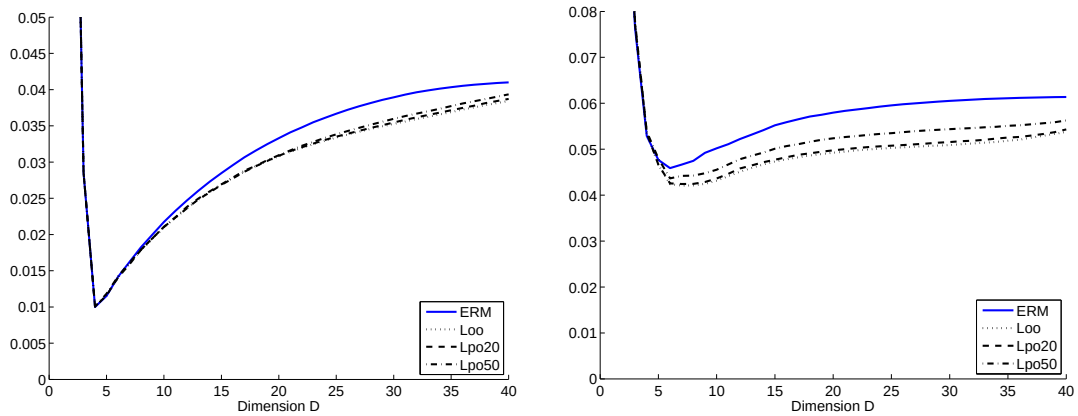


Figure 3: $\mathbb{E}\left[\left\|s-\widehat{s}_{\widehat{m}_{\mathfrak{P}}(D)}\right\|_n^2\right]$ as a function of D for $\mathfrak{P}$ among 'ERM' (empirical risk minimization), 'Loo' (Leave-one-out), 'Lpo(20)' (Lpo_p with $p=20$) and 'Lpo(50)' (Lpo_p with $p=50$). Left: homoscedastic (s_2,σ_c). Right: heteroscedastic $(s_3,\sigma_{pc,3})$. All curves have been estimated from $N=10\,000$ independent samples; error bars are all negligible in front of visible differences (the larger ones are smaller than 8.10^{-5} on the left, and smaller than 2.10^{-4} on the right). The curves $D\mapsto\left\|s-\widehat{s}_{\widehat{m}_{\mathfrak{P}}(D)}\right\|_n^2$ behave similarly to their expectations.

regression functions s and noise-level functions σ, is given in the second authors' thesis [21, Chapter 7]; see also Section 3 of the supplementary material.

3.3.2 Results: Comparison of segmentations for each dimension

The segmentations of each dimension $D\in\mathcal{D}_n$ obtained by empirical risk minimization ('ERM', Procedure 1) and Lpo_p (Procedure 2) for several values of p are compared on Figure 3, through the expected values of the quadratic loss $\mathbb{E}\left[\left\|s-\widehat{s}_{\widehat{m}_{\mathfrak{P}}(D)}\right\|_n^2\right]$ for procedure $\mathfrak{P}$.

On the one hand, when data are homoscedastic (Figure 3, left), all procedures yield similar performances for all dimensions up to twice the best dimension; Lpo_p performs significantly better for larger dimensions. Therefore, unless the dimension is strongly overestimated (whatever the way D is chosen), all procedures are equivalent with homoscedastic data.

On the other hand, when data are heteroscedastic (Figure 3, right), ERM yields significantly worse performance than Lpo for dimensions larger than half the true dimension. As explained in Sections 3.1 and 3.2.4, $\widehat{m}_{\mathrm{ERM}}(D)$ often puts breakpoints inside pure noise for dimensions D smaller than the true dimension, whereas Lpo does not have this drawback. Therefore, whatever the choice of the dimension (except $D\leq 4$, that is for detecting the obvious jumps), Lpo should be prefered to empirical risk minimization as soon as data are heteroscedastic.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

$s.$	$\sigma.$	ERM	Loo	Lpo_{20}	Lpo_{50}
2	c	**2.88** $\pm$ 0.01	2.93 $\pm$ 0.01	2.93 $\pm$ 0.01	2.94 $\pm$ 0.01
	pc,1	1.31 $\pm$ 0.02	1.16 $\pm$ 0.02	1.14 $\pm$ 0.02	**1.11** $\pm$ 0.01
	pc,3	3.09 $\pm$ 0.03	2.52 $\pm$ 0.03	2.48 $\pm$ 0.03	**2.32** $\pm$ 0.03
3	c	**3.18** $\pm$ 0.01	3.25 $\pm$ 0.01	3.29 $\pm$ 0.01	3.44 $\pm$ 0.01
	pc,1	3.00 $\pm$ 0.01	**2.67** $\pm$ 0.02	**2.68** $\pm$ 0.02	2.77 $\pm$ 0.02
	pc,3	4.41 $\pm$ 0.02	**3.97** $\pm$ 0.02	4.00 $\pm$ 0.02	4.11 $\pm$ 0.02

Table 1: Average performance $C_{\text{or}}\left(\llbracket \mathfrak{P}, \text{Id} \rrbracket\right)$ for change-point detection procedures $\mathfrak{P}$ among ERM, Loo and Lpo_p with $p = 20$ and $p = 50$. Several regression functions s and noise-level functions σ have been considered, each time with $N = 10\,000$ independent samples. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$, measuring the uncertainty of the estimated performance.

3.3.3 Results: Comparison of the "best" segmentations

This section focuses on the segmentation obtained with the best possible choice of D, that is the one corresponding to the minimum of $D \mapsto \left\| s - \widehat{s}_{\widehat{m}_{\mathfrak{P}}(D)} \right\|_n^2$ (plotted on Figure 3) for procedures $\mathfrak{P}$ among ERM, Loo, and Lpo_p with $p = 20$ and $p = 50$. Therefore, the performance of a procedure $\mathfrak{P}$ is defined by

$$C_{\text{or}}\left(\llbracket \mathfrak{P}, \text{Id} \rrbracket\right) := \frac{\mathbb{E}\left[\inf_{1 \leq D \leq n} \left\{ \left\| s - \widehat{s}_{\widehat{m}_{\mathfrak{P}}(D)} \right\|_n^2 \right\}\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n} \left\{ \left\| s - \widehat{s}_m \right\|_n^2 \right\}\right]} ,$$

which measures what is lost compared to the oracle when selecting one segmentation $\widehat{m}_{\mathfrak{P}}(D)$ per dimension. Even if the choice of D is a real practical problem—which will be tackled in the next sections—, $C_{\text{or}}\left(\llbracket \mathfrak{P}, \text{Id} \rrbracket\right)$ helps to understand which is the best procedure for selecting a segmentation of a given dimension. The notation $C_{\text{or}}\left(\llbracket \mathfrak{P}, \text{Id} \rrbracket\right)$ has been chosen for consistency with notation used in the next sections (see Section 5.1).

Table 1 confirms the results of Section 3.3.2. On the one hand, when data are homoscedastic, ERM performs slightly better than Loo or Lpo_p. On the other hand, when data are heteroscedastic, Lpo_p often performs better than ERM (whatever p), and the improvement can be large (more than 20% in the setting $(s_2, \sigma_{pc,3})$). Overall, when homoscedasticity of the signal is questionable, Lpo_p appears much more reliable than ERM for localizing a given number of change-points of the mean.

The question of choosing p for optimizing the performance of Lpo_p remains a widely open problem. The simulation experiment summarized with Table 1 only shows that Lpo_p improves ERM whatever p, the optimal value of p depending on s and σ.

4 Estimation of the number of breakpoints

In this section, the number of breakpoints is no longer fixed or known *a priori*. The goal is precisely the estimation of this number, as often needed with real data.

Two main procedures are considered. First, a penalization procedure introduced by Birgé and Massart [15] is analyzed in Section 4.1; this procedure is successful for change-

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

point detection when data are homoscedastic [28, 30]. On the basis of this analysis, V-fold cross-validation (VFCV) is then proposed as an alternative to Birgé and Massart's penalization procedure (BM) when data can be heteroscedastic.

In order to enable the comparison between BM and VFCV when focusing on the question of choosing the number of breakpoints, VFCV is used for choosing among the same segmentations as BM, that is $\{\widehat{m}_{\mathrm{ERM}}(D)\}_{D\in\mathcal{D}_n}$. The combination of VFCV for choosing D with the new procedures proposed in Section 3 will be studied in Section 5.

4.1 Birgé and Massart's penalization

First, let us define precisely the penalization procedure proposed by Birgé and Massart [15] successfully used for change-point detection in [28, 30].

Procedure 3 (Birgé and Massart [15])**.**

1. $\forall m \in \mathcal{M}_n$, $\widehat{s}_m := \mathrm{ERM}\left(S_m; P_n\right)$.
2. $\widehat{m}_{\mathrm{BM}} := \arg\min_{m\in\mathcal{M}_n,\, D_m\in\mathcal{D}_n} \left\{P_n\gamma(\widehat{s}_m) + \mathrm{pen}_{\mathrm{BM}}(m)\right\}$, where for every $m \in \mathcal{M}_n$, the penalty $\mathrm{pen}_{\mathrm{BM}}(m)$ only depends on S_m through its dimension:
$$\mathrm{pen}_{\mathrm{BM}}(m) = \mathrm{pen}_{\mathrm{BM}}(D_m) := \frac{\widehat{C}D_m}{n}\left(5 + 2\log\left(\frac{n}{D_m}\right)\right) \quad , \tag{11}$$
where $\widehat{C}$ is estimated from data using Birgé and Massart's *slope heuristics* [16, 8], as proposed by Lebarbier [30] and by Lavielle [28]. See Section 1 of the supplementary material for a detailed discussion about $\widehat{C}$.
3. $\widetilde{s}_{\mathrm{BM}} := \widehat{s}_{\widehat{m}_{\mathrm{BM}}}$.

All $m \in \mathcal{M}_n(D)$ are penalized in the same way by $\mathrm{pen}_{\mathrm{BM}}(m)$, so that Procedure 3 actually selects a segmentation among $\{\widehat{m}_{\mathrm{ERM}}(D)\}_{D\in\mathcal{D}_n}$. Therefore, Procedure 3 can be reformulated as follows, as noticed in [16, Section 4.3].

Procedure 4 (Reformulation of Procedure 3)**.**

1. $\forall D \in \mathcal{D}_n$, $\widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)} := \mathrm{ERM}\left(\widetilde{S}_D; P_n\right)$ where $\widetilde{S}_D := \bigcup_{m\in\mathcal{M}_n(D)} S_m$.
2. $\widehat{D}_{\mathrm{BM}} := \arg\min_{D\in\mathcal{D}_n} \left\{P_n\gamma(\widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)}) + \mathrm{pen}_{\mathrm{BM}}(D)\right\}$ where $\mathrm{pen}_{\mathrm{BM}}(D)$ is defined by (11).
3. $\widetilde{s}_{\mathrm{BM}} := \widehat{s}_{\widehat{m}_{\mathrm{ERM}}(\widehat{D}_{\mathrm{BM}})}$.

