

Segmentation of the mean of heteroscedastic data via cross-validation

Sylvain Arlot and Alain Celisse

September 20, 2010

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 partial 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 goal is to recover the location of these changes and their number. This problem is motivated by a wide range of applications, such as audio processing [32], financial time-series analysis [36] and Comparative Genomic Hybridization (CGH) data analysis [44]. A large literature exists about change-point detection in many frameworks [see 14, 19, 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 43, for instance]. Looking for multiple change-points is a harder task and has been studied later. For instance, Yao [58] used the BIC criterion for detecting multiple change-points in a Gaussian signal, and Miao and Zhao [41] proposed an approach relying on rank statistics.

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

$$Y_i = s(t_i) + \sigma(t_i)\varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0 \quad \text{and} \quad \text{Var}(\varepsilon_i) = 1. \quad (1)$$

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 change-points. The noise terms $\varepsilon_1, \dots, \varepsilon_n$ are assumed to be independent and identically distributed. No assumption is made on $\sigma : [0, 1] \mapsto [0, \infty)$, except that *it is bounded*. 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 [35], 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 boundaries 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 [35] and by Lebarbier [37] for instance. Indeed, a set of change-points—called a segmentation—is naturally associated with the set of piecewise constant functions that can 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 $\{\hat{s}_m\}_{m \in \mathcal{M}_n}$, the purpose of model selection is to provide a model index \hat{m} such that $\hat{s}_{\hat{m}}$ reaches the “best performance” among all estimators $\{\hat{s}_m\}_{m \in \mathcal{M}_n}$.

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 $C_n \rightarrow 1$ as n tends to infinity. Asymptotic efficiency is the goal of AIC [2, 3] and Mallows' C_p [39], 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 [48] for instance. See also the article by Yang [55] about the distinction between efficiency and model consistency.

In the present paper as in [37], 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 can 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, $\text{Card}(\mathcal{M}_n) \leq Cn^\alpha$ for some constants $C, \alpha \geq 0$. For polynomial collections, procedures like AIC or C_p are proved to satisfy oracle inequalities in various frameworks [10, 17, 11, 18], assuming that data are *homoscedastic*, that is, $\sigma(t_i)$ does not depend on t_i .

However as shown in [7], C_p is suboptimal when data are *heteroscedastic*, that is, when the variance is non-constant. Therefore, other procedures must be used. For instance, resampling penalization is optimal with heteroscedastic data and polynomial collections [6]. Another approach has been explored by Comte and Rozenholc [27] and by Gendre [31], which consists in simultaneously estimating the mean and the variance, using polynomial collections of models.

However in change-point detection, the collection of models is “exponential”, that is, $\text{Card}(\mathcal{M}_n)$ is of order $\exp(\alpha n)$ for some $\alpha > 0$. For such large collections, larger than polynomial, the above penalization procedures fail. Indeed, Birgé and Massart [18] proved that the minimal amount of penalization required for a procedure to satisfy an oracle inequality is of the form

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

where D_m is the dimension of S_m as a vector space, c_1 and c_2 are positive constants, and σ^2 is the variance of the noise, assumed to be constant. Lebarbier [37] 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 [47, 1, 13, 54, 12] and are shown to provide satisfactory results.

Nevertheless, all these results about exponential collections 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 [7] 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 \hat{s}_m (see Section 3.2 about the heuristics underlying CV). Classical examples of CV methods are the leave-one-out [Loo, 34, 52] and V -fold cross-validation [VFCV, 29, 30]. More references on cross-validation can be found in the survey [8].

CV can be used for model selection, by choosing the model S_m for which the CV estimate of the risk of \hat{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 [51, 24, 25] to regression [53, 57] and classification [33, 56]. In particular, Loo is asymptotically equivalent to AIC or 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 38, 49, 28]. In addition, CV methods are robust to heteroscedasticity of data [5], 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 [16, 49, 18, 23] 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 [18] about a penalization procedure is used for solving this issue in the context of change-point detection. Model selection is performed 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. A simulation study and partial theoretical results (Proposition 1) show that this procedure seems to be robust to a possible heteroscedasticity of data 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 [17] when data can be heteroscedastic.

Another major difficulty tackled in this paper is the computational cost of resampling methods when selecting among 2^{n-1} models. Even when the number $(D - 1)$ of change-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 $D > 2$. 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 change-points 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. Partial 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 seem to be robust to a possible heteroscedasticity of data.

Then, the problem of choosing the number of change-points 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 that 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 $t_1 < \dots < t_n \in [0, 1]$ be some deterministic design points, $s \in \mathcal{S}^*$ and $\sigma : [0, 1] \mapsto [0, \infty)$ be some functions and

define

$$\forall i \in \{1, \dots, n\}, \quad Y_i = s(t_i) + \sigma(t_i)\varepsilon_i, \quad (3)$$

where $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed random variables with $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = 1$.

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

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

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

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

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 S is

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

The notation $\text{ERM}(S; P_n)$ 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 $\text{ERM}(S_m; P_n)$.

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 $D \in \{2, \dots, n\}$ and every sequence of integers $\alpha_0 = 1 < \alpha_1 < \alpha_2 < \dots < \alpha_{D-1} \leq n$ (the change-points), $\Lambda_{(\alpha_1, \dots, \alpha_{D-1})}$ denotes the partition

$$[0; t_{\alpha_1}), \dots, [t_{\alpha_{D-2}}; t_{\alpha_{D-1}}), [t_{\alpha_{D-1}}; 1]$$

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

For every $D \in \{2, \dots, n\}$, let $\overline{\mathcal{M}}_n(D)$ denote the set of such sequences $(\alpha_1, \dots, \alpha_{D-1})$ of length $(D-1)$, so that $\{S_m\}_{m \in \overline{\mathcal{M}}_n(D)}$ is the collection of models of piecewise constant functions with $(D-1)$ change-points. When $D = 1$, $\overline{\mathcal{M}}_n(1) := \{\emptyset\}$ where $\Lambda_\emptyset = \{[0; 1]\}$ and the model S_\emptyset is the linear space of constant functions on $[0, 1]$. Remark that for every D and $m \in \overline{\mathcal{M}}_n(D)$, S_m is a vector space of dimension $D_m = D$. Therefore, estimating

the number $(D - 1)$ of change-points (Section 4) is equivalent to choosing the dimension D of a model.

The classical collection of models for change-point detection can now be defined as $\{S_m\}_{m \in \overline{\mathcal{M}}_n}$, where $\overline{\mathcal{M}}_n = \bigcup_{D \in \mathcal{D}_n} \overline{\mathcal{M}}_n(D)$ 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 used. We only consider models S_m such that each element of the partition Λ_m contains at least two design points $(t_j)_{1 \leq j \leq n}$. 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 $m \in \overline{\mathcal{M}}_n(D)$ 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, $\mathcal{D}_n = \{1, \dots, 9n/25\}$ for reasons detailed in Section 4.2 and in Section 3 of the supplementary material.

2.4 Model selection

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

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

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*. Note that when \mathcal{M}_n is exponential, in general, one cannot hope to obtain (4) with $C \ll \ln(n)$ and \widehat{m} completely data-driven (see for instance Corollary 4.12 in [40] in the framework of complete variable selection).

3 Localization of the change-points

A usual strategy for multiple change-point detection [35, 37] is to dissociate the search for the best segmentation given the number of change-points from the choice of the number of change-points.

In this section, the number $(D - 1)$ of change-points 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 exposed by many authors such as Lavielle [35], 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.

$$\widehat{m}_{\text{ERM}}(D) := \arg \min_{m \in \mathcal{M}_n(D)} \{P_n \gamma(\widehat{s}_m)\} ,$$

so that $\widehat{s}_{\widehat{m}_{\text{ERM}}(D)} = \text{ERM}(\widetilde{S}_D; P_n)$ where $\widetilde{S}_D := \bigcup_{m \in \mathcal{M}_n(D)} S_m$

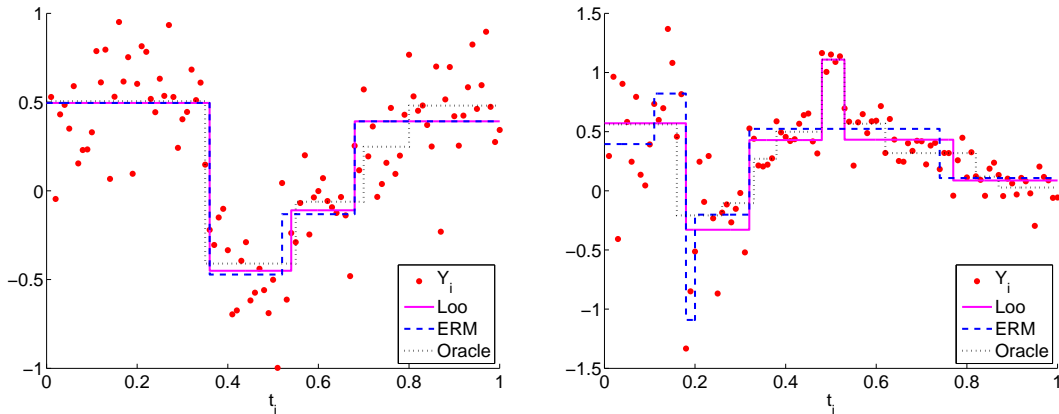


Figure 1: Comparison of $\widehat{s}_{m^*(D)}$ (dotted black line), $\widehat{s}_{\widehat{m}_{\text{ERM}}(D)}$ (dashed blue line) and $\widehat{s}_{\widehat{m}_{\text{Loo}}(D)}$ (plain magenta line, see Section 3.2.2), D being the “optimal” dimension (see Figure 3). Data are generated as described in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

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

Remark 1. Dynamic programming [15] leads to an efficient implementation of Procedure 1 with computational complexity $\mathcal{O}(n^2)$.

