A kernel multiple change-point algorithm via model selection

Sylvain Arlot, Alain Celisse, Zaid Harchaoui

A kernel multiple change-point algorithm via model selection

Sylvain Arlot
sylvain.arlot@math.u-psud.fr
Laboratoire de Mathématiques d'Orsay
Univ. Paris-Sud, CNRS, Université Paris-Saclay
91405 Orsay, France

Alain Celisse
celisse@math.univ-lille1.fr
Laboratoire de Mathématiques Paul Painlevé
UMR 8524 CNRS-Université Lille 1
Modal Project-Team
F-59 655 Villeneuve d'Ascq Cedex, France

Zaid Harchaoui
zaid.harchaoui@nyu.edu
Courant Institute
New York University
715 Broadway
New York

March 18, 2016

Abstract

We tackle the change-point problem with data belonging to a general set. We build a penalty for choosing the number of change-points in the kernel-based method of Harchaoui and Cappé (2007). This penalty generalizes the one proposed by Lebarbier (2005) for one-dimensional signals. We prove a non-asymptotic oracle inequality for the proposed method, thanks to a new concentration result for some function of Hilbert-space valued random variables. Experiments on synthetic data illustrate the accuracy of our method, showing that it can detect changes in the whole distribution of data, even when the mean and variance are constant.

Keywords: model selection, kernel methods, change-point detection, concentration inequality

1 Introduction

The change-point problem has been tackled in numerous papers in the statistics and machine learning literature (Brodsky and Darkhovsky, 1993; Carlstein et al., 1994; Tartakovsky et al., 2014). Given a time series, the goal is to split it into homogeneous segments, in which the marginal distribution of the observations—their mean or their variance, for instance—is constant. When the number of change-points is known, this problem reduces to estimating the change-point locations as precisely as possible; in general, the number of change-points itself must be estimated. This problem arises in a wide range of applications, such as bioinformatics (Picard et al., 2005; Curtis et al., 2012), neuroscience (Park et al., 2015), audio signal processing (Wu and Hsieh, 2006), temporal video segmentation (Koprinska and

Carrato, 2001), hacker attacks detection (Wang et al., 2014), social sciences (Kossinets and Watts, 2006) and econometrics (McCulloh, 2009).

Related work. A large part of the literature on change-point detection deals with observations in $\mathbb{R}$ or $\mathbb{R}^d$ and focuses on detecting changes arising in the mean and/or the variance of the signal (Gijbels et al., 1999; Picard et al., 2005; Arlot and Celisse, 2011; Bertin et al., 2014). To this end, parametric models are often involved to derive change-point detection procedures. For instance, Comte and Rozenholc (2004), Lebarbier (2005), Picard et al. (2011) and Geneus et al. (2014) make a Gaussian assumption, while Frick et al. (2014) and Cleynen and Lebarbier (2014b) consider an exponential family.

The challenging problem of detecting abrupt changes in the full distribution of the data has been recently addressed in the nonparametric setting. However, the corresponding procedures suffer several limitations since they are limited to real-valued data or they assume that the number of true change-points is known. For instance, Zou et al. (2014) design a strategy based on empirical cumulative distribution functions that allows to recover an unknown number of change-points by use of BIC, but only applies to $\mathbb{R}$-valued data. A similar conclusion applies to the strategy of Matteson and James (2014), which is moreover time-consuming due to an intensive permutation use, and only justified in an asymptotic setting. The kernel-based procedure proposed by Harchaoui and Cappé (2007) enables to deal with complex data (not necessarily vectors). But it assumes the number of change-points to recover is known, which reduces its practical interest when no such information is available. Finally, many of these procedures are only theoretically grounded by asymptotic results, which makes their finite-sample performance questionable.

Other attempts have been made to design change-point detection procedures allowing to deal with complex data (that are not necessarily vectors). However, the resulting procedures do not allow to detect more than one or two changes arising in particular features of the distribution. For instance, Chen and Zhang (2015) describe a strategy based on a dissimilarity measure between individuals to compute a graph from which a statistical test allows to detect only one or two change-points. For a graph-valued time series, Wang et al. (2014) design specific scan statistics to test whether one change arises in the connectivity matrix.

Main contributions. We first describe a new efficient multilpe change-point detection procedure allowing to deal with univariate, multivariate or complex data (DNA sequences or graphs, for instance) as soon as a positive semidefinite kernel can be defined for them. Among several assets, this procedure is nonparametric and does not require to know the true number of change-points in advance. Furthermore, it allows to detect abrupt changes arising in the full distribution of the data by using a characteristic kernel; it can also focus on changes in specific features of the distribution by choosing an appropriate kernel.

Secondly, our kernel change-point detection procedure is theoretically grounded with a finite-sample optimality result, namely an oracle inequality in terms of quadratic risk, stating that its performance is almost the same as that of the best one within the class we consider. As argued by Lebarbier (2005) for instance, such a guarantee is what we want for a change-point detection procedure. It means that the procedure detects only

changes that are "large enough" given the noise level and the amount of data available, which is necessary to avoid having many false positives. Note that contrary to previous oracle inequalities in the change-point detection framework, our result requires neither the variance to be constant nor the data to be Gaussian.

Thirdly, we settle a new concentration inequality for the quadratic norm of sums of independent Hilbert-valued vectors with exponential tails, which is a key result to derive our non-asymptotic oracle inequality with a large collection of candidate segmentations. The application domain of our exponential concentration inequality is not limited to change-point detection.

Motivating examples are first provided in Section 2 to highlight the wide applicability of our procedure to various important settings. A comprehensive description of our kernel change-point detection algorithm (namely Algorithm 1) is provided in Section 3, where we also discuss algorithmic aspects as well as the practical choice of influential parameters (Section 3.3). Section 4 exposes some important ideas underlying Algorithm 1 and then states the main theoretical results of the paper (Proposition 1 and Theorem 2). Proofs of these main results have been collected in Section 5, while technical details have been deferred to Appendices A and B. The practical performance of the kernel change-point detection algorithm is illustrated by experiments on synthetic data in Section 6. Section 7 concludes the paper by a short discussion.

Notation. For any $a < b$, we denote by $[\![a, b]\!] := [a, b] \cap \mathbb{N}$ the set of integers between a and b.

2 The change-point problem

Let $\mathcal{X}$ be some measurable set and $X_1, \ldots, X_n \in \mathcal{X}$ a sequence of independent $\mathcal{X}$-valued random variables. For any $i \in \{1, \ldots, n\}$, we denote by P_{X_i} the distribution of X_i. The change-point problem can then be summarized as follows: Given $(X_i)_{1 \leq i \leq n}$, the goal is to find the locations of the abrupt changes along the sequence $P_{X_1}, \ldots, P_{X_n}$. Note that the case of dependent time series is often considered in the change-point literature (Lavielle and Moulines, 2000; Bardet and Kammoun, 2008; Bardet et al., 2012; Chang et al., 2014); as a first step, this paper focuses on the independent case for simplicity.

An important example to have in mind is when X_i corresponds to the observation at time $t_i = i/n$ of some random process on $[0, 1]$, and we assume that this process is stationary over $[t^\star_\ell, t^\star_{\ell+1})$, $\ell = 0, \ldots, D^\star - 1$, for some fixed sequence $0 = t^\star_0 < t^\star_1 < \cdots < t^\star_{D^\star} = 1$. Then, the change-point problem is equivalent to localizing the change-points $t^\star_1, \ldots, t^\star_{D^\star - 1} \in [0, 1]$, which should be possible as the sample size n tends to infinity. Note that we never make such an asymptotic assumption in the paper, where all theoretical results are non-asymptotic.

Let us now detail some motivating examples of the change-point problem.

Example 1 *The set $\mathcal{X}$ is $\mathbb{R}$ or $\mathbb{R}^d$, and the sequence $(P_{X_i})_{1 \leq i \leq n}$ changes only through its mean. This is the most classical setting, for which numerous methods have been proposed and analyzed in the one-dimensional setting (Comte and Rozenholc, 2004; Zhang and Siegmund, 2007; Boysen et al., 2009; Korostelev and Korosteleva, 2011; Fryzlewicz, 2014) as well as*

the multi-dimensional case (Picard et al., 2011; Bleakley and Vert, 2011; Hocking et al., 2013; Soh and Chandrasekaran, 2014; Collilieux et al., 2015).

Example 2 *The set $\mathcal{X}$ is $\mathbb{R}$ or $\mathbb{R}^d$, and the sequence $(P_{X_i})_{1\leq i\leq n}$ changes only through its mean and/or its variance (or covariance matrix). This setting is rather classical, at least in the one-dimensional case, and several methods have been proposed for it (Andreou and Ghysels, 2002; Picard et al., 2005; Fryzlewicz and Subba Rao, 2014).*

Example 3 *The set $\mathcal{X}$ is $\mathbb{R}$ or $\mathbb{R}^d$, and no assumption is made on the changes in the sequence $(P_{X_i})_{1\leq i\leq n}$. For instance, when data are centered and normalized, as in the audio track example (Rabiner and Schäfer, 2007), the mean and the variance of the X_i can be constant, and only higher-order moments of $(P_{X_i})_{1\leq i\leq n}$ are changing. Only a few recent papers deal with (an unknown number of) multiple change-points in a fully nonparametric framework (Zou et al., 2014; Matteson and James, 2014; Biau et al., 2015).*
Note that assuming $\mathcal{X} = \mathbb{R}$ and adding some further restrictions on the maximal order of the moments for which a change can arise in the sequence $(P_{X_i})_{1\leq i\leq n}$, it is nevertheless possible to consider the multivariate sequence $((p_j(X_i))_{0\leq j\leq d})_{1\leq i\leq n}$, where p_j is a polynomial of degree j for $j \in \{0,\dots,d\}$, and to use a method made for detecting changes in the mean (Example 1). For instance with $\mathbb{R}$-valued data, one can take $p_j(X) = X^j$ for every $1 \leq j \leq d$, or p_j equal to the j-th Hermite polynomial, as proposed by Lajugie et al. (2014).