In the following, 'BM' denotes Procedure 4 and
$$\mathrm{crit}_{\mathrm{BM}}(D) := P_n\gamma(\widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)}) + \mathrm{pen}_{\mathrm{BM}}(D)$$
is called the BM criterion.

Procedure 4 clarifies the reason why $\mathrm{pen}_{\mathrm{BM}}$ must be larger than Mallows' C_p penalty. Indeed, for every $m \in \mathcal{M}_n$, Lemma 1 shows that when data are homoscedastic, $P_n\gamma(\widehat{s}_m) + \mathrm{pen}(m)$ is an unbiased estimator of $\|s - \widehat{s}_m\|_n^2$ when $\mathrm{pen}(m) = 2\sigma^2 D_m n^{-1}$, that is Mallows'

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

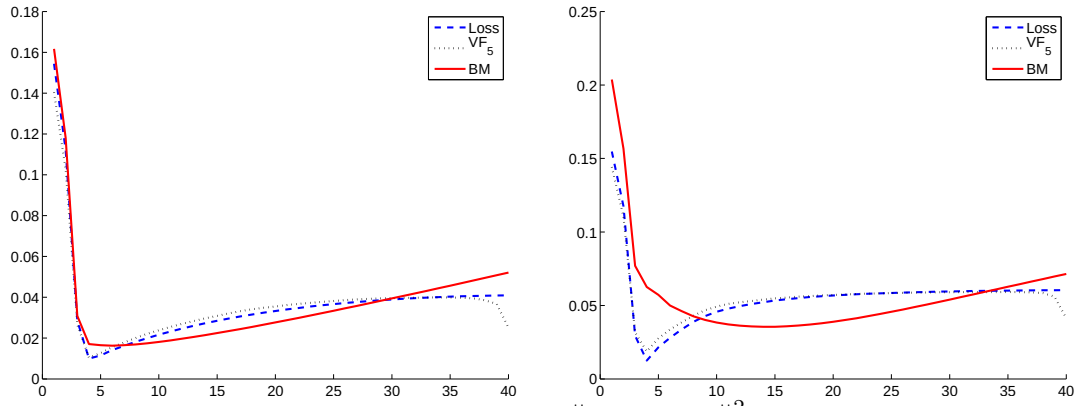


Figure 4: Comparison of the expectations of $\left\|s - \widehat{s}_{\widehat{m}(D)}\right\|_n^2$ ('Loss'), $\mathrm{crit}_{\mathrm{VF}_V}(D)$ ('VF_5') and $\mathrm{crit}_{\mathrm{BM}}(D)$ ('BM'). Data are generated as explained in Section 3.3.1. Left: homoscedastic (s_2, σ_c). Right: heteroscedastic $(s_2, \sigma_{pc,3})$. Expectations have been estimated from $N = 10\,000$ independent samples; error bars are all negligible in front of visible differences (the larger ones are smaller than 5.10^{-4} on the left, and smaller than 2.10^{-3} on the right). Similar behaviours are observed for every single sample, with slightly larger fluctuations for $\mathrm{crit}_{\mathrm{VF}_V}(D)$ than for $\mathrm{crit}_{\mathrm{BM}}(D)$. The curves 'BM' and '$\mathrm{VF}_5$' have been shifted in order to make comparison with 'Loss' easier, without changing the location of the minimum.

C_p penalty. When $\mathrm{Card}(\mathcal{M}_n)$ is at most polynomial in n, Mallows' C_p penalty leads to an efficient model selection procedure, as proved in several regression frameworks [41, 31, 10]. Hence, Mallows' C_p penalty is an adequate measure of the "capacity" of any vector space S_m of dimension D_m, at least when data are homoscedastic.

On the contrary, in the change-point detection framework, $\mathrm{Card}(\mathcal{M}_n)$ grows exponentially with n. The formulation of Procedure 4 points out that $\mathrm{pen}_{\mathrm{BM}}(D)$ has been built so that $\mathrm{crit}_{\mathrm{BM}}(D)$ estimates unbiasedly $\left\|s - \widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)}\right\|_n^2$ for every D, where $\widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)}$ is the empirical risk minimizer over $\widetilde{S}_D$. Hence, $\mathrm{pen}_{\mathrm{BM}}(D)$ measures the "capacity" of $\widetilde{S}_D$, which is much bigger than a vector space of dimension D. Therefore, $\mathrm{pen}_{\mathrm{BM}}$ should be larger than Mallows' C_p, as confirmed by the results of Birgé and Massart [16] on minimal penalties for exponential collections of models.

Simulation experiments support the fact that $\mathrm{crit}_{\mathrm{BM}}(D)$ is an unbiased estimator of $\left\|s - \widehat{s}_{\widehat{m}(D)}\right\|_n^2$ for every D (up to an additive constant) when data are homoscedastic (Figure 4 left). However, when data are heteroscedastic, theoretical results proved by Birgé and Massart [15, 16] no longer apply, and simulations show that $\mathrm{crit}_{\mathrm{BM}}(D)$ does not always estimate $\left\|s - \widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)}\right\|_n^2$ well (Figure 4 right). This result is consistent with Lemma 1, as well as the suboptimality of penalties proportional to D_m for model selection among a polynomial collection of models when data are heteroscedastic [6].

Therefore, $\mathrm{pen}_{\mathrm{BM}}(D)$ is not an adequate capacity measure of $\widetilde{S}_D$ in general when data are heteroscedastic, and another capacity measure is required.

4.2 Cross-validation

As shown in Section 3.2.2, CV can be used for estimating the quadratic loss $\|s - \mathcal{A}(P_n)\|_n^2$ for any algorithm $\mathcal{A}$. In particular, CV was successfully used in Section 3 for estimating

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

the quadratic risk of $\mathrm{ERM}(S_m;\cdot)$ for all segmentations $m \in \mathcal{M}_n(D)$ with a given number $(D-1)$ of breakpoints (Procedure 2), even when data are heteroscedastic.

Therefore, CV methods are natural candidates for fixing BM's failure. The proposed procedure—with VFCV—is the following.

Procedure 5.

1. $\forall D \in \mathcal{D}_n,\ \widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)} := \mathrm{ERM}\left(\widetilde{S}_D; P_n\right)$,
2. $\widehat{D}_{\mathrm{VF}_V} := \arg\min_{D\in\mathcal{D}_n}\left\{\mathrm{crit}_{\mathrm{VF}_V}(D)\right\}$
   where $\mathrm{crit}_{\mathrm{VF}_V}(D) := \widehat{R}_{\mathrm{VF}_V}\left(\mathrm{ERM}\left(\widetilde{S}_D(\cdot);\cdot\right),\cdot\right)$ and $\widehat{R}_{\mathrm{VF}_V}$ is defined by (9).

Remark 3. In algorithm $(t_i,Y_i)_{1\le i\le n} \mapsto \mathrm{ERM}\left(\widetilde{S}_D; P_n\right)$, the model $\widetilde{S}_D$ depends on the design points. When the training set is $(t_i,Y_i)_{i\notin B_k}$, the model $\widetilde{S}_D$ is the union of the S_m such that $\forall\lambda\in\Lambda_m$, I_λ contains at least two elements of $\{t_i \text{ s.t. } i\notin B_k\}$. Such an m exists as soon as $D \le (n-\max_k\{\mathrm{Card}(B_k)\})/2$ and two consecutive design points t_i, t_{i+1} always belong to different blocks B_k, which is always assumed in this paper. Note that the dynamic programming algorithms [13] quoted in Section 3.2.3 can straightforwardly take into account such constraints when minimizing the empirical risk over $\widetilde{S}_D$.

The dependence of $\widetilde{S}_D$ on the design explains why $\mathrm{crit}_{\mathrm{VF}_V}(D)$ decreases for D close to $n(V-1)/(2V)$, as observed on Figure 4. Indeed, when D is close to $n_t/2$ (where n_t is the size of the design), only a few $\{S_m\}_{m\in\mathcal{M}_{n_t}(D)}$ remain in $\widetilde{S}_D$; for instance, when $D=n_t/2$, $\widetilde{S}_D$ is equal to one of the $\{S_m\}_{m\in\mathcal{M}_{n_t}(D)}$. Therefore, the "capacity" of $\widetilde{S}_D$ decreases in the neighborhood of $D=n_t/2$.

Similar procedures can be defined with Loo and Lpo_p instead of VFCV. The interest of VFCV is its reasonably small computational cost—taking $V\le 10$ for instance—, since no closed-form formula exists for CV estimators of the risk of $\mathrm{ERM}\left(\widetilde{S}_D; P_n\right)$.

4.3 Simulation results

A simulation experiment was performed in the setting presented in Section 3.3.1, for comparing BM and VF_V with $V=5$ blocks. A representative picture of the results is given by Figure 4 and by Table 2 [see 21, Chapter 7, and Section 3 of the supplementary material for additional results].

As illustrated by Figure 4, $\mathrm{crit}_{\mathrm{VF}_V}(D)$ can be used for measuring the capacity of $\widetilde{S}_D$. Indeed, VFCV correctly estimates the risk of empirical risk minimizers over $\widetilde{S}_D$ for every D and for both homoscedastic and heteroscedastic data; $\mathrm{crit}_{\mathrm{VF}_V}(D)$ only underestimates $\left\|s-\widehat{s}_{\widehat{m}(D)}\right\|_n^2$ for dimensions D close to $n(V-1)/(2V)$, for reasons explained at the end of Remark 3. On the contrary, $\mathrm{crit}_{\mathrm{BM}}(D)$ is a poor estimate of $\left\|s-\widehat{s}_{\widehat{m}(D)}\right\|_n^2$ when data are heteroscedastic.

Subsequently, VFCV yields a much smaller performance index

$$C_{\mathrm{or}}\left([\![\mathrm{ERM},\mathfrak{P}]\!]\right) := \frac{\mathbb{E}\left[\left\|s-\widehat{s}_{\widehat{m}_{\mathrm{ERM}}(\widehat{D}_{\mathfrak{P}})}\right\|_n^2\right]}{\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\{\left\|s-\widehat{s}_m(P_n)\right\|_n^2\right\}\right]}$$

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

s.	σ.	Oracle	VF_5	BM
2	c	2.88 ± 0.01	**4.51** ± 0.03	5.27 ± 0.03
	pc,2	2.88 ± 0.02	**6.58** ± 0.06	19.82 ± 0.07
	s	3.01 ± 0.01	**5.21** ± 0.04	9.69 ± 0.40
3	c	3.18 ± 0.01	4.41 ± 0.02	**4.39** ± 0.01
	pc,2	4.06 ± 0.02	**5.99** ± 0.02	7.86 ± 0.03
	s	4.02 ± 0.01	**5.97** ± 0.03	7.59 ± 0.03

Table 2: Performance $C_{\mathrm{or}}\left(\llbracket \mathrm{ERM}, \mathfrak{P} \rrbracket\right)$ for $\mathfrak{P} = \mathrm{Id}$ (that is, choosing the dimension $D^\star := \arg\min_{D \in \mathcal{D}_n} \left\{ \left\| s - \widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)} \right\|_n^2 \right\}$), $\mathfrak{P} = \mathrm{VF}_V$ with $V = 5$ or $\mathfrak{P} = \mathrm{BM}$. Several regression functions s and noise-level functions σ have been considered, each time with $N = 10\,000$ independent samples. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$, measuring the uncertainty of the estimated performance.

than BM when data are heteroscedastic (Table 2); see also the supplementary material (Section 1) for details about the performances of BM and possible ways to improve them. When data are homoscedastic, VFCV and BM have similar performances (maybe with a slight advantage for BM), which is not surprising since BM uses the knowledge that data are homoscedastic. Moreover, BM has been proved to be optimal in the homoscedastic setting [15, 16].