Among models corresponding to segmentations with $(D - 1)$ change-points, the oracle model can be defined as $S_{m^*(D)}$ where

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

Figure 1 illustrates how far $\widehat{m}_{\text{ERM}}(D)$ typically is from $m^*(D)$ depending on variations of the standard-deviation σ . On the one hand, when data are homoscedastic, empirical risk minimization yields a segmentation close to the oracle one (Figure 1, left). On the other hand, when data are heteroscedastic, empirical risk minimization introduces artificial change-points in areas where the noise-level is above average, and misses change-points 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 [23, 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 $\widehat{m}_{\text{ERM}}(D)$ and by $m^*(D)$ over $\mathcal{M}_n(D)$, are close to their respective expectations, as proved by the concentration inequalities of [5, Proposition 9] for instance, under assumptions on ε and σ that are discussed at the end of Section 3.2.4. 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,

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

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

where

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

Lemma 1 is proved in [23]; see also Section 1.1 of the supplementary material. 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 S_m , and $V(m)$ is a variance term, measuring the difficulty of estimating the D_m parameters of 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 $\forall i, \sigma(t_i) = \bar{\sigma}$, $V(m) = D_m \bar{\sigma}^2 n^{-1}$ 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

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

Hence, $\widehat{m}_{\text{ERM}}(D)$ and $m^*(D)$ 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 $(S_m)_{m \in \mathcal{M}_n(D)}$, even though they have the same dimension D . Indeed, $V(m)$ increases when a change-point is moved from an area where σ is small to an area where σ is large. Therefore, the empirical risk minimization algorithm rather puts change-points in noisy areas in order to minimize $-V(m)$ in (5); on the contrary, the oracle segmentation $\Lambda_{m^*(D)}$ tends to have more change-points in areas where σ is small, since it minimizes $+V(m)$ in (6). This is illustrated in the right panel of Figure 1.

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, 52] relies on a data splitting idea: For each candidate algorithm—say $\text{ERM}(S_m; \cdot)$ 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 [29, 30].

In the fixed-design setting, $(t_i, Y_i)_{1 \leq i \leq n}$ are not identically distributed so that CV aims at estimating a quantity slightly different from the usual prediction error. Let T be uniformly distributed over $\{t_1, \dots, t_n\}$ and $Y = s(T) + \sigma(T)\varepsilon$, where ε is independent from $\varepsilon_1, \dots, \varepsilon_n$ with the same distribution. Then, the CV estimator of the risk of $\hat{s} = \hat{s}(P_n)$ estimates

$$\begin{aligned} \mathbb{E}_{(T,Y)} \left[(\hat{s}(T) - Y)^2 \right] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}_\varepsilon \left[(s(t_i) + \sigma(t_i)\varepsilon_i - \hat{s}(t_i))^2 \right] \\ &= \|s - \hat{s}\|_n^2 + \frac{1}{n} \sum_{i=1}^n \sigma(t_i)^2 . \end{aligned}$$

Hence, minimizing the CV estimator of $\mathbb{E}_{(T,Y)} \left[(\hat{s}_m(T) - Y)^2 \right]$ over m amounts to minimize $\|s - \hat{s}_m\|_n^2$, 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 $\|s - \hat{s}_m\|_n^2$. 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 $I^{(t)} \subset \{1, \dots, n\}$, define $I^{(v)} := \{1, \dots, n\} \setminus I^{(t)}$,

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

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

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

The *cross-validation estimators of the risk* of \mathcal{A} are then defined as the average of $\hat{R}_{\text{ho}}(\mathcal{A}, P_n, I_j^{(t)})$ over $j = 1, \dots, B$ where $I_1^{(t)}, \dots, I_B^{(t)}$ are chosen in a predetermined way [8]. 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 $I_1^{(t)}, \dots, I_B^{(t)}$.

- *Leave-one-out* (Loo), often called *ordinary CV* [4, 52], consists in training with the whole sample except one point, used for testing, and repeating this for each data point: $I_j^{(t)} = \{1, \dots, n\} \setminus \{j\}$ for $j = 1, \dots, n$. The Loo estimator of the risk of \mathcal{A} is defined by

$$\hat{R}_{\text{Loo}}(\mathcal{A}, P_n) := \frac{1}{n} \sum_{j=1}^n \left[\left(Y_j - \mathcal{A} \left(t_j; P_n^{(-j)} \right) \right)^2 \right] ,$$

where $P_n^{(-j)} = (n-1)^{-1} \sum_{i, i \neq j} \delta_{(t_i, Y_i)}$.

- *Leave-p-out* (Lpo_p , with any $p \in \{1, \dots, n-1\}$) generalizes Loo. Let \mathcal{E}_p denote the collection of all possible subsets of $\{1, \dots, n\}$ with cardinality $n-p$. Then, Lpo consists in considering every $I^{(t)} \in \mathcal{E}_p$ as training set indices:

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

- *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, \dots, B_V be a partition of $\{1, \dots, n\}$ and $P_n^{(\overline{B_k})} := (n - \text{Card}(B_k))^{-1} \sum_{i \notin B_k} \delta_{(t_i, Y_i)}$ for every $k \in \{1, \dots, V\}$. The VFCV estimator of the risk of \mathcal{A} is defined by

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

The interested reader will find theoretical and experimental results on VFCV and the best way to use it in [5, 23] and references therein, in particular [20].

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

Procedure 2.

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

The segmentations $\Lambda_{\widehat{m}_{\text{Lpo}_p}(D)}$ and $\Lambda_{\widehat{m}_{\text{VFCV}}(D)}$, respectively chosen by Lpo and by VFCV, are defined similarly (see Section 2.3 for the definition of Λ_m).

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

3.2.3 Computational tractability

The computational complexity of $\text{ERM}(S_m; P_n)$ is $\mathcal{O}(n)$ since for every interval $\lambda \in \Lambda_m$, the value of $\widehat{m}(P_n)$ on λ is equal to the mean of $\{Y_i\}_{t_i \in \lambda}$. 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 [24, 22] and regression [23] 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 $\widehat{R}_{\text{Lpo}_p}(m) := \widehat{R}_{\text{Lpo}_p}(\text{ERM}(S_m; \cdot), P_n)$ which can be computed with $\mathcal{O}(n)$ elementary operations.

Theorem 1 (Corollary 3.3.2 in [23]). *Let $m \in \mathcal{M}_n$, S_m and $\widehat{s}_m = \text{ERM}(S_m; \cdot)$ be defined as in Section 2. For every $(t_1, Y_1), \dots, (t_n, Y_n) \in \mathbb{R}^2$ and $\lambda \in \Lambda_m$, define*

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

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

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

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

$$\begin{aligned} n_\lambda &:= \text{Card}(\{i \mid t_i \in \lambda\}) & N_\lambda &:= 1 - \mathbf{1}_{\{p \geq n_\lambda\}} \binom{n - n_\lambda}{p - n_\lambda} / \binom{n}{p} \\ A_\lambda &:= V_\lambda(0) \left(1 - \frac{1}{n_\lambda}\right) - \frac{V_\lambda(1)}{n_\lambda} + V_\lambda(-1) \\ B_\lambda &:= V_\lambda(1) \frac{2 - \mathbf{1}_{n_\lambda \geq 3}}{n_\lambda(n_\lambda - 1)} + \frac{V_\lambda(0)}{n_\lambda - 1} \left[\left(1 + \frac{1}{n_\lambda}\right) \mathbf{1}_{n_\lambda \geq 3} - 2 \right] - \frac{V_\lambda(-1) \mathbf{1}_{n_\lambda \geq 3}}{n_\lambda - 1} \\ \text{and } \forall k \in \{-1, 0, 1\}, \quad V_\lambda(k) &:= \sum_{r=\max\{1, (p-n_\lambda)\}}^{\min\{n_\lambda, (n-p)\}} r^k \frac{\binom{n-p}{r} \binom{p}{n_\lambda-r}}{\binom{n}{n_\lambda}} . \end{aligned}$$

Remark 2. $V_\lambda(k)$ can also be written as $\mathbb{E}[Z^k \mathbf{1}_{Z>0}]$ 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 , $\widehat{m}_{\text{Lpo}_p}(D)$ can be computed with the same computational complexity as $\widehat{m}_{\text{ERM}}(D)$, that is $\mathcal{O}(n^2)$. Indeed, Theorem 1 shows that $\widehat{R}_{\text{Lpo}_p}(m)$ is a sum over $\lambda \in \Lambda_m$ of terms depending only on $\{Y_i\}_{t_i \in \lambda}$, so that dynamic programming [15] can be used for computing the minimizer $\widehat{m}_{\text{Lpo}_p}(D)$ of $\widehat{R}_{\text{Lpo}_p}(m)$ over $m \in \mathcal{M}_n$. Therefore, Lpo and Loo are computationally tractable for change-point detection when the number of change-points is given.