Example 4 *The set $\mathcal{X}$ is the d-dimensional simplex $\{(p_1,\dots,p_d) \in [0,1]^d$ such that $p_1 + \cdots + p_d = 1\}$. For instance, audio and video data are often represented by histogram features (Oliva and Torralba, 2001; Lowe, 2004; Rabiner and Schäfer, 2007); see also an earlier version of the present paper (Arlot et al., 2012, Section 6.2) in which we considered the problem of segmenting video streams. In such cases, it is a bad idea to do as if $\mathcal{X}$ were $\mathbb{R}^d$-valued, since the Euclidean norm on $\mathbb{R}^d$ is usually a bad distance measure between histogram data.*

Example 5 *The set $\mathcal{X}$ is a set of graphs. For instance, the X_i can represent a social network (Kossinets and Watts, 2006) or a biological network (Curtis et al., 2012) that is changing over time (Chen and Zhang, 2015). Then, detecting meaningful changes in the structure of a time-varying network is a change-point problem. In the case of social networks, this can be used for detecting the rise of an economic crisis (McCulloh, 2009).*

Example 6 *The set $\mathcal{X}$ is a set of texts (strings). For instance, text analysis can try to localize possible changes of authorship within a given text (Chen and Zhang, 2015).*

Example 7 *The set $\mathcal{X}$ is a subset of $\{A,T,C,G\}^{\mathbb{N}}$, the set of DNA sequences. For instance, an important question in phylogenetics is to find recombination events from the genome of individuals of a given species (Knowles and Kubatko, 2010; Ané, 2011). This can be achieved from a multiple alignment of DNA sequences (Schölkopf et al., 2004) by detecting abrupt changes (change-points) in the phylogenetic tree at each DNA position, that is, by solving a change-point problem.*

Example 8 *The set $\mathcal{X}$ is an infinite-dimensional functional space. Such functional data arise in various fields (see for instance Ferraty and Vieu, 2006, Chapter 2), and the problem*

of testing whether there is a change or not in a functional time series has been considered recently (Ferraty and Vieu, 2006; Berkes et al., 2009; Sharipov et al., 2014).

Other kinds of data could be considered, such as counting data (Cleynen and Lebarbier, 2014b; Alaya et al., 2015), qualitative descriptors, as well as composite data, that is, data X_i that are mixing several above examples.

The goal of the paper is to propose a change-point algorithm that is (i) general enough to handle all these situations (up to the choice of an appropriate similarity measure on $\mathcal{X}$), (ii) in a non parametric framework, (iii) with an unknown number of change-points, and (iv) that we can analyze theoretically in all these examples simultaneously.

Note also that we want our algorithm to output a set of change-points that are "close to" the true ones, at least when n is large enough. But in settings where the signal-to-noise ratio is not large enough to recover all true change-points (for a given n), we do not want to have false positives. This motivates the non-asymptotic analysis of our algorithm that we make in this paper. Since our algorithm relies on a model selection procedure, we prove in Section 4 an oracle inequality, as usually done in non-asymptotic model selection theory.

3 Detecting changes in the distribution with kernels

Our approach for solving the general change-point problem uses positive semidefinite kernels. It can be sketched as follows.

3.1 Kernel change-point algorithm

For any integer $D \in [\![1, n+1]\!]$, the set of sequences of $(D-1)$ change-points is defined by

$$\mathcal{T}_n^D := \left\{(\tau_0, \ldots, \tau_D) \in \mathbb{N}^{D+1} \,/\, 0 = \tau_0 < \tau_1 < \tau_2 < \cdots < \tau_D = n\right\} \tag{1}$$

where $\tau_1, \ldots, \tau_{D-1}$ are the change-points, and τ_0, τ_D are just added for notational convenience. Any $\tau \in \mathcal{T}_n^D$ is called a *segmentation* (of $\{1, \ldots, n\}$) into $D_\tau := D$ segments.

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semidefinite kernel, that is, a measurable function $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that for any $x_1, \ldots, x_n \in \mathcal{X}$, the $n \times n$ matrix $(k(x_i, x_j))_{1 \leq i,j \leq n}$ is positive semidefinite. Examples of such kernels are given in Section 3.2. Then, we measure the quality of any candidate segmentation $\tau \in \mathcal{T}_n^D$ with the *kernel least-squares criterion* introduced by Harchaoui and Cappé (2007):

$$\widehat{\mathcal{R}}_n(\tau) := \frac{1}{n}\sum_{i=1}^{n} k(X_i, X_i) - \frac{1}{n}\sum_{\ell=1}^{D}\left[\frac{1}{\tau_\ell - \tau_{\ell-1}}\sum_{i=\tau_{\ell-1}+1}^{\tau_\ell}\sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} k(X_i, X_j)\right]. \tag{2}$$

In particular when $\mathcal{X} = \mathbb{R}$ and $k(x, y) = xy$, we recover the usual least-squares criterion

$$\widehat{\mathcal{R}}_n(\tau) = \frac{1}{n}\sum_{\ell=1}^{D}\sum_{i=\tau_{\ell-1}+1}^{\tau_\ell}\left(X_i - \overline{X}_{[\![\tau_{\ell-1}+1, \tau_\ell]\!]}\right)^2 \quad \text{where} \quad \overline{X}_{[\![\tau_{\ell-1}+1, \tau_\ell]\!]} := \frac{1}{\tau_\ell - \tau_{\ell-1}}\sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} X_j\,.$$

Note that Eq. (6) in Section 4.1 provides an equivalent formula for $\widehat{\mathcal{R}}_n(\tau)$, which is helpful for understanding its meaning. Given the criterion (2), we cast the choice of τ as a model selection problem (as thoroughly detailed in Section 4), which leads to Algorithm 1 below, that we now briefly comment on.

Input: observations: $X_1, \ldots, X_n \in \mathcal{X}$,
kernel: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$,
constants: $c_1, c_2 > 0$ and $D_{\max} \in [\![1, n-1]\!]$.

Step 1: $\forall D \in [\![1, D_{\max}]\!]$, compute:
$$\widehat{\tau}(D) \in \operatorname{argmin}_{\tau \in \mathcal{T}_n^D} \big\{ \widehat{\mathcal{R}}_n(\tau) \big\} \quad \text{and} \quad \widehat{\mathcal{R}}_n\big(\widehat{\tau}(D)\big)$$
(dynamic programming).

Step 2: Find:
$$\widehat{D} \in \operatorname{argmin}_{1 \leq D \leq D_{\max}} \left\{ \widehat{\mathcal{R}}_n\big(\widehat{\tau}(D)\big) + \frac{D}{n}\Big(c_1 \log\Big[\frac{n}{D}\Big] + c_2 \Big) \right\}.$$

Output: sequence of change-points: $\widehat{\tau} = \widehat{\tau}\big(\widehat{D}\big)$.

Algorithm 1: kernel change-point algorithm (KCP)

- Step 1 of Algorithm 1 consists in choosing the "best" segmentation with D segments, that is, the minimizer of the kernel least-squares criterion $\widehat{\mathcal{R}}_n(\cdot)$ over $\mathcal{T}_n^D$, for every $D \in [\![1, D_{\max}]\!]$.

- Step 2 of Algorithm 1 chooses D by model selection, using a penalized empirical criterion. A major contribution of this paper lies in the building and theoretical justification of the penalty $\frac{D}{n}(c_1 \log(\frac{n}{D}) + c_2)$, see Sections 4–5.

- Practical issues (computational complexity and choice of constants $c_1, c_2, D_{\max}$) are discussed in Section 3.3. Let us only emphasize here that Algorithm 1 is tractable; its most expensive part is the minimization problem of Step 1, which can be done by dynamic programming (see Harchaoui and Cappé, 2007; Celisse et al., 2016).

3.2 Examples of kernels

Algorithm 1 can be used with various sets $\mathcal{X}$ (not necessarily vector spaces) as long as a positive semidefinite kernel on $\mathcal{X}$ is available. An important issue is to design relevant kernels, that are able to capture important features of the data for a given change-point problem, including non-vectorial data—for instance, simplicial data (histograms), texts or graphs (networks), see Section 2. The question of choosing a kernel is discussed in Section 7.2.