Overall, VFCV appears to be a reliable alternative to BM when no prior knowledge guarantees that data are homoscedastic.

5 New change-point detection procedures via cross-validation

Sections 3 and 4 showed that when data are heteroscedastic, CV can be used successfully instead of penalized criteria for detecting breakpoints given their number, as well as for estimating the number of breakpoints. Nevertheless, in Section 4, the segmentations compared by CV were obtained by empirical risk minimization, so that they can be suboptimal according to the results of Section 3.

The next step for obtaining reliable change-point detection procedures for heteroscedastic data is to combine the two ideas, that is, to use CV twice. The goal of the present section is to properly define such procedures (with various kinds of CV) and to assess their performances.

5.1 Definition of a family of change-point detection procedures

The general strategy used in this article for change-point detection relies on two steps: First, detect where $(D-1)$ breakpoints should be located for every $D \in \mathcal{D}_n$; second, estimate the number $(D-1)$ of breakpoints. This strategy can be summarized with the following procedure:

Procedure 6 (General two-step change-point detection procedure)**.**

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

1. $\forall D \in \mathcal{D}_n$, $\mathcal{A}_D(P_n) := \widehat{s}_{\widehat{m}(D)} = \arg\min_{m\in\mathcal{M}_n(D)} \{\mathrm{crit}_1(S_m, P_n)\}$ where for every model S, $\mathrm{crit}_1(S, P_n) \in \mathbb{R}$ estimates $\|s - \mathrm{ERM}(S; P_n)\|_n^2$ and $\widehat{s}_m = \mathrm{ERM}(S_m; P_n)$ is defined as in Section 3.1.

2. $\widehat{D} = \arg\min_{D\in\mathcal{D}_n} \{\mathrm{crit}_2(\mathcal{A}_D, P_n)\}$, where for every algorithm $\mathcal{A}_D$, $\mathrm{crit}_2(\mathcal{A}_D, P_n) \in \mathbb{R}$ estimates $\|s - \mathcal{A}_D(P_n)\|_n^2$.

3. Output: the segmentation $\widehat{m}(\widehat{D})$ and the corresponding estimator $\widehat{s}_{\widehat{m}(\widehat{D})}$ of s.

Let us now detail which are the candidate criteria crit_1 and crit_2 for being used in Procedure 6. For the first step:

- The empirical risk ('ERM') is

$$\mathrm{crit}_{1,\mathrm{ERM}}(S, P_n) := P_n\gamma\left(\mathrm{ERM}\left(S; P_n\right)\right)$$

- The Leave-p-out estimator of the risk ('Lpo_p') is, for every $p \in \{1, \ldots, n-1\}$,

$$\mathrm{crit}_{1,\mathrm{Lpo}}(S, P_n, p) := \widehat{R}_{\mathrm{Lpo}_p}(\mathrm{ERM}(S; \cdot), P_n)$$

- For comparison, the ideal criterion ('Id') is defined by $\mathrm{crit}_{1,Id}(S, P_n) := \|s - \mathrm{ERM}(S; P_n)\|_n^2$.

As in Section 3, Loo denotes Lpo_1. The VFCV estimator of the risk $\widehat{R}_{\mathrm{VF}_V}$ could also be used as crit_1; it will not be considered in the following because it is computationally more expensive and more variable than Lpo (see Section 3.2).

For the second step:

- Birgé and Massart's penalization criterion ('BM') is

$$\mathrm{crit}_{2,\mathrm{BM}}(\mathcal{A}_D, P_n) := P_n\gamma\left(\mathcal{A}_D\left(P_n\right)\right) + \mathrm{pen}_{\mathrm{BM}}(D) \ ,$$

  where $\mathrm{pen}_{\mathrm{BM}}(D)$ is defined by (11) with $c_1 = 5$, $c_2 = 2$ and $\widehat{C}$ is chosen by the slope heuristics (see Section 1 of the supplementary material).

- The V-fold cross-validation estimator of the risk ('VF_V') is, for every $V \in \{1, \ldots, n\}$,

$$\mathrm{crit}_{2,\mathrm{VF}_V}(\mathcal{A}_D, P_n) := \widehat{R}_{\mathrm{VF}_V}(\mathcal{A}_D, P_n) \ ,$$

  where $\widehat{R}_{\mathrm{VF}_V}$ is defined by (9) and the blocks $B_1, \ldots, B_V$ are chosen as in Procedure 5 (see Remark 3).

- For comparison, the ideal criterion ('Id') is defined by $\mathrm{crit}_{2,Id}(\mathcal{A}_D, P_n) := \|s - \mathcal{A}_D(P_n)\|_n^2$.

Remark 4. For crit_2, definitions using Lpo could theoretically be considered. They are not investigated here because they are computationally intractable.

In the following, the notation $[\![\alpha, \beta]\!]$ is used as a shortcut for "Procedure 6 with $\mathrm{crit}_{1,\alpha}$ and $\mathrm{crit}_{2,\beta}$", and the outputs of $[\![\alpha, \beta]\!]$ are denoted by $\widehat{m}_{[\![\alpha,\beta]\!]} \in \mathcal{M}_n$ and $\widetilde{s}_{[\![\alpha,\beta]\!]} \in \mathcal{S}^*$. For instance, BM coincides with $[\![\mathrm{ERM}, \mathrm{BM}]\!]$; Procedures $[\![\alpha, Id]\!]$ are compared for several α in Section 3; Procedures $[\![\mathrm{ERM}, \beta]\!]$ are compared for $\beta \in \{\mathrm{Id}, \mathrm{BM}, \mathrm{VF}_5\}$ in Section 4.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

$s.$	$\sigma.$	$[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$	$[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$	$[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$	$[\![\mathrm{ERM}, \mathrm{BM}]\!]$
1	c	5.40 ± 0.05	5.03 ± 0.05	5.10 ± 0.05	**3.91** ± 0.03
	pc,1	11.96 ± 0.03	**10.25** ± 0.03	10.28 ± 0.03	12.85 ± 0.04
	pc,3	4.96 ± 0.05	**4.82** ± 0.04	**4.79** ± 0.05	13.08 ± 0.04
	s	7.33 ± 0.06	**6.82** ± 0.05	6.99 ± 0.06	9.41 ± 0.04
2	c	**4.51** ± 0.03	**4.55** ± 0.03	**4.50** ± 0.03	5.27 ± 0.03
	pc,1	11.67 ± 0.09	**10.26** ± 0.08	**10.29** ± 0.08	19.36 ± 0.07
	pc,3	6.66 ± 0.06	**5.81** ± 0.06	**5.74** ± 0.06	20.12 ± 0.06
	s	**5.21** ± 0.04	**5.19** ± 0.03	**5.17** ± 0.03	9.69 ± 0.04
3	c	4.41 ± 0.02	4.54 ± 0.02	4.62 ± 0.02	**4.39** ± 0.01
	pc,1	4.91 ± 0.02	**4.40** ± 0.02	**4.44** ± 0.02	6.50 ± 0.02
	pc,3	6.32 ± 0.02	**5.74** ± 0.02	5.81 ± 0.02	8.47 ± 0.03
	s	5.97 ± 0.02	**5.72** ± 0.02	5.86 ± 0.02	7.59 ± 0.03

Table 3: Performance $C_{\mathrm{or}}(\mathfrak{P})$ for several change-point detection procedures $\mathfrak{P}$ in several settings (s, σ). Each time, $N = 10\,000$ independent samples have been generated. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$.

5.2 Simulation study

A simulation experiment compares procedures $[\![\alpha, \mathrm{VF}_5]\!]$ for several α and $[\![\mathrm{ERM}, \mathrm{BM}]\!]$, in the setting described in Section 3.3.1. A representative picture of the results is given by Table 3 [see 21, Chapter 7, for additional results]. The (statistical) performance of each competing procedure $\mathfrak{P}$ is measured by

$$C_{\mathrm{or}}(\mathfrak{P}) := \frac{\mathbb{E}\left[\|s - \widetilde{s}_{\mathfrak{P}}(P_n)\|_n^2\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n}\left\{\|s - \widehat{s}_m(P_n)\|_n^2\right\}\right]} ,$$

both expectations being evaluated by averaging over $N = 10\,000$ independent samples.

Remark 5. Birgé and Massart's penalization procedure is the only classical change-point detection procedure considered in this experiment for two reasons. First, change-point detection procedure looking for changes in the distribution of Y_i would clearly fail to detect changes in the mean of the signal, as soon as the noise-level σ varies inside areas where the mean is constant. Second, among procedures detecting changes in the mean of a signal in a setting comparable to the setting of the paper (that is, frequentist, parametric, off-line, with no information on the number of change-points), BM appears to be the most reliable procedure according to recent papers [28, 30]. The question of the calibration of $\widehat{C}$ is addressed in Section 1 of the supplementary material.

First, BM is consistently outperformed by the other procedures, except in the homoscedastic settings in which it confirms its strength.

Second, empirical risk minimization (ERM) slightly outperforms CV (Loo and Lpo_{20}) when data are homoscedastic. On the contrary, when data are heteroscedastic, Loo and Lpo_{20} clearly outperform ERM, often by a margin larger than 10% (for instance, when $\sigma = \sigma_{pc,1}$). Therefore, the results of Section 3 are confirmed when using VF_5 (instead of Id) for choosing the dimension.

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Framework	A	B	C
$[\![\mathrm{ERM}, \mathrm{BM}]\!]$	6.82 ± 0.03	7.21 ± 0.04	13.49 ± 0.07
$[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$	4.78 ± 0.03	5.09 ± 0.03	7.17 ± 0.05
$[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$	**4.65** ± 0.03	**4.88** ± 0.03	6.61 ± 0.05
$[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$	4.78 ± 0.03	**4.91** ± 0.03	**6.49** ± 0.05
$[\![\mathrm{Lpo}_{50}, \mathrm{VF}_5]\!]$	4.97 ± 0.03	5.18 ± 0.04	6.69 ± 0.05

Table 4: Performance $C_{\mathrm{or}}^{(R)}(\mathfrak{P})$ of several model selection procedures $\mathfrak{P}$ in frameworks A, B, C with sample size $n = 100$. In each framework, $N = 10,000$ independent samples have been considered. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$.