Dynamic programming also applies to $\widehat{m}_{\text{VFCV}}$ with a computational complexity $\mathcal{O}(Vn^2)$, since each term appearing in $\widehat{R}_{\text{VFCV}}(m)$ 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, $\widehat{m}_{\text{Lpo}_p}$ will be preferred to $\widehat{m}_{\text{VFCV}}(D)$ 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 change-points, let us recall a consequence of Theorem 1, which can also be found in [23, Lemma 7.2.1 and Proposition 7.2.3]. The proof of this result has been deferred to Section 1.3 of the supplementary material.

Proposition 1. *Let $m \in \mathcal{M}_n$. With the notation of Lemma 1, let us assume that some $K > 0$ exists such that $n_\lambda \geq Kn$ and $n_\lambda \geq p + 1$ for every $\lambda \in \Lambda_m$, with $p \rightarrow +\infty$ such that $p^2/n \rightarrow 0$ as n tends to infinity. Furthermore, if $\bar{\sigma}^2, M > 0$ exist such that*

$$\text{(BN)} \quad \forall i, \quad \sigma^2(t_i) \leq \bar{\sigma}^2 \quad \text{and} \quad \text{(BV)} \quad \forall i, \quad (s(t_i) - s_m(t_i))^2 \leq M ,$$

then,

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

where $1 \leq C_p \leq p$ may only depend on $K, M, \bar{\sigma}^2$.

The comparison of (6) and (10) shows that Lpo_p yields an almost unbiased estimator of $\mathbb{E}[\|s - \widehat{s}_m\|_n^2]$: The only difference is that the factor $1/n$ in the variance term $V(m)$ has been changed into $C_p/(n-p)$, with

$$\frac{1}{n-p} \leq \frac{C_p}{n-p} \leq \frac{p}{n-p} .$$

Therefore, minimizing the Lpo_p estimator of the risk instead of the empirical risk allows to automatically take into account heteroscedasticity of data.

As already noticed, Proposition 1 is not sufficient to derive theoretical guarantees for $\widehat{m}_{\text{Lpo}_p}(D)$. In particular, precise concentration inequalities for $\widehat{R}_{\text{Lpo}_p}(m)$ would be needed. According to the assumptions under which similar results have been obtained for polynomial collections of models [5, 6, 23], we can conjecture that (at least) the following assumptions would be required to derive an upper bound on $\|s - \widehat{s}_{\widehat{m}_{\text{Lpo}_p}(D)}\|_n^2$:

- moment inequalities for the errors ε_i : $\forall q \geq 2, \|\varepsilon\|_q \leq Cq^\beta$ for some $C, \beta > 0$ (for instance, ε sub-Gaussian),
- a uniform upper bound on σ (see **(BN)** in Proposition 1), and probably a uniform lower bound or mild smoothness assumptions on $\sigma(\cdot)$,
- a relationship between the minimum number of points in the intervals of a segmentation and the smoothness of the regression function s . Note that assumption **(BV)** in Proposition 1 entails that for every $\lambda \in \Lambda_m$ and $t_i \in \lambda$, $\sum_{j=1}^n (s(t_i) - s(t_j))^2 \leq Mn_\lambda$.

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.

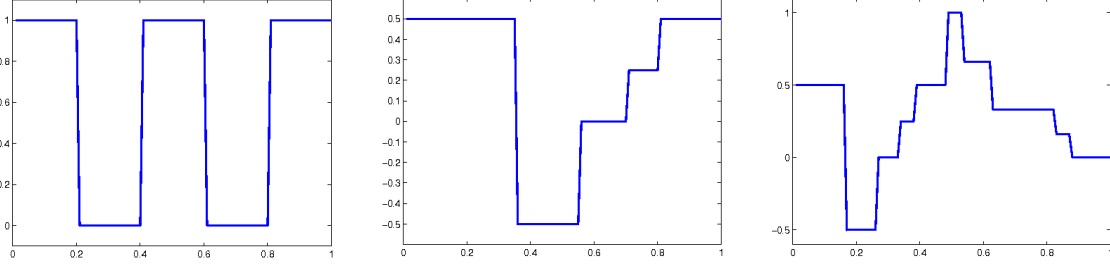


Figure 2: Regression functions s_1, s_2, s_3 ; s_1 and s_2 are piecewise constant with 4 jumps; s_3 is piecewise constant with 9 jumps.

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, except in Table 4 where an exponential noise is used. 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 $\mathcal{D}_n = \{1, \dots, 9n/25\}$. 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 σ 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 regression functions s and noise-level functions σ , is given in the second authors' thesis [23, Chapter 7]; see also Section 5 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 Lp_o_p (Procedure 2) for several values of p are compared on Figure 3, through the expected values of the quadratic loss $\mathbb{E}[\|s - \widehat{s}_{\widehat{m}_{\mathfrak{P}}(D)}\|_n^2]$ 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; Lp_o_p performs significantly better for larger dimensions. Therefore, unless the dimension is strongly over-estimated (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 Lp_o for dimensions larger than half the true dimension.

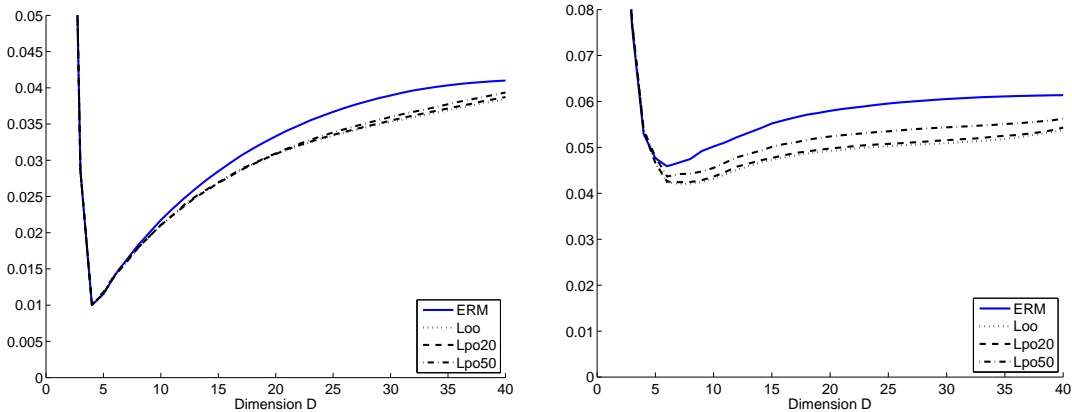


Figure 3: $\mathbb{E} \left[\left\| s - \widehat{s}_{\widehat{m}_{\mathfrak{P}}(D)} \right\|_n^2 \right]$ as a function of D for \mathfrak{P} among ‘ERM’ (empirical risk minimization), ‘Loo’ (Leave-one-out), ‘Lpo(20)’ (Lpo $_p$ with $p = 20$) and ‘Lpo(50)’ (Lpo $_p$ with $p = 50$). Left: homoscedastic, true model of dimension 5, (s_2, σ_c) . Right: heteroscedastic, true model of dimension 10, $(s_3, \sigma_{pc,3})$. 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 $8 \cdot 10^{-5}$ on the left, and smaller than $2 \cdot 10^{-4}$ on the right).

As explained in Sections 3.1 and 3.2.4, $\widehat{m}_{\text{ERM}}(D)$ often puts change-points 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 preferred to empirical risk minimization as soon as data are heteroscedastic.