Classical kernels can be found in the books by Schölkopf and Smola (2001), Shawe-Taylor and Cristianini (2004) and Schölkopf et al. (2004) for instance. Let us mention a few of them:

- When $\mathcal{X} = \mathbb{R}^d$, $k^{\mathrm{lin}}(x, y) = \langle x, y \rangle_{\mathbb{R}^d}$ defines the *linear kernel*. When $d = 1$, Algorithm 1 coincides with the algorithm proposed by Lebarbier (2005).
- When $\mathcal{X} = \mathbb{R}^d$, $k_h^{\mathrm{G}}(x, y) = \exp[-\|x - y\|^2 / (2h^2)]$ defines the *Gaussian kernel* with bandwidth $h > 0$, which is used in the experiments of Section 6.
- When $\mathcal{X} = \mathbb{R}^d$, $k_h^L(x, y) = \exp[-\|x - y\|/(2h^2)]$ defines the *Laplace kernel* with bandwidth $h > 0$.
- When $\mathcal{X} = \mathbb{R}^d$, $k_h^{\mathrm{e}}(x, y) = \exp(\langle x, y \rangle_{\mathbb{R}^d}/h)$ defines the *exponential kernel* with bandwidth $h > 0$. Note that, unlike the Gaussian and Laplace kernels, the exponential kernel is not translation-invariant.
- When $\mathcal{X} = \mathbb{R}$, $k_h^{\mathrm{H}}(x, y) = \sum_{j=1}^5 H_{j,h}(x) H_{j,h}(y)$, corresponds to the Hermite kernel, where $H_{j,h}(x) = 2^{j+1}\sqrt{\pi j!}\mathrm{e}^{-x^2/(2h^2)}(-1)^j \mathrm{e}^{-x^2/2}(\partial/\partial x)^j(\mathrm{e}^{-x^2/2})$ denotes the j-th Hermite function with bandwidth $h > 0$. This kernel is used in Section 6.
- When $\mathcal{X}$ is the d-dimensional simplex as in Example 4, the χ^2-kernel can be defined by $k_{\chi^2}(x, y) = \exp\left(-\frac{1}{2}\sum_{i=1}^d \frac{(x_i - y_i)^2}{x_i + y_i}\right)$. An illustration of its behavior is provided in the simulation experiments of Section 6.

Note that more generally, Sejdinovic et al. (2013) proved that positive semidefinite kernels can be defined on any set $\mathcal{X}$ for which a semimetric of negative type is used to measure closeness between points. The so-called *energy distance* between probability measures is an example (Matteson and James, 2014). In addition, specific kernels have been designed for various kinds of structured data, including all the examples of Section 2 (Cuturi et al., 2005; Shervashidze, 2012).

Let us finally remark that Algorithm 1 can also be used when k is not a positive semidefinite kernel; its computational complexity remains unchanged, but we might loose the theoretical guarantees of Section 4.

3.3 Practical issues

Computational complexity. The discrete optimization problem at Step 1 of Algorithm 1 is apparently hard to solve since, for each D, there are $\binom{n-1}{D-1}$ segmentations of $\{1, \ldots, n\}$ into D segments. Fortunately, as suggested by Harchaoui and Cappé (2007), this optimization problem can be solved efficiently by dynamic programming (Auger and Lawrence, 1989; Kay, 1993): denoting by $\mathcal{C}_k$ the cost of computing $k(x, y)$ for some given $x, y \in \mathcal{X}$, the computational cost of Step 1 then is $\mathcal{O}(\mathcal{C}_k n^2 + D_{\max} n^4)$ in time and $\mathcal{O}(D_{\max} n + n^2)$ in space. Note that the $\mathcal{O}(D_{\max} n^4)$ part of the time complexity results from the necessary computation of the so-called *cost matrix*. The coefficient (i, j) of this $n \times n$ cost matrix is equal to the statistical cost of the segment $[\![i, j - 1]\!]$, which involves itself summing over a quadratic number of terms of the Gram matrix. By a careful optimization of the interplay between dynamic programming and the cost matrix computation, Celisse et al. (2016) reduce the computational complexity to $\mathcal{O}((\mathcal{C}_k + D_{\max})n^2)$ in time and $\mathcal{O}(D_{\max} n)$ in space.

For given constants $D_{\max}$ and c_1, c_2, Step 2 is straightforward since it consists in a minimization problem among $D_{\max}$ terms already stored in memory. Therefore, the overall complexity of Algorithm 1 is at most $\mathcal{O}((\mathcal{C}_k + D_{\max})n^2)$ in time and $\mathcal{O}(D_{\max} n)$ in space.

Setting the constants c_1, c_2. At Step 2 of Algorithm 1, two constants $c_1, c_2 > 0$ appear in the penalty term. Theoretical guarantees (Theorem 2 in Section 4) suggest to take $c_1 = c_2 = c$ large enough, but the lower bound on c in Theorem 2 is pessimistic, and the optimal value of c certainly depends on unknown features of the data such as their "variance", as discussed after Theorem 2. In practice the constants c_1, c_2 must be chosen from data. To do so, we propose a fully data-driven method, based upon the "slope heuristic" (Baudry et al., 2012), that is explained in Section 6.2. Another way of choosing c_1, c_2 is described in supplementary material (Section B.3).

Setting the constant $D_{\max}$. Algorithm 1 requires to specify the maximal dimension $D_{\max}$ of the segmentations considered, a choice that has three main consequences. First, the computational complexity of Algorithm 1 is affine in $D_{\max}$, as discussed above. Second, if $D_{\max}$ is too small—smaller than the number of true change-points that can be detected—, the segmentation $\widehat{\tau}$ provided by the algorithm will necessarily be too coarse. Third, when the slope heuristic is used for choosing c_1, c_2, taking $D_{\max}$ larger than the true number of change-points might not be sufficient: better values for c_1, c_2 can be obtained by taking $D_{\max}$ larger, up to n. From our experiments, it seems that $D_{\max} \approx n/\sqrt{\log n}$ is large enough to provide good results.

3.4 Related change-point algorithms

In addition to the references given in the Introduction, let us mention a few change-point algorithms to which Algorithm 1 is more closely related.

First, some two-sample (or homogeneity) tests based on kernels have been suggested. They tackle a simpler problem than the general change-point problem described in Section 2. Among them, Gretton et al. (2012a) proposed a two-sample test based on a U-statistic of order two, called the maximum mean discrepancy (MMD). A related family of two-sample tests, called B-tests, has been proposed by Zaremba et al. (2013); B-tests have also been used by Li et al. (2015) for localizing a single change-point. Harchaoui et al. (2008) proposed a studentized kernel-based test statistic for testing homogeneity. Resampling methods—(block) bootstrap and permutations—have also been proposed for choosing the threshold of several kernel two-sample tests (Fromont et al., 2012; Chwialkowski et al., 2014; Sharipov et al., 2014).

Second, Harchaoui and Cappé (2007) proposed a kernel change-point algorithm when the true number of segments $D^\star$ is known, which corresponds to Step 1 of Algorithm 1. The present paper proposes a data-driven choice of D for which theoretical guarantees are proved.

Third, when $\mathcal{X} = \mathbb{R}$ and $k(x, y) = xy$, $\widehat{\mathcal{R}}_n(\tau)$ is the usual least-squares risk and Step 2 of Algorithm 1 is similar to the penalization procedures proposed by Comte and Rozenholc (2004) and Lebarbier (2005) for detecting changes in the mean of a one-dimensional signal. We refer readers familiar with model selection techniques to Section 4.1 for an equivalent formulation of Algorithm 1—in more abstract terms—that clearly emphasizes the links between Algorithm 1 and these penalization procedures.

4 Theoretical analysis

We now provide theoretical guarantees for Algorithm 1. We start by reformulating it in an abstract way, which enlightens how it works.

4.1 Abstract formulation of the algorithm

Let $\mathcal{H} = \mathcal{H}_k$ denote the reproducing kernel Hilbert space (RKHS) associated with the positive semidefinite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The canonical feature map $\Phi : \mathcal{X} \mapsto \mathcal{H}$ is then defined by $\Phi(x) = k(x, \cdot) \in \mathcal{H}$ for every $x \in \mathcal{X}$. A detailed presentation of positive semidefinite kernels and related notions can be found in several books (Schölkopf and Smola, 2001; Cucker and Zhou, 2007; Steinwart and Christmann, 2008).

Let us define $Y_i = \Phi(X_i) \in \mathcal{H}$ for every $i \in \{1, \dots, n\}$, $Y = (Y_i)_{1 \leq i \leq n} \in \mathcal{H}^n$, $\mathcal{T}_n := \bigcup_{D=1}^{n} \mathcal{T}_n^D$ the set of segmentations (see Eq. (1)), and for every $\tau \in \mathcal{T}_n$,

$$F_\tau := \left\{ f = (f_1, \dots, f_n) \in \mathcal{H}^n \text{ s.t. } f_{\tau_{\ell-1}+1} = \cdots = f_{\tau_\ell} \quad \forall 1 \leq \ell \leq D_\tau \right\} \ , \tag{3}$$

which is a linear subspace of $\mathcal{H}^n$. We also define on $\mathcal{H}^n$ the canonical scalar product by $\langle f, g \rangle := \sum_{i=1}^{n} \langle f_i, g_i \rangle_{\mathcal{H}}$ for $f, g \in \mathcal{H}^n$, and we denote by $\|\cdot\|$ the corresponding norm. Then, for any $g \in \mathcal{H}^n$,

$$\Pi_\tau g := \operatorname{argmin}_{f \in F_\tau} \left\{ \|f - g\|^2 \right\} \tag{4}$$

is the orthogonal projection of $g \in \mathcal{H}^n$ onto F_τ, and satisfies

$$\forall g \in \mathcal{H}^n, \forall 1 \leq \ell \leq D_\tau \, , \forall i \in [\![\tau_{\ell-1} + 1, \tau_\ell]\!], \quad (\Pi_\tau g)_i = \frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} g_j \ . \tag{5}$$

The proof of this statement has been deferred to Appendix A.1.