Third, the comparison between $[\![\mathrm{Lpo}_p, \mathrm{VF}_5]\!]$ for several values of p is less clear. Even though $p = 1$ (that is, Loo) mostly outperforms $p = 20$ (as well as $p = 50$, see the supplementary material), differences are small and often not significant despite the large number of samples generated. The conclusion of the simulation experiment on this question is that all values of p between 1 and $n/2$ all perform almost equally well, with a small advantage to $p = 1$ which may not be general. Let us mention here that the choice of p for Lpo_p is usually related to overpenalization [see for instance 5, 19, 21], but it seems difficult to characterize the settings for which overpenalization is needed for detecting change-points given their number.

5.3 Random frameworks

In order to assess the generality of the results of Table 3, the procedures considered in Section 5.2 have been compared in three random settings. The following process has been repeated $N = 10,000$ times. First, piecewise constant functions s and σ are randomly chosen (see Section 2 of the supplementary material for details). Then, given s and σ, a data sample $(t_i, Y_i)_{1\le i\le n}$ is generated as described in Section 3.3.1, and the same collection of models is used. Finally, each procedure $\mathfrak{P}$ is applied to the sample $(t_i, Y_i)_{1\le i\le n}$, and its loss $\|s - \widetilde{s}_{\mathfrak{P}}(P_n)\|_n^2$ is measured, as well as the loss of the oracle $\inf_{m\in\mathcal{M}_n}\left\{\|s - \widehat{s}_m\|_n^2\right\}$.

To summarize the results, the quality of each procedure is measured by the ratio

$$C_{\mathrm{or}}^{(R)}(\mathfrak{P}) = \frac{\mathbb{E}_{s,\sigma,\epsilon_1,\ldots,\epsilon_n}\left[\|s - \widetilde{s}_{\mathfrak{P}}(P_n)\|_n^2\right]}{\mathbb{E}_{s,\sigma,\epsilon_1,\ldots,\epsilon_n}\left[\inf_{m\in\mathcal{M}_n}\left\{\|s - \widehat{s}_m\|_n^2\right\}\right]} .$$

The notation $C_{\mathrm{or}}^{(R)}(\mathfrak{P})$ differs from $C_{\mathrm{or}}(\mathfrak{P})$ to emphasize that each expectation includes the randomness of s and σ, in addition to the one of $(\epsilon_i)_{1\le i\le n}$.

The results of this experiment—which are reported in Table 4—mostly confirm the results of the previous section (except that all the frameworks are heteroscedastic here), that is, whatever p, $[\![\mathrm{Lpo}_p, \mathrm{VF}_5]\!]$ outperforms $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$, which strongly outperforms $[\![\mathrm{ERM}, \mathrm{BM}]\!]$. Similar results—not reported here—have been obtained with a sample size $n = 200$ and $N = 1\,000$ samples.

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

Moreover, the difference between the performances of $[\![\mathrm{Lpo}_p, \mathrm{VF}_5]\!]$ and $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$ is the largest in setting C and the smallest in setting A. This fact confirms the interpretation given in Section 3 for the failure of ERM for localizing a given number of change-points. Indeed, the main differences between frameworks A, B and C—which are precisely defined in Section 2 of the supplementary material— can be sketched as follows:

A the partitions on which s is built is often close to regular, and σ is chosen independently from s.

B the partitions on which s is built are often irregular, and σ is chosen independently from s.

C the partitions on which s is built are often irregular, and σ depends on s, so that the noise-level is smaller where s jumps more often.

In other words, frameworks A, B and C have been built so that for any $D \in \mathcal{D}_n$, the largest variations over $\mathcal{M}_n(D)$ of $V(m)$ (defined by (7)) occur in framework C, and the smallest variations occur in framework A. As a consequence, variations of the performance of $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$ compared to $[\![\mathrm{Lpo}_p, \mathrm{VF}_5]\!]$ according to the framework certainly come from the local overfitting phenomenon presented in Section 3.

6 Application to CGH microarray data

In this section, the new change-point detection procedures proposed in the paper are applied to CGH microarray data.

6.1 Biological context

The purpose of Comparative Genomic Hybridization (CGH) microarray experiments is to detect and map chromosomal aberrations. For instance, a piece of chromosome can be *amplified*, that is appear several times more than usual, or *deleted*. Such aberrations are often related to cancer disease.

Roughly, CGH profiles give the log-ratio of the DNA copy number along the chromosomes, compared to a reference DNA sequence [see 35–37, for details about the biological context of CGH data].

The goal of CGH data analysis is to detect abrupt changes in the mean of a signal (the log-ratio of copy numbers), and to estimate the mean in each segment. Hence, change-point detection procedures are needed.

Moreover, assuming that CGH data are homoscedastic is often unrealistic. Indeed, changes in the chemical composition of the sequence are known to induce changes in the variance of the observed CGH profile, possibly independently from variations of the true copy number. Therefore, procedures robust to heteroscedasticity, such as the ones proposed in Section 5, should yield better results—in terms of detecting changes of copy number—than procedures assuming homoscedasticity.

The data set considered in this section is based on the Bt474 cell lines, which denote epithelial cells obtained from human breast cancer tumors of a sixty-year-old woman [36]. A test genome of Bt474 cell lines is compared to a normal reference male genome. Even

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

though several chromosomes are studied in these cell lines, this section focuses on chromosomes 1 and 9. Chromosome 1 exhibits a putative heterogenous variance along the CGH profile, and chromosome 9 is likely to meet the homoscedasticity assumption. Log-ratios of copy numbers have been measured at 119 locations for chromosome 1 and at 93 locations for chromosome 9.

6.2 Procedures used in the CGH literature

Before applying Procedure 6 to the analysis of Bt474 CGH data, let us recall the definition of two change-point detection procedures, which were the most successful for analyzing the same data according to the literature [36].

The first procedure is a simplified version of BM proposed by Lavielle [28, Section 2] and first used on CGH data in [36]. Note that BM would give similar results on the data of Figure 5.

The second procedure—denoted by 'PML' for penalized maximum likelihood—aims at detecting changes in either the mean or the variance, that is breakpoints for (s, σ). The selected model is defined as the minimizer over $m \in \mathcal{M}_n$ of

$$\mathrm{crit}_{\mathrm{PML}}(m) := \sum_{\lambda \in \Lambda_m} n_\lambda \log \left(\frac{1}{n_\lambda} \sum_{t_i \in I_\lambda} (Y_i - \widehat{s}_m(t_i; P_n))^2 \right) + \widehat{C}'' D_m \ ,$$

where $n_\lambda = \mathrm{Card}\,\{t_i \in I_\lambda\}$ and $\widehat{C}''$ is estimated from data by the slope heuristics algorithm [28, 30].

6.3 Results

Results obtained with BMsimple, PML, $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$ and $[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$ on the Bt474 data set are reported on Figure 5.

For chromosome 9, BMsimple and PML yield (almost) the same segmentation, so that the homoscedasticity assumption is certainly not much violated. As expected, $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$ and $[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$ also yield very similar segmentations, which confirms the reliability of these procedures for homoscedastic signal [see 21, Section 7.6 for details].

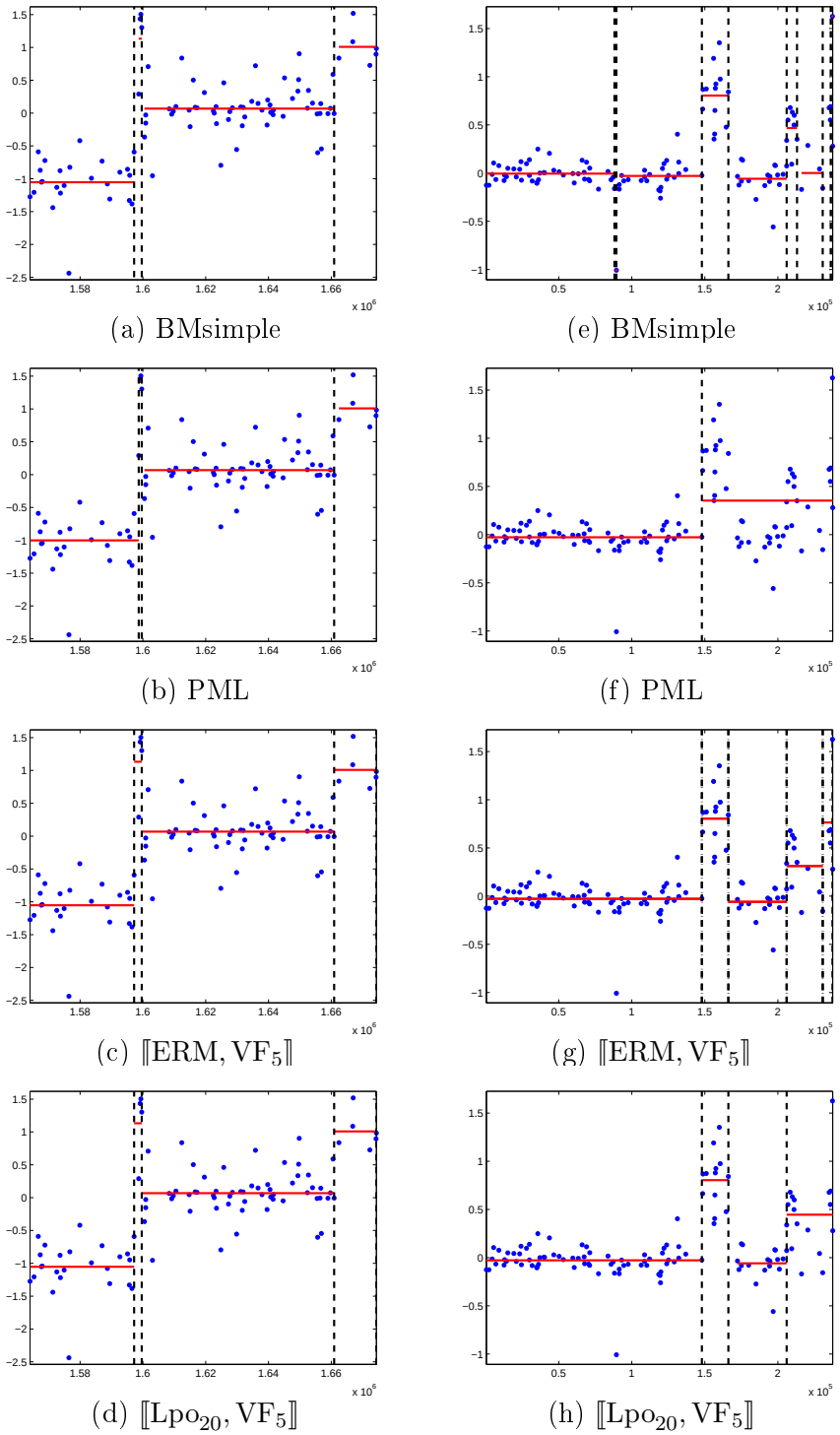
The picture is quite different for chromosome 1. Indeed, as shown by Figure 5 (right), BMsimple selects a segmentation with 7 breakpoints, whereas PML selects a segmentation with only one breakpoint. The major difference between BMsimple and PML supports at least the idea that these data must be heteroscedastic.