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 $D \mapsto \|s - \widehat{s}_{\widehat{m}_{\mathfrak{P}}(D)}\|_n^2$ (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

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

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—, $C_{\text{or}}(\llbracket \mathfrak{P}, \text{Id} \rrbracket)$ helps to understand which is the best procedure for selecting a segmentation of a given dimension. The notation $C_{\text{or}}(\llbracket \mathfrak{P}, \text{Id} \rrbracket)$ 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 ho-

$s.$	$\sigma.$	ERM	Loo	Lpo ₂₀	Lpo ₅₀
2	c	2.87 \pm 0.01	2.89 \pm 0.01	2.90 \pm 0.01	2.96 \pm 0.01
	pc,1	1.33 \pm 0.02	1.15 \pm 0.02	1.14 \pm 0.01	1.11 \pm 0.01
	pc,3	3.14 \pm 0.03	2.52 \pm 0.02	2.47 \pm 0.02	2.36 \pm 0.02
3	c	3.18 \pm 0.01	3.25 \pm 0.01	3.29 \pm 0.01	3.44 \pm 0.01
	pc,1	3.04 \pm 0.02	2.70 \pm 0.02	2.71 \pm 0.02	2.79 \pm 0.02
	pc,3	4.44 \pm 0.02	3.98 \pm 0.02	4.00 \pm 0.02	4.14 \pm 0.02

Table 1: Average performance $C_{\text{or}}(\llbracket \mathfrak{P}, \text{Id} \rrbracket)$ 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. See also Table 3 of the supplementary material.

moscedasticity 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 change-points

In this section, the number of change-points is no longer fixed or *a priori* known. The goal is precisely to estimate this number, as often needed when analyzing data.

Two main procedures are considered. First, a penalization procedure introduced by Birgé and Massart [17] is analyzed in Section 4.1; this procedure is successful for change-point detection when data are homoscedastic [35, 37], but seems to fail when data are heteroscedastic. 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 change-points, VFCV is used for choosing among the same estimators as BM, that is $\{\hat{s}_{\widehat{m}_{\text{ERM}}(D)}\}_{D \in \mathcal{D}_n}$. 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 [17] successfully used for change-point detection in [35, 37].

Procedure 3 (Birgé and Massart [17]).

1. $\forall m \in \mathcal{M}_n, \hat{s}_m := \text{ERM}(S_m; P_n)$.
2. $\widehat{m}_{\text{BM}} := \arg \min_{m \in \mathcal{M}_n, D_m \in \mathcal{D}_n} \{P_n \gamma(\hat{s}_m) + \text{pen}_{\text{BM}}(m)\}$, where for every $m \in \mathcal{M}_n$,

the penalty $\text{pen}_{\text{BM}}(m)$ only depends on S_m through its dimension:

$$\text{pen}_{\text{BM}}(m) = \text{pen}_{\text{BM}}(D_m) := \frac{\widehat{C}D_m}{n} \left(5 + 2 \log \left(\frac{n}{D_m} \right) \right), \quad (11)$$

where \widehat{C} is an estimator of $\mathbb{E}[\sigma^2(X)]$ —in [17], σ is assumed to be constant. In the paper, we choose $\widehat{C} = \widehat{\sigma^2} = n^{-1} \sum_{i=1}^{n/2} (Y_{2i} - Y_{2i-1})^2$, assuming that $t_1 < \dots < t_n$. Note that the conclusions of the present simulation experiment remain similar with several other choices of \widehat{C} , for instance using the slope heuristics [18, 9], as proposed by Lebarbier [37] and by Lavielle [35]. See Section 4 of the supplementary material for a detailed discussion about \widehat{C} .

$$3. \quad \widetilde{s}_{\text{BM}} := \widehat{s}_{\widehat{m}_{\text{BM}}}.$$

All $m \in \mathcal{M}_n(D)$ are penalized in the same way by $\text{pen}_{\text{BM}}(m)$, so that Procedure 3 actually selects among $\{\widehat{s}_{\widehat{m}_{\text{ERM}}(D)}\}_{D \in \mathcal{D}_n}$. Therefore, Procedure 3 can be reformulated as follows, as noticed in [18, Section 4.3].

Procedure 4 (Reformulation of Procedure 3).

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

In the following, ‘BM’ denotes Procedure 4 and

$$\text{crit}_{\text{BM}}(D) := P_n \gamma(\widehat{s}_{\widehat{m}_{\text{ERM}}(D)}) + \text{pen}_{\text{BM}}(D)$$

is called the BM criterion.

Procedure 4 clarifies the reason why pen_{BM} must be larger than the C_p penalty. Indeed, for every $m \in \mathcal{M}_n$, Lemma 1 shows that when data are homoscedastic, $P_n \gamma(\widehat{s}_m) + \text{pen}(m)$ is an unbiased estimator of $\mathbb{E}[\|s - \widehat{s}_m\|_n^2]$ when $\text{pen}(m) = 2\sigma^2 D_m n^{-1}$, that is, the C_p penalty. When $\text{Card}(\mathcal{M}_n)$ is at most polynomial in n , C_p is an efficient model selection procedure, as proved in several regression frameworks [50, 38, 11]. Hence, the 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, $\text{Card}(\mathcal{M}_n)$ grows exponentially with n . The formulation of Procedure 4 points out that $\text{pen}_{\text{BM}}(D)$ has been built so that $\text{crit}_{\text{BM}}(D)$ estimates unbiasedly $\mathbb{E}[\|s - \widehat{s}_{\widehat{m}_{\text{ERM}}(D)}\|_n^2]$ for every D , where $\widehat{s}_{\widehat{m}_{\text{ERM}}(D)}$ is the empirical risk minimizer over \widetilde{S}_D . Hence, $\text{pen}_{\text{BM}}(D)$ measures the ‘‘capacity’’ of \widetilde{S}_D , which is much bigger than a vector space of dimension D . Therefore, pen_{BM} should be larger than the C_p penalty, as confirmed by the results of Birgé and Massart [18] on minimal penalties for exponential collections of models.

Simulation experiments support the fact that $\text{crit}_{\text{BM}}(D)$ is an unbiased estimator of $\mathbb{E}[\|s - \widehat{s}_{\widehat{m}(D)}\|_n^2]$ for every D (up to an additive constant) when data are homoscedastic

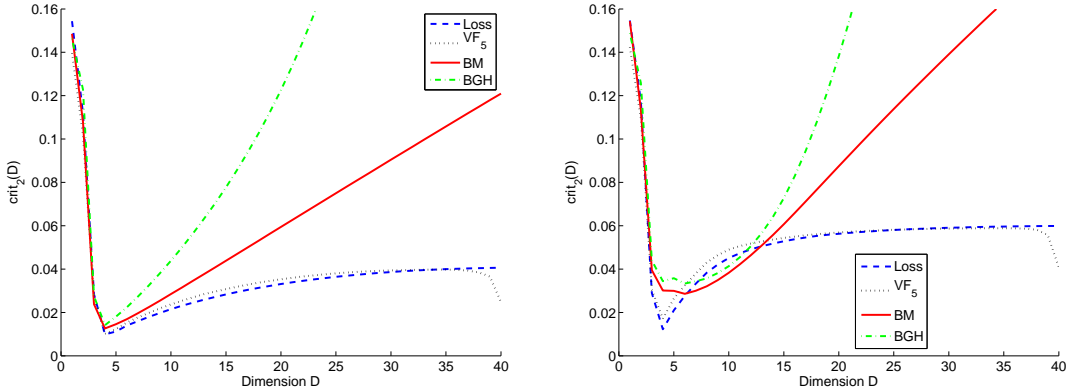


Figure 4: Comparison of the expectations of $\|s - \hat{s}_{\hat{m}(D)}\|_n^2$ (‘Loss’), $\text{crit}_{\text{VF}_5}(D)$ (‘VF₅’) and $\text{crit}_{\text{BM}}(D)$ (‘BM’); $\text{crit}_{\text{BGH}}(D)$ has been included for completeness (see Section 5.2). Data are generated as explained in Section 3.3.1. Left: homoscedastic (s_2, σ_c). Right: heteroscedastic ($s_2, \sigma_{pc,3}$). In both cases, the dimension of the true model is 5. 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 $2 \cdot 10^{-3}$ on the left, and smaller than $4 \cdot 10^{-3}$ on the right). Similar behaviours are observed for every single sample, with slightly larger fluctuations for $\text{crit}_{\text{VF}_5}(D)$ than for $\text{crit}_{\text{BM}}(D)$. The curves ‘BM’, ‘VF₅’ and ‘BGH’ have been shifted (with the same shift value) in order to make comparison with ‘Loss’ easier, without changing the location of the minimum.

(Figure 4 left). However, when data are heteroscedastic, theoretical results proved by Birgé and Massart [17, 18] no longer apply, and simulations show that $\text{crit}_{\text{BM}}(D)$ does not always estimate $\|s - \hat{s}_{\hat{m}_{\text{ERM}}(D)}\|_n^2$ 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 [7].