Following Harchaoui and Cappé (2007), the empirical risk $\widehat{\mathcal{R}}_n(\tau)$ defined by Eq. (2) can be rewritten as

$$\widehat{\mathcal{R}}_n(\tau) = \frac{1}{n} \|Y - \widehat{\mu}_\tau\|^2 \qquad \text{where} \qquad \widehat{\mu}_\tau = \Pi_\tau Y \ , \tag{6}$$

as proved in Appendix A.1.

For each $D \in [\![1, D_{\max}]\!]$, Step 1 of Algorithm 1 consists in finding a segmentation $\widehat{\tau}(D)$ in D segments such that

$$\widehat{\tau}(D) \in \operatorname{argmin}_{\tau \in \mathcal{T}_n^D} \left\{ \left\| Y - \widehat{\mu}_\tau \right\|^2 \right\} = \operatorname{argmin}_{\tau \in \mathcal{T}_n^D} \left\{ \inf_{f \in F_\tau} \sum_{i=1}^{n} \left\| \Phi(X_i) - f_i \right\|^2 \right\} \ ,$$

which is the "kernelized" version of the classical least-squares change-point algorithm (Lebarbier, 2005). Since the penalized criterion of Step 2 is similar to that of Comte and Rozenholc (2004) and Lebarbier (2005), we can see Algorithm 1 as a "kernelization" of these penalized least-squares change-point procedures.

Let us emphasize that building a theoretically-grounded penalty for such a kernel least-squares change-point algorithm is not straightforward. For instance, we cannot apply the model selection results by Birgé and Massart (2001) that were used by Comte and Rozenholc (2004) and Lebarbier (2005). Indeed, a Gaussian homoscedastic assumption is not realistic for general Hilbert-valued data, and we have to consider possibly heteroscedastic data for which we assume only that $Y_i = \Phi(X_i)$ is bounded in $\mathcal{H}$ (see Assumption (**Db**) in Section 4.3). Note that unbounded data X_i can satisfy Assumption (**Db**), for instance by choosing a bounded kernel such as the Gaussian or Laplace ones. In addition, dealing with Hilbert-valued random variables instead of (multivariate) real variables requires a new concentration inequality, see Proposition 1 in Section 4.4.

4.2 Intuitive analysis

Section 4.1 shows that Algorithm 1 can be seen as a kernelization of change-point algorithms focusing on changes of the mean of the signal (Lebarbier, 2005, for instance). Therefore, Algorithm 1 is looking for changes in the "mean" of $Y_i = \Phi(X_i) \in \mathcal{H}$, provided that such a notion can be defined.

If $\mathcal{H}$ is separable and $\mathbb{E}[k(X_i, X_i)] < +\infty$, we can define the (Bochner) mean $\mu_i^\star \in \mathcal{H}$ of $\Phi(X_i)$ (Ledoux and Talagrand, 1991), also called the mean element of P_{X_i}, by

$$\forall g \in \mathcal{H}, \qquad \langle \mu_i^\star,\, g \rangle_{\mathcal{H}} = \mathbb{E}\big[g(X_i)\big] = \mathbb{E}\big[\langle Y_i,\, g \rangle_{\mathcal{H}}\big] \ . \tag{7}$$

Then, we can write

$$\forall 1 \leq i \leq n, \qquad Y_i = \mu_i^\star + \varepsilon_i \in \mathcal{H} \qquad \text{where} \qquad \varepsilon_i := Y_i - \mu_i^\star \ .$$

The variables $(\varepsilon_i)_{1 \leq i \leq n}$ are independent and centered—that is, $\forall g \in \mathcal{H}$, $\mathbb{E}[\langle \varepsilon_i,\, g \rangle_{\mathcal{H}}] = 0$. So, we can understand $\widehat{\mu}_\tau$ as the least-squares estimator over F_τ of $\mu^\star = (\mu_1^\star, \ldots, \mu_n^\star) \in \mathcal{H}^n$.

An interesting case is when k is a *characteristic kernel* (Fukumizu et al., 2008), or equivalently, when $\mathcal{H}_k$ is *probability-determining* (Fukumizu et al., 2004a,b). Then any change in the distribution P_{X_i} induces a change in the mean element $\mu_i^\star$. In such settings, we can expect Algorithm 1 to be able to detect *any change* in the distribution P_{X_i}, at least asymptotically. For instance the Gaussian kernel is characteristic (Fukumizu et al., 2004b, Theorem 4), and general sufficient conditions for k to be characteristic are known (Sriperumbudur et al., 2010, 2011).

Note that Sharipov et al. (2014) suggest to use $k_\leq(x, y) = \mathbf{1}_{x \leq y}$ as a "kernel" within a two-sample test, in order to look for any change of the distribution of real-valued data X_i (Example 3). This idea is similar to our proposal of using Algorithm 1 with a characteristic kernel for tackling Example 3, even if we do not advise to take $k = k_\leq$ within Algorithm 1. Indeed, when $k = k_\leq$, $\widehat{\mathcal{R}}_n(\tau) = \frac{1}{2} - \frac{D_\tau}{2n}$ as soon as the X_i are all different so that Algorithm 1 becomes useless. This illustrates that using a kernel which is not symmetric positive definite should be done cautiously.

4.3 Notation and assumptions

Throughout the paper, we assume that $\mathcal{H}$ *is separable*, which is kind of a minimal assumption for two reasons: it allows to define the mean element (see Eq. (7)), and most reasonable

examples satisfy this requirement (Dieuleveut and Bach, 2014, p. 4). Let us further assume

$$\exists M \in (0,+\infty)\,, \quad \forall i \in \{1,\dots,n\}, \qquad \|Y_i\|_{\mathcal{H}}^2 = \|\Phi(X_i)\|_{\mathcal{H}}^2 = k(X_i,X_i) \leq M^2 \quad \text{a.s.} \quad \textbf{(Db)}$$

For every $1 \leq i \leq n$, we also define the "variance" of Y_i by

$$v_i := \mathbb{E}\Big[\big\|\Phi(X_i) - \mu_i^\star\big\|_{\mathcal{H}}^2\Big] = \mathbb{E}\big[k(X_i,X_i)\big] - \big\|\mu_i^\star\big\|_{\mathcal{H}}^2 = \mathbb{E}\big[k(X_i,X_i) - k(X_i,X_i')\big] \qquad (8)$$

where X_i' is an independent copy of X_i, and $v_{\max} := \max_{1\leq i\leq n} v_i$. Let us make a few remarks.

- If (**Db**) holds true, then the mean element $\mu_i^\star$ exists since $\mathbb{E}[\sqrt{k(X_i,X_i)}] < \infty$, the variances v_i are finite and $v_{\max} \leq M^2$.
- If (**Db**) holds true, then Y_i admits a covariance operator Σ_i that is trace-class and $v_i = \operatorname{tr}(\Sigma_i)$.
- If k is translation invariant, that is, $\mathcal{X}$ is a vector space and $k(x,x') = \overline{k}(x-x')$ for every $x, x' \in \mathcal{X}$, and some measurable function $\overline{k} : \mathcal{X} \to \mathbb{R}$, then (**Db**) holds true with $M^2 = k(0)$ and $v_i = k(0) - \|\mu_i^\star\|_{\mathcal{H}}^2$. For instance the Gaussian and Laplace kernels are translation invariant (see Section 3.2).
- Let us consider the case of the linear kernel $(x,y) \mapsto \langle x,\, y\rangle$ on $\mathcal{X} = \mathbb{R}^d$. If $\mathbb{E}[\|X_i\|_{\mathbb{R}^d}^2] < \infty$, then, $v_i = \operatorname{tr}(\Sigma_i)$ where Σ_i is the covariance matrix of X_i. In addition, (**Db**) holds true if and only if $\|X_i\|_{\mathbb{R}^d} \leq M$ a.s. for all i.

4.4 Concentration inequality for some quadratic form of Hilbert-valued random variables

Our main theoretical result, stated in Section 4.5, relies on two concentration inequalities for some linear and quadratic functionals of Hilbert-valued vectors. Here we state the concentration result that we prove for the quadratic term, which is significantly different from existing results and can be of independent interest.