Nevertheless, none of the segmentations chosen by BMsimple and PML are entirely satisfactory: BMsimple relies on an assumption which is certainly violated; PML may use a change in the estimated variance for explaining several changes in the mean.

CV-based procedures $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$ and $[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$ yield two other segmentations, with a medium number of breakpoints, respectively 4 and 3. In view of the simulation experiments of the previous sections, the segmentation obtained via $[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$ should be the most reliable one since data are heteroscedastic. Therefore, the right of Figure 5 can be interpretated as follows: The noise-level is small in the first part of chromosome 1, then higher, but not as high as estimated by PML. In particular, the copy number changes

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript



(a) BMsimple (e) BMsimple

(b) PML (f) PML

(c) $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$ (g) $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$

(d) $[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$ (h) $[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$

Figure 5: Change-points locations along Chromosome 9 (Left) and Chromosome 1 (Right). The mean on each homogeneous region is indicated by plain horizontal lines.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

twice inside the second part of chromosome 1 (as defined by the segmentation obtained with PML), indicating that two putative amplified regions of chromosome 1 have been detected.

Note however that choosing among the segmentations obtained with $[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$ and $[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$ is not an easy task without additional data. A definitive answer would need further biological experiments.

7 Conclusion

7.1 Results summary

Cross-validation (CV) methods have been used to build reliable procedures (Procedure 6) for detecting changes in the mean of a signal whose variance may not be constant.

First, when the number of breakpoints is given, empirical risk minimization has been proved to fail for some heteroscedastic problems, from both theoretical and experimental points of view. On the contrary, the Leave-p-out (Lpo_p) remains robust to heteroscedasticity while being computationally efficient thanks to closed-form formulas given in Section 3.2.3 (Theorem 1).

Second, for choosing the number of breakpoints, the commonly used penalization procedure proposed by Birgé and Massart in the homoscedastic framework should not be applied to heteroscedastic data. V-fold cross-validation (VFCV) turns out to be a reliable alternative—both with homoscedastic and heteroscedastic data—, leading to much better segmentations in terms of quadratic risk when data are heteroscedastic. Furthermore, unlike usual deterministic penalized criteria, VFCV efficiently chooses among segmentations obtained by either Lpo or empirical risk minimization, without any specific change in the procedure.

To conclude, the combination of Lpo (for choosing a segmentation for each possible number of breakpoints) and VFCV yields the most reliable procedure for detecting changes in the mean of a signal which is not *a priori* known to be homoscedastic. The resulting procedure is computationally tractable for small values of V, since its computational complexity is of order $\mathcal{O}(Vn^2)$, which is similar to many comparable change-point detection procedures. The influence of V on the statistical performance of the procedure is not studied specifically in this paper; nevertheless, considering $V = 5$ only was sufficient to obtain a better statistical performance than Birgé and Massart's penalization procedure when data are heteroscedastic. When applied to real data (CGH profiles in Section 6), the proposed procedure turns out to be quite useful and effective, for a data set on which existing procedures highly disagree because of heteroscedasticity.

7.2 Prospects

The general form of Procedure 6 could be used with several other criteria, at both steps of the change-point detection procedure. For instance, resampling penalties [5] could be used at the first step, for localizing the change-points given their number. At the second step, V-fold penalization [6] could also be used instead of VFCV, with the same computational cost and possibly an improved statistical performance.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

Comparing precisely these resampling-based criteria for optimizing the performance of Procedure 6 would be of great interest and deserves further works. Simultaneously, several values of V should be compared for the second step of Procedure 6, and the precise influence of p when Lpo_p is used at the first step should be further investigated. Preliminary results in this direction can already be found in [21, Chapter 7].

References

[1] F. Abramovich, Y. Benjamini, D. Donoho, and I. Johnstone. Adapting to Unknown Sparsity by controlling the False Discovery Rate. *The Annals of Statistics*, 34(2):584–653, 2006.

[2] H. Akaike. Statistical predictor identification. *Ann. Inst. Statisti. Math.*, 22:203–217, 1969.

[3] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[4] David M. Allen. The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, 16:125–127, 1974.

[5] Sylvain Arlot. Model selection by resampling penalization, 2008. hal-00262478.

[6] Sylvain Arlot. Suboptimality of penalties proportional to the dimension for model selection in heteroscedastic regression, December 2008. arXiv:0812.3141.

[7] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

[8] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[9] Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.

[10] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

[11] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory and Relat. Fields*, 113:301–413, 1999.

[12] M. Basseville and N. Nikiforov. *The Detection of Abrupt Changes - Theory and Applications*. Prentice-Hall: Information and System Sciences Series, 1993.

[13] R. E. Bellman and S. E. Dreyfus. *Applied Dynamic Programming*. Princeton, 1962.

[14] L. Birgé and P. Massart. From model selection to adaptive estimation. In D. Pollard, E. Torgensen, and G. Yang, editors, *In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 55–87. Springer-Verlag, New York, 1997.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

[15] L. Birgé and P. Massart. Gaussian model selection. *J. European Math. Soc.*, 3(3):203–268, 2001.

[16] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[17] B. Brodsky and B. Darkhovsky. *Methods in Change-point problems*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.

[18] P. Burman. Comparative study of Ordinary Cross-Validation, v-Fold Cross-Validation and the repeated Learning-Testing Methods. *Biometrika*, 76(3):503–514, 1989.

[19] A. Celisse. Density estimation via cross-validation: Model selection point of view. Technical report, arXiv, 2008.

[20] A. Celisse and S. Robin. Nonparametric density estimation by exact leave-p-out cross-validation. *Computational Statistics and Data Analysis*, 52(5):2350–2368, 2008.

[21] Alain Celisse. *Model selection via cross-validation in density estimation, regression and change-points detection*. PhD thesis, University Paris-Sud 11, December 2008. oai:tel.archives-ouvertes.fr:tel-00346320_v1.

[22] S. Dudoit and M. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

[23] S. Geisser. A predictive approach to the random effect model. *Biometrika*, 61(1):101–107, 1974.

[24] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.

[25] Xavier Gendre. Simultaneous estimation of the mean and the variance in heteroscedastic gaussian regression, 2008.

[26] M. Kearns, Y. Mansour, A. Y. Ng, and D. Ron. An Experimental and Theoretical Comparison of Model Selection Methods. *Machine Learning*, 27:7–50, 1997.

[27] P. A. Lachenbruch and M. R. Mickey. Estimation of Error Rates in Discriminant Analysis. *Technometrics*, 10(1):1–11, 1968.

[28] M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85:1501–1510, 2005.

[29] M. Lavielle and G. Teyssière. Detection of Multiple Change-Points in Multivariate Time Series. *Lithuanian Mathematical Journal*, 46:287–306, 2006.

[30] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

[31] K.-C. Li. Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. *The Annals of Statistics*, 15(3):958–975, 1987.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

[32] C. L. Mallows. Some comments on C_p. *Technometrics*, 15:661–675, 1973.

[33] B. Q. Mia and L. C. Zhao. On detection of change points when the number is unknown. *Chinese J. Appl. Probab. Statist.*, 9(2):138–145, 1993.

[34] D. Picard. Testing and estimating change points in time series. *J. Appl. Probab.*, 17:841–867, 1985.

[35] F. Picard. *Process segmentation/clustering Application to the analysis of array CGH data*. PhD thesis, Université Paris-Sud 11, 2005.

[36] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J-J. Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 27(6):electronic access, 2005.

[37] Franck Picard, Stéphane Robin, Émilie Lebarbier, and Jean-Jacques Daudin. A segmentation/clustering model for the analysis of array cgh data. *Biometrics*, 2007. To appear. doi:10.1111/j.1541-0420.2006.00729.x.

[38] J. Rissanen. Universal Prior for Integers and Estimation by Minimum Description Length. *The Annals of Statistics*, 11(2):416–431, 1983.

[39] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[40] J. Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

[41] R. Shibata. An optimal selection of regression variables. *Biometrika*, 68:45–54, 1981.

[42] C.J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297, 1984.

[43] M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974.

[44] M. Stone. An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike's Criterion. *JRSS B*, 39(1):44–47, 1977.

[45] R. Tibshirani and K. Knight. The Covariance Inflation Criterion for Adaptive Model Selection. *JRSS B*, 61(3):529–546, 1999.

[46] Y. Yang. Regression with multiple candidate model: selection or mixing? *Statist. Sinica*, 13:783–809, 2003.

[47] Y. Yang. Comparing Learning Methods for Classification. *Statistica Sinica*, 16:635–657, 2006.

[48] Y. Yang. Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.

[49] Y. Yao. Estimating the number of change-points via Schwarz criterion. *Statist. Probab. Lett.*, 6:181–189, 1988.

arXiv:0902.3977v2 [stat.ME] 8 Apr 2009

Manuscrit d'auteur / Author manuscript

Manuscrit d'auteur / Author manuscript

Manuscrit d'auteur / Author manuscript

Postprint

Version définitive du manuscrit publié dans / Final version of the manuscript published in :

Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Supplementary material for Segmentation of the mean of heteroscedastic data via cross-validation

Sylvain Arlot and Alain Celisse

April 8, 2009

1 Calibration of Birgé and Massart's penalization

Birgé and Massart's penalization makes use of the penalty

$$\mathrm{pen}_{\mathrm{BM}}(D) := \frac{\widehat{C} D}{n} \left(5 + 2 \log \left(\frac{n}{D} \right) \right) \quad .$$

In a previous version of this work [6, Chapter 7], $\widehat{C}$ was defined as suggested in [7, 8], that is, $\widehat{C} = 2\widehat{K}_{\mathrm{max.jump}}$ with the notation below. This yielded poor performances, which seemed related to the definition of $\widehat{C}$. Therefore, alternative definitions for $\widehat{C}$ have been investigated, leading to the choice $\widehat{C} = 2\widehat{K}_{\mathrm{thresh.}}$ throughout the paper, where $\widehat{K}_{\mathrm{thresh.}}$ is defined by (2) below. The present appendix intends to motivate this choice.

Two main approaches have been considered in the literature for defining $\widehat{C}$ in the penalty $\mathrm{pen}_{\mathrm{BM}}$:

- Use $\widehat{C} = \widehat{\sigma^2}$ any estimate of the noise-level, for instance,

$$\widehat{\sigma^2} := \frac{1}{n} \sum_{i=1}^{n/2} \left(Y_{2i} - Y_{2i-1} \right)^2 \quad , \tag{1}$$

  assuming n is even and $t_1 < \cdots < t_n$.