Therefore, $\text{pen}_{\text{BM}}(D)$ is not an adequate measure of the capacity of \tilde{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 the quadratic risk of $\text{ERM}(S_m; \cdot)$ for all $m \in \mathcal{M}_n(D)$ with a given number ($D - 1$) of change-points (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. $\forall D \in \mathcal{D}_n, \hat{s}_{\hat{m}_{\text{ERM}}(D)} := \text{ERM}(\tilde{S}_D; P_n),$
2. $\hat{D}_{\text{VF}_V} := \arg \min_{D \in \mathcal{D}_n} \{\text{crit}_{\text{VF}_V}(D)\}$
 where $\text{crit}_{\text{VF}_V}(D) := \hat{R}_{\text{VF}_V}(\text{ERM}(\tilde{S}_D(\cdot); \cdot), \cdot)$ and \hat{R}_{VF_V} is defined by (9).
3. $\tilde{s}_{\llbracket \text{ERM}, \text{VF}_V \rrbracket} =: \hat{s}_{\hat{m}_{\text{ERM}}(\hat{D}_{\text{VF}_V})}.$

$s.$	$\sigma.$	Id	VF ₅	BM	BGH
2	c	2.87 ± 0.01	3.99 ± 0.02	3.58 ± 0.02	3.52 ± 0.02
	pc,2	2.91 ± 0.02	5.62 ± 0.05	9.25 ± 0.06	10.13 ± 0.07
	s	2.98 ± 0.01	4.34 ± 0.03	4.76 ± 0.03	4.88 ± 0.03
3	c	3.18 ± 0.01	4.31 ± 0.02	4.67 ± 0.01	4.47 ± 0.01
	pc,2	4.14 ± 0.02	5.82 ± 0.02	5.90 ± 0.02	5.93 ± 0.02
	s	3.92 ± 0.01	5.61 ± 0.02	5.64 ± 0.02	5.63 ± 0.02

Table 2: Performance $C_{\text{or}}(\llbracket \text{ERM}, \mathfrak{P} \rrbracket)$ for $\mathfrak{P} = \text{Id}$ (that is, choosing the dimension $D^* := \arg \min_{D \in \mathcal{D}_n} \left\{ \|s - \widehat{s}_{\widehat{m}_{\text{ERM}}(D)}\|_n^2 \right\}$), $\mathfrak{P} = \text{VF}_V$ with $V = 5$, $\mathfrak{P} = \text{BM}$ or $\mathfrak{P} = \text{BGH}$. 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. See also Table 4 of the supplementary material.

The notation $\llbracket \text{ERM}, \text{VF}_V \rrbracket$ means that ERM is used first to choose the best segmentation for each dimension, and then VF_V estimates the number of change-points (see the last paragraph of Section 5.1).

Remark 3. The model \widetilde{S}_D depends on the design points, so it should be written $\widetilde{S}_D((t_i)_{1 \leq i \leq n})$. In particular, in the second step of Procedure 5, $\text{crit}_{\text{VF}_V}(D)$ is computed by considering the models $\widetilde{S}_D((t_i)_{i \notin B_k})$ for $k = 1, \dots, V$, which are non-empty only when $D \leq (n - \max_k \{\text{Card}(B_k)\})/2 \approx n(V-1)/(2V)$. This is always assumed in the paper (see Section 3 of the supplementary material for further details).

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

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 23, Chapter 7, and Section 5 of the supplementary material for additional results].

As illustrated by Figure 4, $\text{crit}_{\text{VF}_V}(D)$ 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; $\text{crit}_{\text{VF}_V}(D)$ only underestimates $\|s - \widehat{s}_{\widehat{m}(D)}\|_n^2$ for dimensions D close to $n(V-1)/(2V)$, for reasons explained in Section 3 of the supplementary material. On the contrary, $\text{crit}_{\text{BM}}(D)$ is a poor estimate of $\|s - \widehat{s}_{\widehat{m}(D)}\|_n^2$ when data are heteroscedastic.

Subsequently, VFCV yields a much smaller performance index

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

than BM when data are heteroscedastic (Table 2); see also Section 4 of the supplementary material 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 [17, 18].

Overall, VFCV seems to be a reliable alternative to BM when no prior knowledge can guarantee 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 change-points given their number, as well as for estimating the number of change-points. 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 $(D - 1)$ change-points should be located for every $D \in \mathcal{D}_n$; second, estimate the number $(D - 1)$ of change-points. This strategy can be summarized with the following procedure:

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

1. $\forall D \in \mathcal{D}_n$, $\mathcal{A}_D(P_n) := \widehat{s}_{\widehat{m}(D)} = \arg \min_{m \in \mathcal{M}_n(D)} \{\text{crit}_1(S_m, P_n)\}$ where for every model S , $\text{crit}_1(S, P_n) \in \mathbb{R}$ estimates $\|s - \text{ERM}(S; P_n)\|_n^2$ and $\widehat{s}_m = \text{ERM}(S_m; P_n)$ is defined as in Section 3.1.
2. $\widehat{D} = \arg \min_{D \in \mathcal{D}_n} \{\text{crit}_2(\mathcal{A}_D, P_n)\}$, where for every algorithm \mathcal{A}_D , $\text{crit}_2(\mathcal{A}_D, P_n) \in \mathbb{R}$ estimates $\|s - \mathcal{A}_D(P_n)\|_n^2$.
3. Output: $\widehat{m}(\widehat{D})$ 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

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

- The Leave- p -out estimator of the risk (‘Lpo $_p$ ’) is, for every $p \in \{1, \dots, n-1\}$,

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

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

As in Section 3, Loo denotes Lpo $_1$. The VFCV estimator of the risk $\widehat{R}_{\text{VFCV}}$ 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

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

where $\text{pen}_{\text{BM}}(D)$ is defined by (11) with $c_1 = 5$, $c_2 = 2$ and \widehat{C} is chosen by the slope heuristics (see Section 4 of the supplementary material).

- The V -fold cross-validation estimator of the risk (‘VF $_V$ ’) is, for every $V \in \{1, \dots, n\}$,

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

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

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

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 $\text{crit}_{1,\alpha}$ and $\text{crit}_{2,\beta}$ ”, 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 $[[\text{ERM}, \text{BM}]]$; Procedures $[[\alpha, \text{Id}]]$ are compared for several α in Section 3; Procedures $[[\text{ERM}, \beta]]$ are compared for $\beta \in \{\text{Id}, \text{BM}, \text{VF}_5\}$ in Section 4.

5.2 Alternative competing procedures

Three alternative procedures for change-point detection are also considered in the simulation study of Section 5.3. Some of them follow a completely different strategy from that of Birgé and Massart. Their comparison with procedures proposed in Section 5.1 strengthens the generality level of the observed results.

ZS Zhang and Siegmund [59] proposed a modified version of the BIC criterion, denoted by ZS in the sequel. It is suited to the specific change-point detection setting, where likelihood is not smooth with respect to the location of the change-points, unlike what is

implicitly assumed to derive BIC. The ZS procedure consists in maximizing over $m \in \mathcal{M}_n$ the following criterion:

$$\begin{aligned} \text{crit}_{\text{ZS}}(m) := & \frac{n - D_m + 1}{2} \log \left[1 + \frac{SS_{\log}}{SS_{\text{all}} - SS_{\log}} \right] + \log \left[\Gamma \left(\frac{n - D_m + 1}{2} \right) \Gamma^{-1} \left(\frac{D_m + 1}{2} \right) \right] \\ & + \frac{D_m}{2} \log(SS_{\text{all}}) - \frac{1}{2} \sum_{\lambda \in \Lambda_m} \log n_\lambda + \left(\frac{1}{2} - D_m \right) \log(n) , \end{aligned}$$

where

$$SS_{\log} = \sum_{\lambda \in \Lambda_m} \sum_{i=1}^n \left(\hat{s}_m(t_i) - \sum_{j=1}^n \frac{Y_j}{n} \right)^2 \mathbb{1}_\lambda(t_i) , \quad SS_{\text{all}} = \sum_{i=1}^n \left(Y_i - \sum_{j=1}^n \frac{Y_j}{n} \right)^2$$

and $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$.

Note that this procedure does not assume the variance to be known, but assumes the variance is constant.

BGH In the context of Gaussian regression, Baraud, Giraud and Huet [12] propose a new penalized criterion, which does not depend on the *a priori* knowledge of the variance of the noise: The selected model results from the minimization over $m \in \mathcal{M}_n$ of the criterion

$$\text{crit}_{\text{BGH}}(m) := \left[\sum_{\lambda \in \Lambda_m} \sum_{i=1}^n \left[(Y_i - \hat{s}_m(t_i))^2 \mathbb{1}_\lambda(t_i) \right] \right] \left(1 + \frac{\text{pen}_{\text{BGH}}(m)}{n - D_m} \right) ,$$

where $\text{pen}_{\text{BGH}}(m)$ is defined in [12] and depends on a constant chosen by simulation to be equal to $K = 1.1$.