Proposition 1 (Concentration of the quadratic term) *Let $\tau \in \mathcal{T}_n$ and recall that Π_τ is the orthogonal projection onto F_τ in $\mathcal{H}^n$ defined by Eq. (4). Let $X_1,\dots,X_n$ be independent $\mathcal{X}$-valued random variables and assume that (**Db**) holds true, so that we can define $\varepsilon = (\varepsilon_1,\dots,\varepsilon_n) \in \mathcal{H}^n$ as in Section 4.1. Then for every $x > 0$, with probability at least $1 - \mathrm{e}^{-x}$,*

$$\|\Pi_\tau\varepsilon\|^2 - \mathbb{E}\big[\|\Pi_\tau\varepsilon\|^2\big] \leq \frac{14M^2}{3}\Big(x + 2\sqrt{2xD_\tau}\Big)\ .$$

Proposition 1 is proved in Section 5.4. The proof relies on a combination of Bernstein's and Pinelis-Sakhanenko's inequalities. Note that the proof of Proposition 1 also shows that for every $x > 0$, with probability at least $1 - \mathrm{e}^{-x}$,

$$\|\Pi_\tau\varepsilon\|^2 - \mathbb{E}\big[\|\Pi_\tau\varepsilon\|^2\big] \geq -\frac{14M^2}{3}\Big(x + 2\sqrt{2xD_\tau}\Big)\ .$$

Previous concentration results for quantities such as $\|\Pi_\tau\varepsilon\|^2$ or $\|\Pi_\tau\varepsilon\|$ do not imply Proposition 1—even up to numerical constants. Indeed, they either assume that ε is a Gaussian vector, or they involve much larger deviation terms (see Section 5.4.3 for a detailed discussion of these results).

4.5 Oracle inequality for Algorithm 1

Similarly to the results of Comte and Rozenholc (2004) and Lebarbier (2005) in the one-dimensional case, we state below a non-asymptotic oracle inequality for Algorithm 1. First, we define the quadratic risk of any $\mu \in \mathcal{H}^n$ as an estimator of $\mu^\star$ by

$$\mathcal{R}(\mu) = \frac{1}{n}\|\mu - \mu^\star\|^2 = \frac{1}{n}\sum_{i=1}^{n}\|\mu_i - \mu_i^\star\|_{\mathcal{H}}^2.$$

Theorem 2 *We consider the framework and notation introduced in Sections 2–4. Let $C \geq 0$ be some constant. Assume that* (**Db**) *holds true and that* $\mathrm{pen} : \mathcal{T}_n \to \mathbb{R}$ *is some penalty function satisfying*

$$\forall \tau \in \mathcal{T}_n, \qquad \mathrm{pen}(\tau) \geq \frac{CM^2}{n}\left[D_\tau + \log\binom{n-1}{D_\tau - 1}\right] \ . \tag{9}$$

Then, some numerical constant $L_1 > 0$ exists such that the following holds: if $C \geq L_1$, for every $y \geq 0$, an event of probability at least $1 - \mathrm{e}^{-y}$ exists on which, for every

$$\widehat{\tau} \in \mathrm{argmin}_{\tau \in \mathcal{T}_n}\left\{\widehat{\mathcal{R}}_n(\tau) + \mathrm{pen}(\tau)\right\} \ , \tag{10}$$

we have

$$\mathcal{R}(\widehat{\mu}_{\widehat{\tau}}) \leq 2 \inf_{\tau \in \mathcal{T}_n}\left\{\mathcal{R}(\widehat{\mu}_\tau) + \mathrm{pen}(\tau)\right\} + \frac{83yM^2}{n} \ . \tag{11}$$

Theorem 2 is proved in Section 5.5. In addition, Section 5.1 provides some details about the construction of the penalty suggested by Eq. (9).

Theorem 2 applies to the segmentation $\widehat{\tau}$ provided by Algorithm 1 when $c_1, c_2 \geq 2L_1M^2$ since for any $D \in \{1, \ldots, n\}$,

$$\binom{n-1}{D-1} = \frac{D}{n}\binom{n}{D} \leq \binom{n}{D} \leq \frac{n^D}{D!} \leq \left(\frac{n\mathrm{e}}{D}\right)^D \ .$$

Theorem 2 shows that $\widehat{\mu}_{\widehat{\tau}}$ estimates well the "mean" $\mu^\star \in \mathcal{H}^n$ of the transformed time series $Y_1 = \Phi(X_1), \ldots, Y_n = \Phi(X_n)$. Such a non-asymptotic oracle inequality is the usual way to theoretically validate a model selection procedure (Birgé and Massart, 2001, for instance). It is therefore a natural way to theoretically validate any change-point detection procedure based on model selection. As argued by Lebarbier (2005) for instance, proving such a non-asymptotic result is necessary for taking into account situations where some changes are too small to be detected—they are "below the noise level". By defining the performance of $\widehat{\tau}$ as the quadratic risk of $\widehat{\mu}_{\widehat{\tau}}$ as an estimator of $\mu^\star$, a non-asymptotic oracle inequality such as Eq. (11) is the natural way to prove that Algorithm 1 works well for finite sample size and for a set $\mathcal{X}$ that can have a large dimensionality. The possible consistency of Algorithm 1 for estimating the change-point locations, which is outside the scope of this paper, is discussed in Section 7.1.

The constant 2 in front of the first term in Eq. (11) has no special meaning, and could be replaced by any quantity strictly larger than 1, at the price of enlarging L_1 and 83.

The value $2L_1M^2$ suggested by Theorem 2 for the constants c_1, c_2 in Algorithm 1 should not be used in practice because it is likely to lead to a conservative choice for two reasons. First, the minimal value L_1 for the constant C suggested by the proof of Theorem 2 depends on the numerical constants appearing in the deviation bounds of Propositions 1 and 3, which probably are not optimal. Second, the constant M^2 in the penalty is probably pessimistic in several frameworks. For instance with the linear kernel and Gaussian data belonging to $\mathcal{X} = \mathbb{R}$, (**Db**) is not satisfied, but other similar oracle inequalities have been proved with M^2 replaced by the residual variance (Lebarbier, 2005). In practice, as we do in the experiments of Section 6, we recommend to use a data-driven value for the leading constant C in the penalty, as explained in Section 3.3.

A nice feature of Theorem 2 is that it holds under mild assumptions: we only need the data X_i to be independent and to have (**Db**) satisfied. As noticed in Section 4.3, (**Db**) holds true for translation-invariant kernel such as the Gaussian and Laplace kernels. Compared to previous results (Comte and Rozenholc, 2004; Lebarbier, 2005), we do not need the data to be Gaussian or homoscedastic. Furthermore, the independence assumption can certainly be relaxed: to do so, it would be sufficient to prove concentration inequalities similar to Propositions 1 and 3 for some dependent X_i.

In the particular setting where $\mathcal{X} = \mathbb{R}$ and k is the linear kernel $(x, y) \mapsto xy$, Theorem 2 provides an oracle inequality similar to the one proved by Lebarbier (2005) for Gaussian and homoscedastic real-valued data. The price to pay for extending this result to heteroscedastic Hilbert-valued data is rather mild: we only assume (**Db**) and replace the residual variance by M^2.

Apart from the results already mentioned, a few oracle inequalities have been proved for change-point procedures, for real-valued data with a multiplicative penalty (Baraud et al., 2009), for discrete data (Akakpo, 2011), for counting data with a total-variation penalty (Alaya et al., 2015), for counting data with a penalized maximum-likelihood procedure (Cleynen and Lebarbier, 2014b) and for data distributed according to an exponential family (Cleynen and Lebarbier, 2014a). Among these oracle inequalities, only the result by Akakpo (2011) is more precise than Theorem 2 (there is no $\log(n)$ factor compared to the oracle loss), at the price of using a smaller (dyadic) collection of possible segmentations, hence a worse oracle performance in general.

5 Main proofs

We now prove the main results of the paper, Theorem 2 and Proposition 1.

5.1 Outline of the proof of Theorem 2

As usual for proving an oracle inequality (see Arlot, 2014, Section 2.2), we remark that by Eq. (10), for every $\tau \in \mathcal{T}_n$,

$$\widehat{\mathcal{R}}_n(\widehat{\mu}_{\widehat{\tau}}) + \mathrm{pen}(\widehat{\tau}) \leq \widehat{\mathcal{R}}_n(\widehat{\mu}_{\tau}) + \mathrm{pen}(\tau) \enspace .$$

Therefore,

$$\mathcal{R}(\widehat{\mu}_{\widehat{\tau}}) + \mathrm{pen}(\widehat{\tau}) - \mathrm{pen}_{\mathrm{id}}(\widehat{\tau}) \leq \mathcal{R}(\widehat{\mu}_\tau) + \mathrm{pen}(\tau) - \mathrm{pen}_{\mathrm{id}}(\tau) \tag{12}$$

$$\text{where} \qquad \forall \tau \in \mathcal{T}, \qquad \mathrm{pen}_{\mathrm{id}}(\tau) := \mathcal{R}(\widehat{\mu}_\tau) - \widehat{\mathcal{R}}_n(\widehat{\mu}_\tau) + \frac{1}{n}\left\|\varepsilon\right\|^2 \ . \tag{13}$$

The idea of the proof is that if we had $\mathrm{pen}(\tau) \geq \mathrm{pen}_{\mathrm{id}}(\tau)$ for every $\tau \in \mathcal{T}_n$, we would get an oracle inequality similar to Eq. (11). What remains to obtain is a deterministic upper bound on the *ideal penalty* $\mathrm{pen}_{\mathrm{id}}(\tau)$ that holds true simultaneously for all $\tau \in \mathcal{T}_n$ on a large probability event. To this aim, our approach is to compute $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(\tau)]$ and to show that $\mathrm{pen}_{\mathrm{id}}(\tau)$ concentrates around its expectation for every $\tau \in \mathcal{T}_n$ (Sections 5.2–5.4). Then we use a union bound as detailed in Section 5.5. A similar strategy has been used for instance by Comte and Rozenholc (2004) and Lebarbier (2005) in the specific context of change-point detection.