- Use Birgé and Massart's *slope heuristics*, that is, compute the sequence

$$\widehat{D}(K) := \arg\min_{D \in \mathcal{D}_n} \left\{ P_n \gamma(\widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)}) + \frac{KD}{n} \left(5 + 2 \log \left(\frac{n}{D} \right) \right) \right\} \quad ,$$

  find the (unique) $K = \widehat{K}_{\mathrm{jump}}$ at which $\widehat{D}(K)$ jumps from large to small values, and define $\widehat{C} = 2\widehat{K}_{\mathrm{jump}}$.

The first approach follows from theoretical and experimental results [4, 8] which show that $\widehat{C}$ should be close to σ^2 when the noise-level is constant; (1) is a classical estimator of the variance used for instance by Baraud [3] for model selection in a different setting.

The optimality (in terms of oracle inequalities) of the second approach has been proved for regression with homoscedastic Gaussian noise and possibly exponential collections of

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

$s.$	$\sigma.$	$2\widehat{K}_{\text{max.jump}}$	$2\widehat{K}_{\text{thresh.}}$	$\widehat{\sigma^2}$	σ^2_{true}
1	c	6.85 ± 0.12	3.91 ± 0.03	1.74 ± 0.02	2.05 ± 0.02
	pc,3	17.56 ± 0.15	13.08 ± 0.04	4.42 ± 0.04	10.43 ± 0.05
	s	20.07 ± 0.31	9.41 ± 0.04	2.18 ± 0.03	1.66 ± 0.02
2	c	6.02 ± 0.03	5.27 ± 0.03	3.58 ± 0.02	3.54 ± 0.02
	pc,3	17.76 ± 0.10	20.12 ± 0.07	10.58 ± 0.07	16.64 ± 0.08
	s	10.17 ± 0.05	9.69 ± 0.04	5.28 ± 0.03	10.95 ± 0.02
3	c	4.97 ± 0.02	4.39 ± 0.01	4.62 ± 0.01	4.21 ± 0.01
	pc,3	8.66 ± 0.03	8.47 ± 0.03	6.64 ± 0.02	8.00 ± 0.03
	s	8.50 ± 0.04	7.59 ± 0.03	5.94 ± 0.02	15.50 ± 0.04
A		7.52 ± 0.04	6.82 ± 0.03	4.86 ± 0.03	5.55 ± 0.03
B		7.89 ± 0.04	7.21 ± 0.04	5.18 ± 0.03	5.77 ± 0.03
C		12.81 ± 0.08	13.49 ± 0.07	8.93 ± 0.06	12.44 ± 0.07

Table 1: Performance $C_{\text{or}}(\text{BM})$ with four different definitions of $\widehat{C}$ (see text), in some of the simulation settings considered in the paper. In each setting, $N = 10\,000$ independent samples have been generated. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$.

models [5], as well as in a heteroscedastic framework with polynomial collections of models [2]. In the context of change-point detection with homoscedastic data, Lavielle [7] and Lebarbier [8] showed that $\widehat{C} = 2\widehat{K}_{\text{max.jump}}$ can even perform better than $\widehat{C} = \sigma^2$ when $\widehat{K}_{\text{max.jump}}$ corresponds to the highest jump of $\widehat{D}(K)$.

Alternatively, it was proposed in [2] to define $\widehat{C} = 2\widehat{K}_{\text{thresh.}}$ where

$$\widehat{K}_{\text{thresh.}} := \min\left\{K \text{ s.t. } \widehat{D}(K) \leq D_{\text{thresh.}} := \left\lfloor \frac{n}{\ln(n)} \right\rfloor\right\} \quad . \tag{2}$$

These three definitions of $\widehat{C}$ have been compared with $\widehat{C} = \sigma^2_{\text{true}} := n^{-1}\sum_{i=1}^n \sigma(t_i)^2$ in the settings of the paper. A representative part of the results is reported in Table 1. The main conclusions are the following.

- $2\widehat{K}_{\text{thresh.}}$ almost always beats $2\widehat{K}_{\text{max.jump}}$, even in homoscedastic settings. This confirms some simulation results reported in [2].
- σ^2_{true} often beats slope heuristics-based definitions of $\widehat{C}$, but not always, as previously noticed by Lebarbier [8]. Differences of performance can be huge (in particular when $\sigma = \sigma_s$), but not always in favour of σ^2_{true} (for instance, when $s = s_3$).
- $\widehat{\sigma^2}$ yields significantly better performance than σ^2_{true} in most settings (but not all), with huge margins in some heteroscedastic settings.

The latter result actually comes from an artefact, which can be explained as follows. First,

$$\mathbb{E}\left[\widehat{\sigma^2}\right] = \frac{1}{n}\sum_{i=1}^n \sigma(t_i)^2 + \frac{1}{n}\sum_{i=1}^n \left(s(t_{2i}) - s(t_{2i-1})\right)^2 \geq \frac{1}{n}\sum_{i=1}^n \sigma(t_i)^2 = \sigma^2_{\text{true}} \quad .$$

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

The difference between these expectations is not negligible in all the settings of the paper. For instance, when $n = 100$, $t_i = i/n$ and $s = s_1$, $n^{-1}\sum_i (s(t_{2i}) - s(t_{2i-1}))^2 = 0.04$ whereas σ^2_{true} varies between 0.015 (when $\sigma = \sigma_{pc,1}$) to 0.093 (when $\sigma = \sigma_{pc,3}$). Nevertheless, $\widehat{\sigma^2}$ would not overestimate σ^2_{true} at all in a very close setting: Shifting the jumps of s_1 by 1/100 is sufficient to make $n^{-1}\sum_i (s(t_{2i}) - s(t_{2i-1}))^2$ equal to zero, and the performances of BM with $\widehat{C} = \widehat{\sigma^2}$ would then be very close to the performances of BM with $\widehat{C} = \sigma_{\text{true}}$.

Second, overpenalization turns out to improve the results of BM in most of the heteroscedastic settings considered in the paper. The reason for this phenomenon is illustrated by the right panel of Figure 4. Indeed, pen_{BM} is a poor penalty when data are heteroscedastic, underpenalizing dimensions close to the oracle but overpenalizing the largest dimensions (remember that $\widehat{C} = 2\widehat{K}_{\text{thresh.}}$ on Figure 4). Then, in a setting like $(s_2, \sigma_{pc,3})$ multiplying pen_{BM} by a factor $C_{\text{over}} > 1$ helps decreasing the selected dimension; the same cause has different consequences in other settings, such as $(s_1, \sigma_s$ or (s_3, σ_c). Nevertheless, even choosing $\widehat{C}$ using both P_n and s, $(\text{crit}_{\text{BM}}(D))_{D>0}$ remains a poor estimate of $\left(\left\|s - \widehat{s}_{\widehat{m}_{\text{ERM}}(D)}\right\|_n^2\right)_{D>0}$ in most heteroscedastic settings (even up to an additive constant).

To conclude, pen_{BM} with $\widehat{C} = \widehat{\sigma^2}$ is not a reliable change-point detection procedure, and the apparently good performances observed in Table 1 could be misleading. This leads to the remaining choice $\widehat{C} = 2\widehat{K}_{\text{thresh.}}$ which has been used throughout the paper, although this calibration method may certainly be improved.

Results of Table 1 for $\widehat{C} = \sigma^2_{\text{true}}$ indicate how far the performances of pen_{BM} could be improved without overpenalization. According to Tables 4 and 5, BM with $\widehat{C} = \sigma^2_{\text{true}}$ only has significantly better performances than $[\![\text{ERM}, \text{VF}_5]\!]$ or $[\![\text{Loo}, \text{VF}_5]\!]$ in the three homoscedastic settings and in setting (s_1, σ_s).

Finally, overpenalization could be used to improve BM, but choosing the overpenalization factor from data is a difficult problem, especially without knowing *a priori* whether the signal is homoscedastic or heteroscedastic. This question deserves a specific extensive simulation experiment. To be completely fair with CV methods, such an experiment should also compare BM with overpenalization to V-fold penalization [1] with overpenalization, for choosing the number of change-points.

2 Random frameworks generation

The purpose of this appendix is to detail how piecewise constant functions s and σ have been generated in the frameworks A, B and C of Section 5.3. In each framework, s and σ are of the form

$$s(x) = \sum_{j=0}^{K_s-1} \alpha_j \mathbb{1}_{[a_j;a_{j+1})} + \alpha_{K_s}\mathbb{1}_{[a_{K_s};a_{K_s+1}]} \qquad \text{with } a_0 = 0 < a_1 < \cdots < a_{K_s} = 1$$

$$\sigma(x) = \sum_{j=0}^{K_\sigma-1} \beta_j \mathbb{1}_{[b_j;b_{j+1})} + \beta_{K_\sigma}\mathbb{1}_{[b_{K_\sigma};b_{K_\sigma+1}]} \qquad \text{with } b_0 = 0 < b_1 < \cdots < b_{K_\sigma} = 1$$

for some positive integers K_s, K_σ and real numbers $\alpha_0, \ldots, \alpha_{K_s} \in \mathbb{R}$ and $\beta_0, \ldots, \beta_{K_\sigma} > 0$.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

Remark 1. The frameworks A, B and C depend on the sample size n, through the distribution of K_s, K_σ, and of the size of the intervals $[a_j; a_{j+1})$ and $[b_j; b_{j+1})$. This ensures that the signal-to-noise ratio remains rather small, so that the quadratic risk remains an adequate performance measure for change-point detection.

When the signal-to-noise ratio is larger (that is, when all jumps of s are much larger than the noise-level, and the number of jumps of s is small compared to the sample size), the change-point detection problem is of different nature. In particular, the number of change-points would be better estimated with procedures targeting identification (such as BIC, or even larger penalties) than efficiency (such as VFCV).

2.1 Framework A

In framework A, s and σ are generated as follows:

- K_s, the number of jumps of s, has uniform distribution over $\{3, \ldots, \lfloor\sqrt{n}\rfloor\}$.
- For $0 \leq j \leq K_s$,
$$a_{j+1} - a_j = \Delta^s_{\min} + \frac{(1-(K_s+1)\Delta^s_{\min})U_j}{\sum_{k=0}^{K_s} U_k}$$
with $\Delta^s_{\min} = \min\{5/n, 1/(K_s+1)\}$ and $U_0, \ldots, U_{K_s}$ are i.i.d. with uniform distribution over $[0; 1]$.
- $\alpha_0 = V_0$ and for $1 \leq j \leq K_s$, $\alpha_j = \alpha_{j-1} + V_j$ where $V_0, \ldots, V_{K_s}$ are i.i.d. with uniform distribution over $[-1; -0.1] \cup [0.1; 1]$.
- K_σ, the number of jumps of σ, has uniform distribution in $\{5, \ldots, \lfloor\sqrt{n}\rfloor\}$.
- For $0 \leq j \leq K_\sigma$,
$$b_{j+1} - b_j = \Delta^\sigma_{\min} + \frac{(1-(K_\sigma+1)\Delta^\sigma_{\min})U'_j}{\sum_{k=0}^{K_s} U'_k}$$
with $\Delta^\sigma_{\min} = \min\{5/n, 1/(K_\sigma+1)\}$ and $U'_0, \ldots, U'_{K_\sigma}$ are i.i.d. with uniform distribution over $[0; 1]$.
- $\beta_0, \ldots, \beta_{K_\sigma}$ are i.i.d. with uniform distribution over $[0.05; 0.5]$.