PML The so-called PML procedure—for penalized maximum likelihood—was introduced in the bioinformatics community to deal with the particular problem of change-point detection in CGH profiles [45]. PML aims at detecting changes in either the mean or the variance, that is, change-points for (s, σ) . The selected model is defined as a minimizer over $m \in \mathcal{M}_n$ of

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

where $n_\lambda = \text{Card} \{t_i \in \lambda\}$ and \hat{C}'' is estimated from data by the slope heuristics algorithm [35, 37].

5.3 Simulation study

A simulation experiment compares procedures $[\alpha, \text{VF}_5]$ for several α , $[\text{ERM}, \text{BM}]$, and also ZS, BGH, and PML (defined Section 5.2). The experimental setting is the same as that of Section 3.3.1. Two cases are considered for the errors ε_i : standard Gaussian, and centered normalized exponential. A representative picture of the results is given by

Tables 3–4; additional results can be found in Section 5 of the supplementary material and in [23, Chapter 7]. The (statistical) performance of each competing procedure \mathfrak{P} is measured by

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

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

(s, σ)	(s_1, σ_c)	$(s_1, \sigma_{pc,2})$	$(s_1, \sigma_{pc,3})$	(s_1, σ_s)
[[Loo, VF ₅]]	2.40 ± 0.02	3.17 ± 0.03	3.40 ± 0.03	2.59 ± 0.03
[[ERM, VF ₅]]	2.38 ± 0.02	3.20 ± 0.03	3.42 ± 0.03	2.73 ± 0.03
[[ERM, BM]]	1.70 ± 0.02	3.10 ± 0.03	3.81 ± 0.03	2.08 ± 0.02
BGH	1.85 ± 0.02	6.38 ± 0.05	6.51 ± 0.04	3.83 ± 0.03
ZS	1.71 ± 0.02	3.92 ± 0.04	4.41 ± 0.04	2.46 ± 0.03
PML	2.79 ± 0.02	3.42 ± 0.03	3.97 ± 0.04	2.72 ± 0.03

(s, σ)	(s_2, σ_c)	$(s_2, \sigma_{pc,2})$	$(s_2, \sigma_{pc,3})$	(s_2, σ_s)
[[Loo, VF ₅]]	4.02 ± 0.02	4.95 ± 0.05	5.24 ± 0.05	4.32 ± 0.03
[[ERM, VF ₅]]	3.99 ± 0.02	5.62 ± 0.05	5.94 ± 0.06	4.34 ± 0.03
[[ERM, BM]]	3.58 ± 0.02	9.25 ± 0.06	8.79 ± 0.06	4.76 ± 0.03
BGH	3.52 ± 0.02	10.13 ± 0.07	9.77 ± 0.07	4.88 ± 0.03
ZS	3.62 ± 0.02	6.50 ± 0.05	6.92 ± 0.06	4.16 ± 0.02
PML	4.34 ± 0.02	2.73 ± 0.03	2.84 ± 0.03	4.32 ± 0.02

(s, σ)	(s_3, σ_c)	$(s_3, \sigma_{pc,2})$	$(s_3, \sigma_{pc,3})$	(s_3, σ_s)
[[Loo, VF ₅]]	4.42 ± 0.02	5.24 ± 0.02	5.59 ± 0.02	5.35 ± 0.02
[[ERM, VF ₅]]	4.31 ± 0.02	5.82 ± 0.02	6.13 ± 0.02	5.61 ± 0.02
[[ERM, BM]]	4.67 ± 0.01	5.90 ± 0.02	6.24 ± 0.02	5.64 ± 0.02
BGH	4.47 ± 0.01	5.93 ± 0.02	6.31 ± 0.02	5.63 ± 0.02
ZS	5.46 ± 0.02	6.63 ± 0.02	6.61 ± 0.02	6.31 ± 0.02
PML	5.05 ± 0.02	4.67 ± 0.03	4.99 ± 0.03	5.52 ± 0.02

Table 3: Performance $C_{\text{or}}(\mathfrak{P})$ for several change-point detection procedures \mathfrak{P} . Several regression functions s and noise-level functions σ have been considered, each time with $N = 10000$ 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)	$(s_2, \sigma_{pc,2})$	$(s_2, \sigma_{pc,3})$	$(s_3, \sigma_{pc,2})$	$(s_3, \sigma_{pc,3})$
[[Loo, VF ₅]]	4.47 ± 0.05	4.69 ± 0.06	4.80 ± 0.03	5.11 ± 0.03
[[ERM, VF ₅]]	5.98 ± 0.07	6.31 ± 0.07	5.82 ± 0.04	6.22 ± 0.04
[[ERM, BM]]	10.81 ± 0.09	10.31 ± 0.09	6.09 ± 0.04	6.45 ± 0.04
BGH	11.67 ± 0.09	11.15 ± 0.09	5.94 ± 0.03	6.42 ± 0.04
ZS	9.34 ± 0.09	9.29 ± 0.09	6.60 ± 0.04	6.83 ± 0.04
PML	5.04 ± 0.06	5.00 ± 0.06	5.17 ± 0.03	5.40 ± 0.03

Table 4: Same as Table 3 with ε_i having an exponential distribution.

Let us now comment the results reported in Tables 3–4.

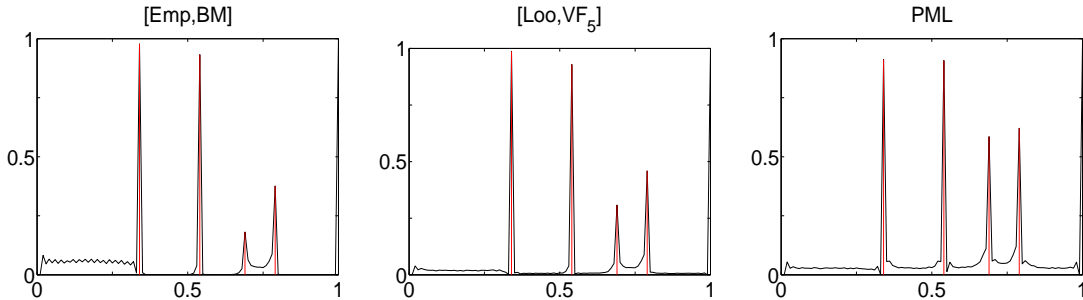


Figure 5: Probability to put a change-point at t_i , $i = 1, \dots, n$, for three procedures in setting $(s_2, \sigma_{pc,3})$ with exponential errors. The four vertical red lines correspond to the four true change-points of $s = s_2$.

First, BM is consistently outperformed by other procedures and can lead to poor estimations, except in the homoscedastic settings in which it confirms its strength. In the few heteroscedastic frameworks where BM performs well ($s = s_1$ and $\sigma = \sigma_{pc,2}$ or σ_s), these good performances are mainly due to a somehow lucky overestimation of $\mathbb{E}[\sigma^2(X)]$, as detailed in Section 4 of the supplementary material. Comparing BM and $\llbracket \text{Loo}, \text{VF}_5 \rrbracket$, Figure 5 shows that BM often puts false change-points in $[0, 1/3)$, where the noise-level is higher, and that BM detects less often the two true change-points of s at $t = 0.7$ and $t = 0.8$. This confirms the phenomenon pointed out in Section 3.

Second, empirical risk minimization (ERM) slightly outperforms CV (Loo) when data are homoscedastic. On the contrary, when data are heteroscedastic, Loo clearly outperforms ERM, often by a margin larger than 10%, for instance in setting $(s_2, \sigma_{pc,2})$. Since ERM puts more change-points than CV in noisy regions, it leads to poor approximations of the oracle segmentation. Therefore, the results of Section 3 are confirmed when using VF_5 (instead of Id) for choosing the dimension.

Third, considering the overall simulation study, $\llbracket \text{Loo}, \text{VF}_5 \rrbracket$ remarkably achieves the best performance. It is sometimes outperformed by other procedures like PML, BGH or ZS in specific settings. But these three procedures can yield very poor results in different contexts, whereas CV remains reliable. For instance, BGH performs badly in settings $(s_2, \sigma_{pc,\cdot})$ compared with CV, and ZS performs significantly worse than $\llbracket \text{Loo}, \text{VF}_5 \rrbracket$ in all heteroscedastic settings but (s_1, σ_s) and (s_2, σ_s) . The good performances achieved by PML in Table 3 with s_2 , s_3 and $\sigma_{pc,\cdot}$ are balanced by the overfitting of PML, which chooses segmentations with too many change-points, as illustrated by Table 5 and by Figures 4–5 of the supplementary material. Furthermore, PML relies on the assumption that errors are Gaussian: with an exponential noise (Table 4), CV outperforms PML and other procedures. Figure 5 also shows that PML detects less often than CV the two “easy” change-points of s_2 at 0.35 and 0.55, and that PML often puts false change-points uniformly over $[0, 1]$.