Note that we prove below a slightly weaker result than $\mathrm{pen}(\tau) \geq \mathrm{pen}_{\mathrm{id}}(\tau)$, which is nevertheless sufficient to obtain Eq. (11). Remark also that Eq. (12) would be true if the constant $n^{-1}\left\|\varepsilon\right\|^2$ in the definition (13) of $\mathrm{pen}_{\mathrm{id}}$ was replaced by any quantity independent from τ; the reasons for this specific choice appear in the computations below.

5.2 Computation of the ideal penalty

From Eq. (13) it results that for every $\tau \in \mathcal{T}_n$,

$$\begin{aligned}
n \times \mathrm{pen}_{\mathrm{id}}(\tau) &= \left\|\widehat{\mu}_\tau - \mu^\star\right\|^2 - \left\|\widehat{\mu}_\tau - Y\right\|^2 + \left\|\varepsilon\right\|^2 \\
&= \left\|\widehat{\mu}_\tau - \mu^\star\right\|^2 - \left\|\widehat{\mu}_\tau - \mu^\star - \varepsilon\right\|^2 + \left\|\varepsilon\right\|^2 \\
&= 2\left\langle \widehat{\mu}_\tau - \mu^\star,\, \varepsilon \right\rangle \\
&= 2\big\langle \Pi_\tau(\mu^\star + \varepsilon) - \mu^\star,\, \varepsilon \big\rangle \\
&= 2\left\langle \Pi_\tau \mu^\star - \mu^\star,\, \varepsilon \right\rangle + 2\left\langle \Pi_\tau \varepsilon,\, \varepsilon \right\rangle \\
&= 2\left\langle \Pi_\tau \mu^\star - \mu^\star,\, \varepsilon \right\rangle + 2\left\|\Pi_\tau \varepsilon\right\|^2
\end{aligned} \tag{14}$$

since Π_τ is an orthogonal projection. The next two sections focus separately on the two terms appearing in Eq. (14).

5.3 Concentration of the linear term

We prove in Section A.2 the following concentration inequality for the linear term in Eq. (14), mostly by applying Bernstein's inequality.

Proposition 3 (Concentration of the linear term) *If* (**Db**) *holds true, then for every* $x > 0$*, with probability at least* $1 - 2\mathrm{e}^{-x}$*,*

$$\forall \theta > 0\,, \quad \left|\big\langle (I - \Pi_\tau)\mu^\star,\, \Phi(\mathbf{X}) - \mu^\star \big\rangle\right| \leq \theta \left\|\Pi_\tau \mu^\star - \mu^\star\right\|^2 + \left(\frac{v_{\max}}{2\theta} + \frac{4M^2}{3}\right) x \ . \tag{15}$$

5.4 Dealing with the quadratic term

We now focus on the quadratic term in the right-hand side of Eq. (14).

5.4.1 Preliminary computations

We start by providing a useful closed-form formula for $\|\Pi_\tau \varepsilon\|^2$ and by computing its expectation. First, a straightforward consequence of Eq. (5) is that

$$\|\Pi_\tau \varepsilon\|^2 = \sum_{\ell=1}^{D_\tau} \left[\frac{1}{\tau_\ell - \tau_{\ell-1}} \left\| \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_i \right\|_{\mathcal{H}}^2 \right] \tag{16}$$

$$= \sum_{\ell=1}^{D_\tau} \left[\frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{\tau_{\ell-1}+1 \le i,j \le \tau_\ell} \langle \varepsilon_i, \varepsilon_j \rangle_{\mathcal{H}} \right]. \tag{17}$$

Second, we remark that for every $i, j \in \{1, \dots, n\}$,

$$\begin{aligned}
\mathbb{E}\left[\langle \varepsilon_i, \varepsilon_j \rangle_{\mathcal{H}}\right] &= \mathbb{E}\left[\langle \Phi(X_i), \Phi(X_j) \rangle_{\mathcal{H}}\right] - \mathbb{E}\left[\langle \mu_i^\star, \Phi(X_j) \rangle_{\mathcal{H}}\right] - \mathbb{E}\left[\left\langle \Phi(X_i), \mu_j^\star \right\rangle_{\mathcal{H}}\right] + \left\langle \mu_i^\star, \mu_j^\star \right\rangle_{\mathcal{H}} \\
&= \mathbb{E}\left[\langle \Phi(X_i), \Phi(X_j) \rangle_{\mathcal{H}}\right] - \left\langle \mu_i^\star, \mu_j^\star \right\rangle_{\mathcal{H}} \\
&= \mathbf{1}_{i=j} \left(\mathbb{E}\left[k(X_i, X_i)\right] - \|\mu_i^\star\|_{\mathcal{H}}^2 \right) = \mathbf{1}_{i=j} v_i.
\end{aligned} \tag{18}$$

Combining Eq. (18) and (17), we get

$$\mathbb{E}\left[\|\Pi_\tau \varepsilon\|^2\right] = \sum_{\ell=1}^{D_\tau} \left[\frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} v_i \right] = \sum_{\ell=1}^{D_\tau} v_\ell^\tau \ , \tag{19}$$

where $v_\ell^\tau := \frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} v_i$.

5.4.2 Concentration: proof of Proposition 1

This proof is inspired from that of a concentration inequality by Sauvé (2009) in the context of regression with real-valued non Gaussian noise. Let us define

$$T_\ell := \frac{1}{\tau_\ell - \tau_{\ell-1}} \left\| \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_j \right\|_{\mathcal{H}}^2, \qquad \text{so that} \qquad \|\Pi_\tau \varepsilon\|^2 = \sum_{1 \le \ell \le D_\tau} T_\ell$$

by Eq. (16). Since the real random variables $(T_\ell)_{1 \le \ell \le D_\tau}$ are independent, we get a concentration inequality for their sum $\|\Pi_\tau \varepsilon\|^2$ via Bernstein's inequality (Theorem 6) as long as T_ℓ satisfies some moment conditions. The rest of the proof consists in showing such moment bounds by using Pinelis-Sakhanenko's deviation inequality (Proposition 7).

First, note that (**Db**) implies that $\|\varepsilon_i\|_{\mathcal{H}} \le 2M$ almost surely for every i by Lemma 5, hence $\|\sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_i\|_{\mathcal{H}} \le 2(\tau_\ell - \tau_{\ell-1})M$ a.s. for every $1 \le \ell \le D_\tau$. Then for every $q \ge 2$ and $1 \le \ell \le D_\tau$,

$$\mathbb{E}\left[T_\ell^q\right] = \frac{1}{(\tau_\ell - \tau_{\ell-1})^q} \int_0^{2(\tau_\ell - \tau_{\ell-1})M} 2q x^{2q-1} \mathbb{P}\left[\left\| \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_i \right\|_{\mathcal{H}} \ge x \right] \mathrm{d}x \ . \tag{20}$$

Second, since $\|\varepsilon_i\|_{\mathcal{H}} \leq 2M$ almost surely and $\mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{H}}^2\right] = v_i \leq M^2$ for every i, we get that for every $p \geq 2$ and $1 \leq \ell \leq D_\tau$,

$$\sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{H}}^p\right] \leq \frac{p!}{2}\left(\sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} v_j\right)\left(\frac{2M}{3}\right)^{p-2} \leq \frac{p!}{2} \times (\tau_\ell - \tau_{\ell-1})M^2 \times \left(\frac{2M}{3}\right)^{p-2} \ .$$

Hence, the assumptions of Pinelis-Sakhanenko's deviation inequality (Pinelis and Sakhanenko, 1986)—which is recalled by Proposition 7—are satisfied with $c = 2M/3$ and $\sigma^2 = (\tau_\ell - \tau_{\ell-1})M^2$, and we get that for every $x \in [0,\ 2(\tau_\ell - \tau_{\ell-1})M]$

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_i\right\|_{\mathcal{H}} \geq x\right) &\leq 2\exp\left(-\frac{x^2}{2\left[(\tau_\ell - \tau_{\ell-1})M^2 + \frac{2Mx}{3}\right]}\right) \\ &\leq 2\exp\left(-\frac{3x^2}{14(\tau_\ell - \tau_{\ell-1})M^2}\right) \ .\end{aligned}$$