Two examples of a function s and a sample (t_i, Y_i) generated in framework A are plotted on Figure 1.

2.2 Framework B

The only difference with framework A is that $U_0, \ldots, U_{K_s}$ are i.i.d. with the same distribution as $Z = |10Z_1 + Z_2|$ where Z_1 has Bernoulli distribution with parameter $1/2$ and Z_2 has a standard Gaussian distribution. Two examples of a function s and a sample (t_i, Y_i) generated in framework B are plotted on Figure 2.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript

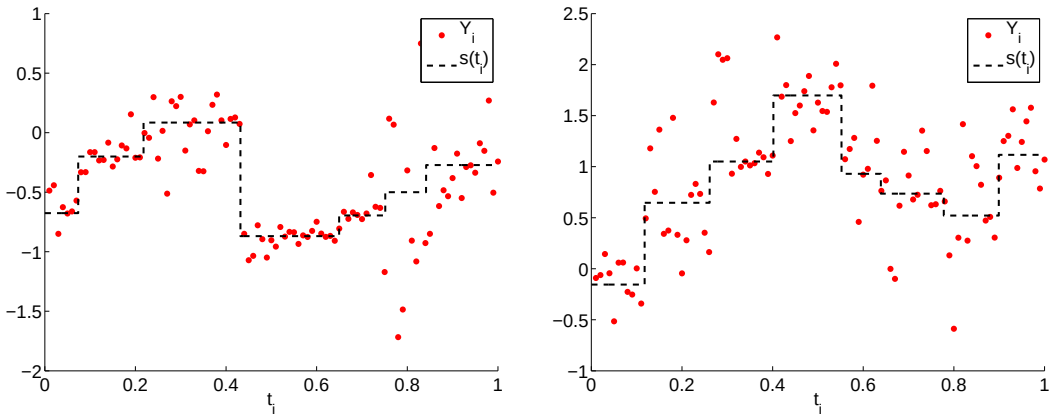


Figure 1: Random framework A: two examples of a sample $(t_i, Y_i)_{1 \leq i \leq 100}$ and the corresponding regression function s.

Manuscrit d'auteur / Author manuscript

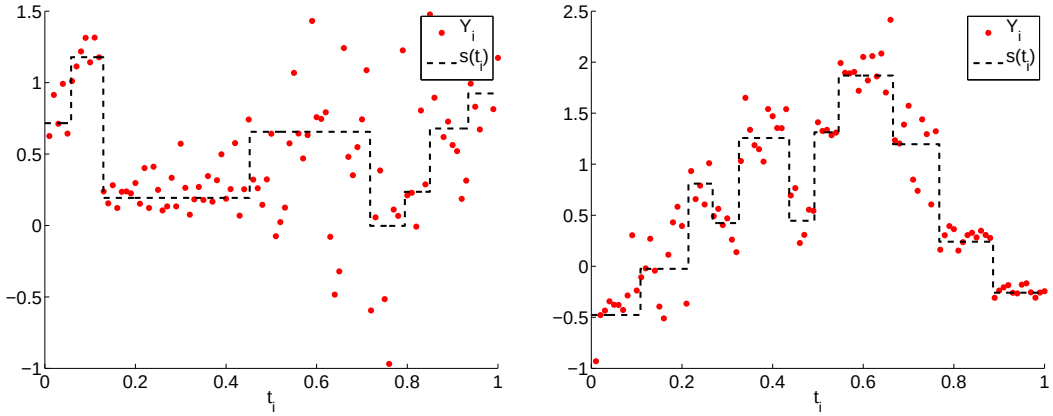


Figure 2: Random framework B: two examples of a sample $(t_i, Y_i)_{1 \leq i \leq 100}$ and the corresponding regression function s.

Manuscrit d'auteur / Author manuscript

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

2.3 Framework C

The main difference between frameworks C and B is that $[0;1]$ is split into two regions: $a_{K_{s,1}+1} = 1/2$ and $K_s = K_{s,1} + K_{s,2} + 1$ for some positive integers $K_{s,1}, K_{s,2}$, and the bounds of the distribution of β_j are larger when $b_j \geq 1/2$ and smaller when $b_j < 1/2$. Two examples of a function s and a sample (t_i, Y_i) generated in framework C are plotted on Figure 3. More precisely, s and σ are generated as follows:

- $K_{s,1}$ has uniform distribution over $\{2, \ldots, K_{\max,1}\}$ with $K_{\max,1} = \lfloor\sqrt{n}\rfloor - 1 - \lfloor(\lfloor\sqrt{n} - 1\rfloor)/3\rfloor$.
- $K_{s,2}$ has uniform distribution over $\{0, \ldots, K_{\max,2}\}$ with $K_{\max,2} = \lfloor(\lfloor\sqrt{n} - 1\rfloor)/3\rfloor$.
- Let $U_0, \ldots, U_{K_s}$ be i.i.d. random variables with the same distribution as $Z = |10Z_1 + Z_2|$ where Z_1 has Bernoulli distribution with parameter $1/2$ and Z_2 has a standard Gaussian distribution.
- For $0 \leq j \leq K_{s,1}$,

$$a_{j+1} - a_j = \Delta^{s,1}_{\min} + \frac{(1 - (K_{s,1} + 1)\Delta^{s,1}_{\min})U_j}{\sum_{k=0}^{K_{s,1}} U_k}$$

  with $\Delta^{s,1}_{\min} = \min\{5/n, 1/(K_{s,1} + 1)\}$.
- For $K_{s,1} + 1 \leq j \leq K_s$,

$$a_{j+1} - a_j = \Delta^{s,2}_{\min} + \frac{(1 - (K_{s,2} + 1)\Delta^{s,2}_{\min})U_j}{\sum_{k=K_{s,1}+1}^{K_s} U_k}$$

  with $\Delta^{s,2}_{\min} = \min\{5/n, 1/(K_{s,2} + 1)\}$.
- $\alpha_0 = V_0$ and for $1 \leq j \leq K_s$, $\alpha_j = \alpha_{j-1} + V_j$ where $V_0, \ldots, V_{K_s}$ are i.i.d. with uniform distribution over $[-1; -0.1] \cup [0.1; 1]$.
- K_σ, $(b_{j+1} - b_j)_{0 \leq j \leq K_\sigma}$ are distributed as in frameworks A and B.
- $\beta_0, \ldots, \beta_{K_\sigma}$ are independent.
  When $b_j < 1/2$, β_j has uniform distribution over $[0.025; 0.2]$.
  When $b_j \geq 1/2$, β_j has uniform distribution over $[0.1; 0.8]$.

3 Additional results from the simulation study

In the next pages are presented extended versions of the Tables of the main paper, as well as an extended version of Table 1 (Table 7).

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript
Manuscrit d'auteur / Author manuscript

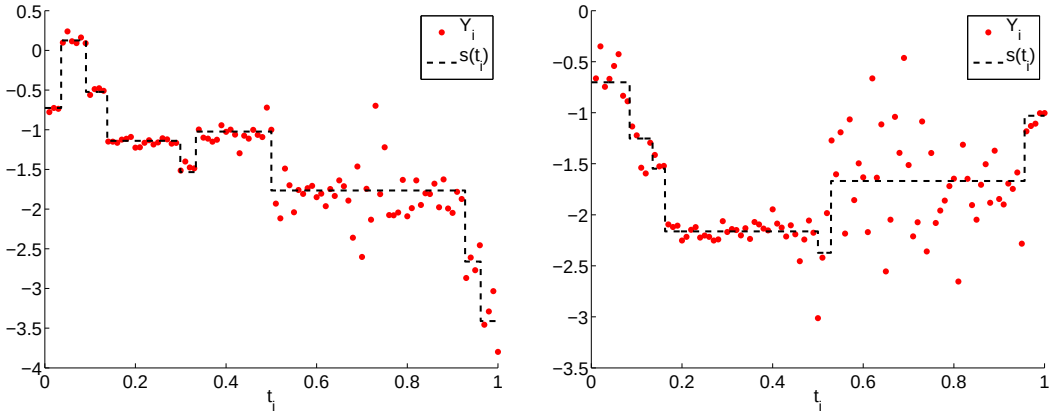


Figure 3: Random framework C: two examples of a sample $(t_i, Y_i)_{1 \leq i \leq 100}$ and the corresponding regression function s.

References

[1] Sylvain Arlot. V-fold cross-validation improved: V-fold penalization, February 2008. arXiv:0802.0566v2.

[2] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.

[3] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.

[4] L. Birgé and P. Massart. Gaussian model selection. *J. European Math. Soc.*, 3(3):203–268, 2001.

[5] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.

[6] Alain Celisse. *Model selection via cross-validation in density estimation, regression and change-points detection.* PhD thesis, University Paris-Sud 11, December 2008. oai:tel.archives-ouvertes.fr:tel-00346320_v1.

[7] M. Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85:1501–1510, 2005.

[8] E. Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

s.	σ.	ERM	Loo	Lpo_{20}	Lpo_{50}
1	c	**1.59** ± 0.01	1.60 ± 0.02	**1.58** ± 0.01	**1.58** ± 0.01
	pc,1	**1.04** ± 0.01	1.06 ± 0.01	1.06 ± 0.01	1.06 ± 0.01
	pc,2	**1.89** ± 0.02	**1.87** ± 0.02	**1.87** ± 0.02	**1.87** ± 0.02
	pc,3	**2.05** ± 0.02	**2.05** ± 0.02	**2.05** ± 0.02	**2.07** ± 0.02
	s	1.54 ± 0.02	**1.52** ± 0.02	**1.52** ± 0.02	**1.51** ± 0.02
2	c	**2.88** ± 0.01	2.93 ± 0.01	2.93 ± 0.01	2.94 ± 0.01
	pc,1	1.31 ± 0.02	1.16 ± 0.02	1.14 ± 0.02	**1.11** ± 0.01
	pc,2	2.88 ± 0.02	2.24 ± 0.02	2.19 ± 0.02	**2.13** ± 0.02
	pc,3	3.09 ± 0.03	2.52 ± 0.03	2.48 ± 0.03	**2.32** ± 0.03
	s	**3.01** ± 0.01	**3.03** ± 0.01	3.05 ± 0.01	3.13 ± 0.01
3	c	**3.18** ± 0.01	3.25 ± 0.01	3.29 ± 0.01	3.44 ± 0.01
	pc,1	3.00 ± 0.01	**2.67** ± 0.02	**2.68** ± 0.02	2.77 ± 0.02
	pc,2	4.06 ± 0.02	**3.63** ± 0.02	**3.64** ± 0.02	3.78 ± 0.02
	pc,3	4.41 ± 0.02	**3.97** ± 0.02	4.00 ± 0.02	4.11 ± 0.02
	s	4.02 ± 0.01	**3.82** ± 0.01	3.85 ± 0.01	3.98 ± 0.01