Fourth, C_{or} often decreases when the noise-level increases, such as in settings $(s_1, \sigma_{pc,\cdot})$ and $(s_2, \sigma_{pc,\cdot})$ in Table 3. Indeed, C_{or} is a ratio to the performance of the risk of the oracle, which becomes so small when the noise-level decreases that it becomes harder to keep C_{or} at the same level. In these frameworks, only overfitting can make C_{or} larger when the noise-level is larger (see the performances of $\llbracket \text{ERM}, \text{BM} \rrbracket$).

		Med(\widehat{D})	$\mathbb{P}[\widehat{D} < D^*]$	$\mathbb{P}[\widehat{D} = D^*]$	$\mathbb{P}[\widehat{D} > D^*]$
$(s_1, \sigma_{pc,3})$	[[Loo, VF ₅]]	5	0	0.68	0.32
	BGH	7	0	0.30	0.70
	ZS	5	0	0.56	0.44
	PML	7	0	0.14	0.86
$(s_2, \sigma_{pc,2})$	[[Loo, VF ₅]]	5	0.50	0.31	0.19
	BGH	6	0.37	0.10	0.53
	ZS	4	0.66	0.13	0.21
	PML	7	0	0.25	0.75

Table 5: Empirical probabilities of choosing the best possible dimension D^* for several change-point detection procedures. Several regression functions s and noise-level functions σ have been considered. $\text{Med}(\widehat{D})$ denotes the empirical median of \widehat{D} over $N = 10000$ independent samples.

In order to assess the generality of the results of Table 3, the procedures considered in Section 5.3 have been compared in three random settings. The results of these experiments are reported in Section 2 of the supplementary material. They show in particular that when the noise-level happens to be much larger in areas where s does not jump than in areas where s does jump, then the local overfitting phenomenon is more frequent, making $C_{\text{or}}(\llbracket \text{Lpo}, \text{VF}_5 \rrbracket)$ significantly smaller than $C_{\text{or}}(\llbracket \text{ERM}, \text{VF}_5 \rrbracket)$.

Finally, the dependence on p of the performance of $\llbracket \text{Lpo}_p, \text{VF}_5 \rrbracket$ is assessed in Section 5 of the supplementary material. All values of p between 1 and $n/2$ 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 6, 22, 23], but it seems difficult to characterize the settings for which overpenalization is needed for detecting change-points given their number.

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 44–46, 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 [45]. A test genome of Bt474 cell lines is compared to a normal reference male genome. Even 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 [45].

The first procedure is a simplified version of BM, called ‘BMsimple’, proposed by Lavielle [35, Section 2] and first used on CGH data in [45]. Note that BM would give similar results on the data of Figure 6.

The second procedure is PML, as defined in Section 5.2.

6.3 Results

Results obtained with BMsimple, PML, $\llbracket\text{ERM}, \text{VF}_5\rrbracket$ and $\llbracket\text{Lpo}_{20}, \text{VF}_5\rrbracket$ on the Bt474 data set are reported on Figure 6.

For chromosome 9, BMsimple and PML yield (almost) the same segmentation, so that the homoscedasticity assumption is certainly not much violated. As expected, $\llbracket\text{ERM}, \text{VF}_5\rrbracket$ and $\llbracket\text{Lpo}_{20}, \text{VF}_5\rrbracket$ also yield very similar segmentations, which confirms the reliability of these procedures for homoscedastic signal [see 23, Section 7.6 for details].

The picture is quite different for chromosome 1. Indeed, as shown by Figure 6 (right), BMsimple selects a segmentation with 7 change-points, whereas PML selects a segmentation with only one change-point. 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 $\llbracket\text{ERM}, \text{VF}_5\rrbracket$ and $\llbracket\text{Lpo}_{20}, \text{VF}_5\rrbracket$ yield two other segmentations, with a medium number of change-points, respectively 4 and 3. In view of the simulation experiments of the previous sections, the segmentation obtained via $\llbracket\text{Lpo}_{20}, \text{VF}_5\rrbracket$ should be the most reliable one since data can be heteroscedastic. Therefore, the right of Figure 6 can be interpreted as follows: The noise-level is small in the first part of chromosome 1,

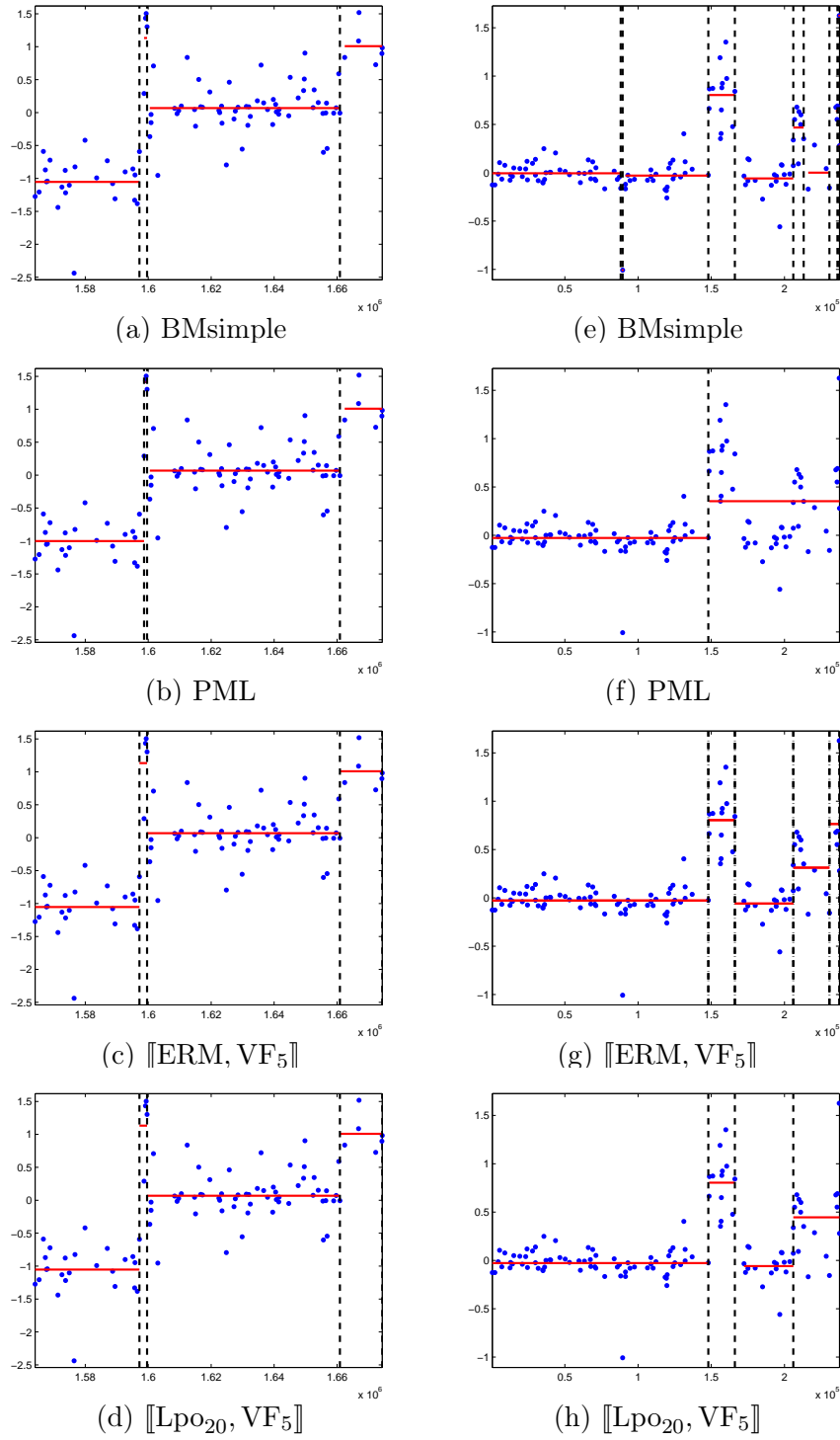


Figure 6: Change-points locations along Chromosome 9 (Left) and Chromosome 1 (Right). The mean on each homogeneous region is indicated by plain horizontal lines.

then higher, but not as high as estimated by PML. In particular, the copy number changes 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 $\llbracket \text{ERM}, \text{VF}_5 \rrbracket$ and $\llbracket \text{Lpo}_{20}, \text{VF}_5 \rrbracket$ 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 change-points is given, empirical risk minimization seems to fail for some heteroscedastic problems, as shown by the simulation study and by partial theoretical results. On the contrary, the Leave- p -out (Lpo_p) seems robust to a possible heteroscedasticity while being computationally efficient thanks to closed-form formulas given in Section 3.2.3 (Theorem 1).