Together with Eq. (20), we obtain that

$$\begin{aligned}\mathbb{E}\left[T_\ell^q\right] &\leq \frac{4q}{(\tau_\ell - \tau_{\ell-1})^q}\int_0^{2(\tau_\ell - \tau_{\ell-1})M} x^{2q-1}\exp\left[-\frac{3x^2}{14(\tau_\ell - \tau_{\ell-1})M^2}\right]\,\mathrm{d}x \\ &\leq 4q\left(\frac{7M^2}{3}\right)^q \int_0^{+\infty} u^{2q-1}\exp\left[-\frac{u^2}{2}\right]\,\mathrm{d}u \\ &= 2^{q-1}(q-1)! \times 4q\left(\frac{7M^2}{3}\right)^q \\ &= 2 \times (q!)\left[\frac{14M^2}{3}\right]^q \ , \end{aligned} \tag{21}$$

since for every $q \geq 1$,

$$\int_0^{+\infty} u^{2q-1}\exp(-u^2/2)\,\mathrm{d}u = 2^{q-1}(q-1)! \ .$$

Finally summing Eq. (21) over $1 \leq \ell \leq D_\tau$, it comes

$$\begin{aligned}\sum_{1\leq\ell\leq D_\tau} \mathbb{E}\left[T_\ell^q\right] &\leq 2 \times (q!)\left[\frac{14M^2}{3}\right]^q D_\tau \\ &= \frac{q!}{2} \times D_\tau \left[\frac{28M^2}{3}\right]^2 \times \left[\frac{14M^2}{3}\right]^{q-2} \ .\end{aligned}$$

Then, condition (33) of Bernstein's inequality holds true with

$$v = D_\tau\left[\frac{28M^2}{3}\right]^2 \quad \text{and} \quad c = \frac{14M^2}{3} \ .$$

Therefore, Bernstein's inequality (Massart, 2007, Proposition 2.9)—which is recalled by Proposition 6—shows that for every $x > 0$, with probability at least $1 - \mathrm{e}^{-x}$,

$$\begin{aligned}\|\Pi_\tau \varepsilon\|^2 - \mathbb{E}\left[\|\Pi_\tau \varepsilon\|^2\right] &\leq \sqrt{2vx} + cx \\ &= \sqrt{2D_\tau x}\frac{28M^2}{3} + \frac{14M^2}{3}x \\ &= \frac{14M^2}{3}\left(2\sqrt{2D_\tau x} + x\right) \quad . \qquad \blacksquare\end{aligned}$$

5.4.3 Why do we need a new concentration inequality?

We now review previous concentration results for quantities such as $\|\Pi_\tau \varepsilon\|^2$ or $\|\Pi_\tau \varepsilon\|$, showing that they are not sufficient for our needs, hence requiring a new result such as Proposition 1.

First, when $\varepsilon \in \mathbb{R}^n$ is a Gaussian isotropic vector, $\|\Pi_\tau \varepsilon\|^2$ is a chi-square random variable for which concentration tools have been developed. Such results have been used by Birgé and Massart (2001) and by Lebarbier (2005) for instance. They cannot be applied here since ε cannot be assumed Gaussian, and the ε_j do not necessarily have the same variance.

Second, Eq. (17) shows that $\|\Pi_\tau \varepsilon\|^2$ is a U-statistic of order 2. Some tight exponential concentration inequalities exist for such quantities when $\varepsilon_j \in \mathbb{R}$ (Houdré and Reynaud-Bouret, 2003) and when ε_j belongs to a general measurable set (Giné and Nickl, 2016, Theorem 3.4.8). In both results, a term of order M^2x^2 appears in the deviations, which is too large because the proof of Theorem 2 relies on Proposition 1 with $x \gg D_\tau$: we really need a smaller deviation term, as in Proposition 1 where it is proportional to M^2x.

Third, since

$$\|\Pi_\tau \varepsilon\| = \sup_{f \in \mathcal{H}^n, \|f\|=1} \left|\langle f,\, \Pi_\tau \varepsilon\rangle\right| = \sup_{f \in \mathcal{H}^n, \|f\|=1} \left|\sum_{i=1}^n \langle f_i,\, (\Pi_\tau \varepsilon)_i\rangle_{\mathcal{H}}\right| \ ,$$

Talagrand's inequality (Boucheron et al., 2013, Corollary 12.12) provides a concentration inequality for $\|\Pi_\tau \varepsilon\|$ around its expectation. More precisely, we can get the following result, which is proved in supplementary material (Section B.2).

Proposition 4 *If* (**Db**) *holds true, then for every* $x > 0$ *with probability at least* $1 - 2\mathrm{e}^{-x}$,

$$\left|\|\Pi_\tau \varepsilon\| - \mathbb{E}\big[\|\Pi_\tau \varepsilon\|\big]\right| \leq \sqrt{2x\left(4M\mathbb{E}\big[\|\Pi_\tau \varepsilon\|\big] + \max_{1 \leq \ell \leq D_\tau} v_\ell^\tau\right)} + \frac{2Mx}{3} \quad . \tag{22}$$

Therefore, in order to get a concentration inequality for $\|\Pi_\tau \varepsilon\|^2$, we have to square Eq. (22) and we necessarily get a deviation term of order M^2x^2. As with the U-statistics approach, this is too large for our needs.

Fourth, given Eq. (16), it is also natural to think of Pinelis-Sakhanenko's inequality (Pinelis and Sakhanenko, 1986), but this result alone is not precise enough because it is a *deviation* inequality, and not a *concentration* inequality. It is nevertheless a key ingredient in our proof of Proposition 1.

5.5 Oracle inequality: proof of Theorem 2

We now end the proof of Theorem 2 as explained in Section 5.1.

Upper bound on $\mathrm{pen}_{\mathrm{id}}(\tau)$ for every $\tau \in \mathcal{T}_n$. First, by Eq. (14) for every $\tau \in \mathcal{T}_n$,

$$\mathrm{pen}_{\mathrm{id}}(\tau) = \frac{1}{n}\left(\|\widehat{\mu}_\tau - \mu^\star\|^2 - \|\widehat{\mu}_\tau - Y\|^2 + \|\varepsilon\|^2\right) = \frac{2}{n}\|\Pi_\tau \varepsilon\|^2 - \frac{2}{n}\langle (I - \Pi_\tau)\mu^\star, \varepsilon\rangle \ . \quad (23)$$

In other words, $\mathrm{pen}_{\mathrm{id}}(\tau)$ is the sum of two terms, for which Propositions 1 and 3 provide concentration inequalities.

On the one hand, by Proposition 1 under (**Db**), for every $\tau \in \mathcal{T}_n$ and $x \geq 0$, with probability at least $1 - \mathrm{e}^{-x}$ we have

$$\frac{2}{n}\|\Pi_\tau \varepsilon\|^2 \leq \frac{2}{n}\left(\mathbb{E}\left[\|\Pi_\tau \varepsilon\|^2\right] + \frac{14M^2}{3}\left(x + 2\sqrt{2xD_\tau}\right)\right) \quad (24)$$

$$\leq \frac{2M^2}{n}\left(D_\tau + \frac{14x}{3} + \frac{28}{3}\sqrt{2xD_\tau}\right) \quad (25)$$

since

$$\mathbb{E}\left[\|\Pi_\tau \varepsilon\|^2\right] = \sum_{j=1}^{D_\tau} v_j^\tau \leq D_\tau M^2$$

by Eq. (19). On the other hand, by Proposition 3 under (**Db**), for every $\tau \in \mathcal{T}_n$ and $x \geq 0$, with probability at least $1 - 2\mathrm{e}^{-x}$ we have

$$\forall \theta > 0, \qquad \frac{2}{n}\left|\langle (I - \Pi_\tau)\mu^\star, \varepsilon\rangle\right| \leq \frac{2\theta}{n}\|\Pi_\tau \mu^\star - \mu^\star\|^2 + \frac{2}{n}\left(\frac{v_{\max}}{2\theta} + \frac{4M^2}{3}\right)x$$

$$\leq \frac{2\theta}{n}\|\Pi_\tau \mu^\star - \mu^\star\|^2 + \frac{xM^2}{n}\left(\theta^{-1} + \frac{8}{3}\right) \ . \quad (26)$$

For every $\tau \in \mathcal{T}_n$ and $x \geq 0$, let Ω_x^τ be the event on which Eq. (25) and (26) hold true. A union bound shows that $\mathbb{P}(\Omega_x^\tau) \geq 1 - 3\mathrm{e}^{-x}$. Furthermore, combining Eq. (23), (25) and (26) shows that on Ω_x^τ, for every $\theta > 0$,

$$\mathrm{pen}_{\mathrm{id}}(\tau) \leq \frac{2M^2}{n}\left(D_\tau + \frac{14x}{3} + \frac{28}{3}\sqrt{2xD_\tau}\right) + \frac{2\theta}{n}\|\Pi_\tau \mu^\star - \mu^\star\|^2 + \frac{xM^2}{n}\left(\theta^{-1} + \frac{8}{3}\right)$$

$$\leq 2\theta\mathcal{R}(\widehat{\mu}_\tau) + \frac{M^2}{n}\left[2D_\tau + \left(\theta^{-1} + \frac{36}{3}\right)x + \frac{56}{3}\sqrt{2xD_\tau}\right] \quad (27)$$

using that $n^{-1}\|\Pi_\tau \mu^\star - \mu^\star\|^2 = \mathcal{R}(\Pi_\tau \mu^\star) \leq \mathcal{R}(\widehat{\mu}_\tau)$ by definition of the orthogonal projection Π_τ, and

$$\mathrm{pen}_{\mathrm{id}}(\tau) \geq -\frac{2}{n}\langle (I - \Pi_\tau)\mu^\star, \varepsilon\rangle$$

$$\geq -\frac{2\theta}{n}\|\Pi_\tau \mu^\star - \mu^\star\|^2 - \frac{xM^2}{n}\left(\theta^{-1} + \frac{8}{3}\right)$$

$$\geq -2\theta\mathcal{R}(\widehat{\mu}_\tau) - \frac{xM^2}{n}\left(\theta^{-1} + \frac{8}{3}\right) \ . \quad (28)$$