Table 2: Average performance $C_{\mathrm{or}}\left(\llbracket \mathfrak{P}, \mathrm{Id} \rrbracket\right)$ for change-point detection procedures $\mathfrak{P}$ among ERM, Loo and Lpo_p with $p = 20$ and $p = 50$. Several regression functions s and noise-level functions σ have been considered, each time with $N = 10\,000$ independent samples. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$, measuring the uncertainty of the estimated performance.

s.	σ.	Oracle	VF_5	BM
1	c	1.59 ± 0.01	5.40 ± 0.05	**3.91** ± 0.03
	pc,1	1.04 ± 0.01	**11.96** ± 0.03	12.85 ± 0.04
	pc,2	1.89 ± 0.02	**6.43** ± 0.05	13.03 ± 0.04
	pc,3	2.05 ± 0.02	**4.96** ± 0.05	13.08 ± 0.04
	s	1.54 ± 0.02	**7.33** ± 0.06	9.41 ± 0.04
2	c	2.88 ± 0.01	**4.51** ± 0.03	5.27 ± 0.03
	pc,1	1.31 ± 0.02	**11.67** ± 0.09	19.36 ± 0.07
	pc,2	2.88 ± 0.02	**6.58** ± 0.06	19.82 ± 0.07
	pc,3	3.09 ± 0.03	**6.66** ± 0.06	20.12 ± 0.07
	s	3.01 ± 0.01	**5.21** ± 0.04	9.69 ± 0.40
3	c	3.18 ± 0.01	4.41 ± 0.02	**4.39** ± 0.01
	pc,1	3.00 ± 0.01	**4.91** ± 0.02	6.50 ± 0.02
	pc,2	4.06 ± 0.02	**5.99** ± 0.02	7.86 ± 0.03
	pc,3	4.41 ± 0.02	**6.32** ± 0.02	8.47 ± 0.03
	s	4.02 ± 0.01	**5.97** ± 0.03	7.59 ± 0.03

Table 3: Performance $C_{\mathrm{or}}\left(\llbracket \mathrm{ERM}, \mathfrak{P} \rrbracket\right)$ for $\mathfrak{P} = \mathrm{Id}$ (that is, choosing the dimension $D^\star := \arg\min_{D \in \mathcal{D}_n} \left\{ \left\| s - \widehat{s}_{\widehat{m}_{\mathrm{ERM}}(D)} \right\|_n^2 \right\}$), $\mathfrak{P} = \mathrm{VF}_V$ with $V = 5$ or $\mathfrak{P} = \mathrm{BM}$. Several regression functions s and noise-level functions σ have been considered, each time with $N = 10\,000$ independent samples. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$, measuring the uncertainty of the estimated performance.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript Manuscrit d'auteur / Author manuscript

$s.$	$\sigma.$	$[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$	$[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$	$[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$	$[\![\mathrm{Lpo}_{50}, \mathrm{VF}_5]\!]$	$[\![\mathrm{ERM}, \mathrm{BM}]\!]$
1	c	5.40 ± 0.05	5.03 ± 0.05	5.10 ± 0.05	5.24 ± 0.05	**3.91** ± 0.03
	pc,1	11.96 ± 0.03	**10.25** ± 0.03	10.28 ± 0.03	10.66 ± 0.04	12.85 ± 0.04
	pc,2	6.43 ± 0.05	**5.83** ± 0.05	5.99 ± 0.05	6.20 ± 0.05	13.03 ± 0.04
	pc,3	4.96 ± 0.05	**4.82** ± 0.04	**4.79** ± 0.05	5.02 ± 0.05	13.08 ± 0.04
	s	7.33 ± 0.06	**6.82** ± 0.05	6.99 ± 0.06	**6.91** ± 0.06	9.41 ± 0.04
2	c	**4.51** ± 0.03	**4.55** ± 0.03	**4.50** ± 0.03	4.73 ± 0.03	5.27 ± 0.03
	pc,1	11.67 ± 0.09	**10.26** ± 0.08	**10.29** ± 0.08	10.45 ± 0.09	19.36 ± 0.07
	pc,2	6.58 ± 0.06	5.85 ± 0.06	5.85 ± 0.06	**5.49** ± 0.06	19.82 ± 0.07
	pc,3	6.66 ± 0.06	5.81 ± 0.06	**5.74** ± 0.06	**5.66** ± 0.06	20.12 ± 0.06
	s	**5.21** ± 0.04	**5.19** ± 0.03	**5.17** ± 0.03	5.51 ± 0.04	9.69 ± 0.04
3	c	4.41 ± 0.02	4.54 ± 0.02	4.62 ± 0.02	4.94 ± 0.02	**4.39** ± 0.01
	pc,1	4.91 ± 0.02	**4.40** ± 0.02	**4.44** ± 0.02	4.69 ± 0.02	6.50 ± 0.02
	pc,2	5.99 ± 0.02	**5.34** ± 0.02	5.42 ± 0.02	5.75 ± 0.02	7.86 ± 0.03
	pc,3	6.32 ± 0.02	**5.74** ± 0.02	5.81 ± 0.02	6.24 ± 0.02	8.47 ± 0.03
	s	5.97 ± 0.02	**5.72** ± 0.02	5.86 ± 0.02	6.07 ± 0.02	7.59 ± 0.03

Table 4: Performance $C_{\mathrm{or}}(\mathfrak{P})$ for several change-point detection procedures $\mathfrak{P}$. Several regression functions s and noise-level functions σ have been considered, each time with $N = 10\,000$ independent samples. Next to each value is indicated the corresponding empirical standard deviation.

Framework	A	B	C
$[\![\mathrm{ERM}, \mathrm{BM}]\!]$	6.82 ± 0.03	7.21 ± 0.04	13.49 ± 0.07
$[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$	4.78 ± 0.03	5.09 ± 0.03	7.17 ± 0.05
$[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$	**4.65** ± 0.03	**4.88** ± 0.03	6.61 ± 0.05
$[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$	4.78 ± 0.03	**4.91** ± 0.03	**6.49** ± 0.05
$[\![\mathrm{Lpo}_{50}, \mathrm{VF}_5]\!]$	4.97 ± 0.03	5.18 ± 0.04	6.69 ± 0.05

Table 5: Performance $C_{\mathrm{or}}^{(R)}(\mathfrak{P})$ of several model selection procedures $\mathfrak{P}$ in frameworks A, B, C with sample size $n = 100$. In each framework, $N = 10,000$ independent samples have been considered. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$.

Postprint
Version définitive du manuscrit publié dans / Final version of the manuscript published in :
Statistics and Computing, 2010, vol. 21, n° 4, 613-632, 10.1007/s11222-010-9196-x

Manuscrit d'auteur / Author manuscript

Framework	A	B	C
$[\![\mathrm{ERM}, \mathrm{BM}]\!]$	9.04 ± 0.12	11.62 ± 0.14	21.21 ± 0.31
$[\![\mathrm{ERM}, \mathrm{BM}_{\widehat{\sigma}}]\!]$	5.34 ± 0.10	6.24 ± 0.11	11.48 ± 0.22
$[\![\mathrm{ERM}, \mathrm{VF}_5]\!]$	5.10 ± 0.11	5.92 ± 0.11	7.31 ± 0.14
$[\![\mathrm{Loo}, \mathrm{VF}_5]\!]$	**4.90** ± 0.11	**5.63** ± 0.11	**6.89** ± 0.16
$[\![\mathrm{Lpo}_{20}, \mathrm{VF}_5]\!]$	**4.88** ± 0.10	**5.55** ± 0.10	**6.82** ± 0.15
$[\![\mathrm{Lpo}_{50}, \mathrm{VF}_5]\!]$	5.11 ± 0.11	**5.49** ± 0.10	7.14 ± 0.15

Table 6: Performance $C_{\mathrm{or}}^{(R)}(\mathfrak{P})$ of several model selection procedures $\mathfrak{P}$ in frameworks A, B, C with sample size $n = 200$. In each framework, $N = 1,000$ independent samples have been considered. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$.

Manuscrit d'auteur / Author manuscript

$s.$	$\sigma.$	$2\widehat{K}_{\mathrm{max.jump}}$	$2\widehat{K}_{\mathrm{thresh.}}$	$\widehat{\sigma^2}$	σ^2_{true}
1	c	6.85 ± 0.12	3.91 ± 0.03	1.74 ± 0.02	2.05 ± 0.02
	pc,1	70.97 ± 1.18	12.85 ± 0.04	1.13 ± 0.02	10.20 ± 0.05
	pc,2	23.74 ± 0.26	13.03 ± 0.04	3.55 ± 0.04	10.43 ± 0.05
	pc,3	17.56 ± 0.15	13.08 ± 0.04	4.42 ± 0.04	10.43 ± 0.05
	s	20.07 ± 0.31	9.41 ± 0.04	2.18 ± 0.03	1.66 ± 0.02
2	c	6.02 ± 0.03	5.27 ± 0.03	3.58 ± 0.02	3.54 ± 0.02
	pc,1	17.83 ± 0.10	19.36 ± 0.07	8.52 ± 0.06	15.62 ± 0.08
	pc,2	17.63 ± 0.10	19.82 ± 0.07	10.77 ± 0.07	16.56 ± 0.08
	pc,3	17.76 ± 0.10	20.12 ± 0.07	10.58 ± 0.07	16.64 ± 0.08
	s	10.17 ± 0.05	9.69 ± 0.04	5.28 ± 0.03	10.95 ± 0.02
3	c	4.97 ± 0.02	4.39 ± 0.01	4.62 ± 0.01	4.21 ± 0.01
	pc,1	7.18 ± 0.03	6.50 ± 0.02	4.52 ± 0.02	6.70 ± 0.03
	pc,2	8.14 ± 0.03	7.86 ± 0.03	6.22 ± 0.02	7.55 ± 0.03
	pc,3	8.66 ± 0.03	8.47 ± 0.03	6.64 ± 0.02	8.00 ± 0.03
	s	8.50 ± 0.04	7.59 ± 0.03	5.94 ± 0.02	15.50 ± 0.04
A		7.52 ± 0.04	6.82 ± 0.03	4.86 ± 0.03	5.55 ± 0.03
B		7.89 ± 0.04	7.21 ± 0.04	5.18 ± 0.03	5.77 ± 0.03
C		12.81 ± 0.08	13.49 ± 0.07	8.93 ± 0.06	12.44 ± 0.07

Table 7: Performance $C_{\mathrm{or}}(\mathrm{BM})$ with four different definitions of $\widehat{C}$ (see text), in some of the simulation settings considered in the paper. In each setting, $N = 10\,000$ independent samples have been generated. Next to each value is indicated the corresponding empirical standard deviation divided by $\sqrt{N}$.

Manuscrit d'auteur / Author manuscript