Second, for choosing the number of change-points, 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) seems 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 change-points) and VFCV yields a reliable procedure for detecting changes in the mean of a signal which is not *a priori* known to be homoscedastic (contrary to existing change-point detection procedures, that are not well adapted to this problem). The resulting procedure is computationally tractable for small values of V , since its computational complexity is of order $\mathcal{O}(Vn^2 \text{Card}(\mathcal{D}_n)) = \mathcal{O}(Vn^3)$, 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 [6] could be used at the first step, for localizing the change-points given their number. At the second step,

V -fold penalization [7] could also be used instead of VFCV, with the same computational cost and possibly an improved statistical performance.

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 $Lp\alpha_p$ is used at the first step should be further investigated. Preliminary results in this direction can already be found in [23, Chapter 7].

Another important direction for future works concerns the assumption that errors ε_i are independent, which is always made throughout the paper. Indeed, the cross-validation heuristics strongly relies on this to make sure that the validation sample is independent from the training sample. So, one can wonder whether Procedure 6 can be applied when errors are dependent. The literature on CV [see Section 8.1 in 8] shows that dependencies can make CV overfit [see for instance 42], except maybe in very specific settings [e.g., with a particular stationary Markov process 21]. Most approaches to overcome this problem [e.g., “modified CV”, 26] consist in discarding part of the validation sample so that it becomes (almost) independent from the training sample, which obviously can only be done in the change-point detection framework if the segments are much longer than the range of dependencies. Up to the best of our knowledge, no modification of CV has ever been proposed that could deal with more general dependencies for change-point detection. So, we can only make the following suggestion: if two (or more) independent copies of $(t_i, Y_i)_{1 \leq i \leq n}$ are available (which is possible in many applications), use a procedure of the form $[[VF, VF]]$ where in both steps, the training and validation samples are taken from independent parts of the whole sample.

Acknowledgments

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-09-JCJC-0027-01.

References

- [1] Felix Abramovich, Yoav Benjamini, David L. Donoho, and Iain M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.
- [2] Hirotugu Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217, 1970.
- [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. V -fold cross-validation improved: V -fold penalization, February 2008. arXiv:0802.0566v2.

- [6] Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624 (electronic), 2009.
- [7] Sylvain Arlot. Choosing a penalty for model selection in heteroscedastic regression, June 2010. arXiv:0812.3141.
- [8] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010. Doi: 10.1214/09-SS054.
- [9] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10:245–279 (electronic), 2009.
- [10] Yannick Baraud. Model selection for regression on a fixed design. *Probab. Theory Related Fields*, 117(4):467–493, 2000.
- [11] Yannick Baraud. Model selection for regression on a random design. *ESAIM Probab. Statist.*, 6:127–146 (electronic), 2002.
- [12] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2):630–672, 2009.
- [13] Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory and Relat. Fields*, 113:301–413, 1999.
- [14] Michèle Basseville and Igor V. Nikiforov. *Detection of abrupt changes: theory and application*. Prentice Hall Information and System Sciences Series. Prentice Hall Inc., Englewood Cliffs, NJ, 1993.
- [15] Richard E. Bellman and Stuart E. Dreyfus. *Applied dynamic programming*. Princeton University Press, Princeton, N.J., 1962.
- [16] Lucien Birgé and Pascal 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.
- [17] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. European Math. Soc.*, 3(3):203–268, 2001.
- [18] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.
- [19] Boris E. Brodsky and Boris S. Darkhovsky. *Methods in Change-point problems*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1993.
- [20] Prabir Burman. A comparative study of ordinary cross-validation, v -fold cross-validation and the repeated learning-testing methods. *Biometrika*, 76(3):503–514, 1989.
- [21] Prabir Burman and Deborah Nolan. Data-dependent estimation of prediction functions. *J. Time Ser. Anal.*, 13(3):189–207, 1992.

- [22] Alain Celisse. Model selection in density estimation via cross-validation. Technical report, arXiv, 2008. arXiv:0811.0802v2.
- [23] Alain Celisse. *Model selection via cross-validation in density estimation, regression and change-points detection*. PhD thesis, University Paris-Sud 11, December 2008. <http://tel.archives-ouvertes.fr/tel-00346320/>.
- [24] Alain Celisse and Stéphane Robin. Nonparametric density estimation by exact leave- p -out cross-validation. *Comput. Statist. Data Anal.*, 52(5):2350–2368, 2008.
- [25] Alain Celisse and Stéphane Robin. A cross-validation based estimation of the proportion of true null hypotheses. *J. Statist. Plann. Inference*, 2010. doi:10.1016/j.jspi.2010.04.014.
- [26] C.-K. Chu and J. S. Marron. Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, 19(4):1906–1918, 1991.
- [27] F. Comte and Y. Rozenholc. Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Process. Appl.*, 97(1):111–145, 2002.
- [28] Sandrine Dudoit and Mark J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.*, 2(2):131–154, 2005.
- [29] Seymour Geisser. A predictive approach to the random effect model. *Biometrika*, 61(1):101–107, 1974.
- [30] Seymour Geisser. The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.*, 70:320–328, 1975.
- [31] Xavier Gendre. Simultaneous estimation of the mean and the variance in heteroscedastic Gaussian regression. *Electron. J. Stat.*, 2:1345–1372, 2008.
- [32] Z. Harchaoui, F. Vallet, A Lung-Yut-Fong, and O. Cappé. A regularized kernel-based approach to unsupervised audio segmentation. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009.
- [33] Michael Kearns, Yishay Mansour, Andrew Y. Ng, and Dana Ron. An Experimental and Theoretical Comparison of Model Selection Methods. *Mach. Learn.*, 7:7–50, 1997.
- [34] Peter A. Lachenbruch and M. Ray Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10:1–11, 1968.
- [35] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85:1501–1510, 2005.
- [36] Marc Lavielle and Gilles Teyssière. Detection of Multiple Change-Points in Multivariate Time Series. *Lithuanian Mathematical Journal*, 46:287–306, 2006.
- [37] Émilie Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proc.*, 85:717–736, 2005.

- [38] Ker-Chau 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.
- [39] Collin L. Mallows. Some comments on C_p . *Technometrics*, 15:661–675, 1973.
- [40] P. Massart. *Concentration Inequalities and Model Selection*. Lecture Notes in Mathematics. Springer, 2007.
- [41] B. Q. Miao and L. C. Zhao. On detection of change points when the number is unknown. *Chinese J. Appl. Probab. Statist.*, 9(2):138–145, 1993.
- [42] Jean Opsomer, Yuedong Wang, and Yuhong Yang. Nonparametric regression with correlated errors. *Statist. Sci.*, 16(2):134–153, 2001.
- [43] Dominique Picard. Testing and estimating change-points in time series. *Adv. in Appl. Probab.*, 17(4):841–867, 1985.
- [44] Franck Picard. *Process segmentation/clustering Application to the analysis of array CGH data*. PhD thesis, Université Paris-Sud 11, 2005. <http://tel.archives-ouvertes.fr/tel-00116025/fr/>.
- [45] Franck Picard, Stéphane Robin, Marc Lavielle, Christian Vaisse, and Jean-Jacques Daudin. A statistical approach for array CGH data analysis. *BMC Bioinformatics*, 27(6):electronic access, 2005.
- [46] Franck Picard, Stéphane Robin, Émilie Lebarbier, and Jean-Jacques Daudin. A segmentation/clustering model for the analysis of array CGH data. *Biometrics*, 63(3):758–766, 2007.
- [47] Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *Ann. Statist.*, 11(2):416–431, 1983.
- [48] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.
- [49] Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.
- [50] Ritei Shibata. An optimal selection of regression variables. *Biometrika*, 68(1):45–54, 1981.
- [51] Charles J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *The Annals of Statistics*, 12(4):1285–1297, 1984.
- [52] Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974.
- [53] Mervyn Stone. An Asymptotic Equivalence of Choice of Model by Cross-validation and Akaike’s Criterion. *JRSS B*, 39(1):44–47, 1977.
- [54] Robert Tibshirani and Keith Knight. The covariance inflation criterion for adaptive model selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 61(3):529–546, 1999.

- [55] Yuhong Yang. Regression with multiple candidate model: selection or mixing? *Statist. Sinica*, 13:783–809, 2003.
- [56] Yuhong Yang. Comparing Learning Methods for Classification. *Statistica Sinica*, 16:635–657, 2006.
- [57] Yuhong Yang. Consistency of cross-validation for comparing regression procedures. *The Annals of Statistics*, 35(6):2450–2473, 2007.
- [58] Yi-Ching Yao. Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.*, 6(3):181–189, 1988.
- [59] N. R. Zhang and D. O. Siegmund. Modified Bayes Information Criterion with Application to the Analysis of Comparative Genomic Hybridization Data. *Biometrics*, 63:22–32, 2007.