Union bound over the models and conclusion. Let $y \geq 0$ be fixed and let us define the event $\Omega_y = \bigcap_{\tau \in \mathcal{T}_n} \Omega^\tau_{x(\tau,y)}$ where for every $\tau \in \mathcal{T}_n$,

$$x(\tau, y) := y + \log\left(\frac{3}{\mathrm{e}-1}\right) + D_\tau + \log\binom{n-1}{D_\tau - 1} \ .$$

Then, since

$$\mathrm{Card}\left\{\tau \in \mathcal{T}_n \,|\, D_\tau = D\right\} = \binom{n-1}{D-1}$$

for every $D \in \{1, \dots, n\}$, a union bound shows that

$$\begin{aligned}\mathbb{P}(\Omega_y) \geq 1 - \sum_{\tau \in \mathcal{T}_n} \mathbb{P}\big(\overline{\Omega}^\tau_{x(\tau,y)}\big) \geq 1 - 3\sum_{D=1}^{n} \mathrm{e}^{-y-\log\left(\frac{3}{\mathrm{e}-1}\right)-D} &= 1 - (\mathrm{e}-1)\mathrm{e}^{-y}\sum_{D=1}^{n}\mathrm{e}^{-D} \\ &\geq 1 - \mathrm{e}^{-y} \ .\end{aligned}$$

In addition, on Ω_y, for every $\tau \in \mathcal{T}_n$, since Eq. (27) and (28) hold true with $x = x(\tau, y) \geq D_\tau$, taking $\theta = 1/6$, we get that

$$-\frac{26}{3}\frac{M^2 x(\tau,y)}{n} - \frac{1}{3}\mathcal{R}(\widehat{\mu}_\tau) \leq \mathrm{pen}_{\mathrm{id}}(\tau) \leq \frac{1}{3}\mathcal{R}(\widehat{\mu}_\tau) + \left(20 + \frac{56\sqrt{2}}{3}\right)\frac{M^2 x(\tau,y)}{n} \ .$$

Let us define

$$\kappa_1 := 20 + \frac{56\sqrt{2}}{3} \qquad \text{and} \qquad \kappa_2 := \frac{26}{3} \ ,$$

and assume that $C \geq \kappa_1$. Then, using Eq. (9), we have

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(\tau) &\leq \frac{1}{3}\mathcal{R}(\widehat{\mu}_\tau) + \mathrm{pen}(\tau) + \frac{\kappa_1 M^2\left[y + \log(3/(\mathrm{e}-1))\right]}{n} \\ \mathrm{pen}_{\mathrm{id}}(\tau) &\geq -\frac{1}{3}\mathcal{R}(\widehat{\mu}_\tau) - \frac{\kappa_2}{C}\,\mathrm{pen}(\tau) - \frac{\kappa_2 M^2\left[y + \log(3/(\mathrm{e}-1))\right]}{n} \ .\end{aligned}$$

Therefore, by Eq. (12), on Ω_y, for every $\tau \in \mathcal{T}_n$,

$$\frac{2}{3}\mathcal{R}(\widehat{\mu}_{\widehat{\tau}}) - \frac{\kappa_1 M^2\left[y + \log(3/(\mathrm{e}-1))\right]}{n} \leq \frac{4}{3}\mathcal{R}(\widehat{\mu}_\tau) + \left(1 + \frac{\kappa_2}{C}\right)\mathrm{pen}(\tau) + \frac{\kappa_2 M^2\left[y + \log(3/(\mathrm{e}-1))\right]}{n}$$

hence

$$\begin{aligned}\frac{2}{3}\mathcal{R}(\widehat{\mu}_{\widehat{\tau}}) &\leq \frac{4}{3}\mathcal{R}(\widehat{\mu}_\tau) + \left(1 + \frac{\kappa_2}{C}\right)\mathrm{pen}(\tau) + \frac{(\kappa_1+\kappa_2) M^2\left[y + \log(3/(\mathrm{e}-1))\right]}{n} \\ &\leq \frac{4}{3}\mathcal{R}(\widehat{\mu}_\tau) + \left(1 + \frac{\kappa_2 + (\kappa_1+\kappa_2)\log(3/(\mathrm{e}-1))}{C}\right)\mathrm{pen}(\tau) + (\kappa_1+\kappa_2)\frac{M^2 y}{n}\end{aligned}$$

since $\mathrm{pen}(\tau) \geq CM^2/n$ for every $\tau \in \mathcal{T}_n$. Multiplying both sides by $3/2$, we get that if $C \geq \kappa_1$, on Ω_y,

$$\mathcal{R}(\widehat{\mu}_{\widehat{\tau}}) \leq \inf_{\tau \in \mathcal{T}_n}\left\{2\mathcal{R}(\widehat{\mu}_\tau) + \frac{3}{2}\left(1 + \frac{\kappa_2 + (\kappa_1+\kappa_2)\log(3/(\mathrm{e}-1))}{C}\right)\mathrm{pen}(\tau)\right\} + \frac{3(\kappa_1+\kappa_2)}{2}\frac{M^2 y}{n} \ .$$

Let us finally define

$$L_1 := 3\Big[\kappa_2 + (\kappa_1+\kappa_2)\log\big(3/(\mathrm{e}-1)\big)\Big] \geq \kappa_1$$

so that

$$\frac{3}{2}\left(1 + \frac{\kappa_2 + (\kappa_1+\kappa_2)\log(3/(\mathrm{e}-1))}{L_1}\right) = 2 \ .$$

Then, we get that if $C \geq L_1$, on Ω_y,

$$\mathcal{R}(\widehat{\mu}_{\widehat{\tau}}) \leq 2 \inf_{\tau\in\mathcal{T}_n} \{\mathcal{R}(\widehat{\mu}_\tau) + \mathrm{pen}(\tau)\} + \frac{3(\kappa_1+\kappa_2)}{2}\frac{M^2 y}{n}$$

and the result follows. ■

6 Simulation experiments

This section reports the results of some experiments on synthetic data that illustrate the performance of Algorithm 1.

6.1 Data generation process

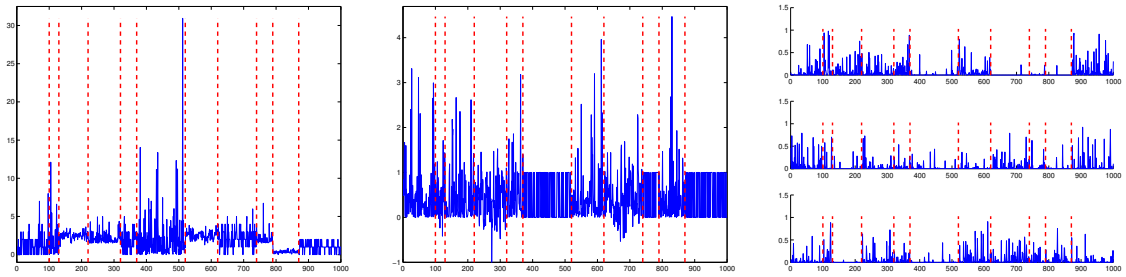
Three scenarios are considered: (*i*) real-valued data with a changing (mean,variance), (*ii*) real-valued data with constant mean and variance, and (*iii*) distribution-valued data as in Example 4.

In the three scenarios, the sample size is $n = 1\,000$ and the true segmentation $\tau^\star$ is made of $D^\star = 11$ segments, with change-points $\tau_1^\star = 100$, $\tau_2^\star = 130$, $\tau_3^\star = 220$, $\tau_4^\star = 320$, $\tau_5^\star = 370$, $\tau_6^\star = 520$, $\tau_7^\star = 620$, $\tau_8^\star = 740$, $\tau_9^\star = 790$, $\tau_{10}^\star = 870$ (see Figure 1). For each sample, we choose randomly the distribution of the X_i within each segment of $\tau^\star$ as detailed below; note that we always make sure that the distribution of X_i does change at each change-point $\tau_\ell^\star$.

For each scenario, we generate $N = 500$ independent samples, from which we estimate all quantities that are reported in Section 6.3.

Scenario 1: Real-valued data with changing (mean, variance). The distribution of $X_i \in \mathbb{R}$ is randomly picked out from: $\mathcal{B}(10, 0.2)$ (Binomial), $\mathcal{NB}(3, 0.7)$ (Negative-Binomial), $\mathcal{H}(10, 5, 2)$ (Hypergeometric), $\mathcal{N}(2.5, 0.25)$ (Gaussian), $\gamma(0.5, 5)$ (Gamma), $\mathcal{W}(5, 2)$ (Weibull) and $\mathcal{P}ar(1.5, 3)$ (Pareto). Note that the pair (mean, variance) in each segment changes from that of its neighbors. Table 1 summarizes its values.

The distribution within segment $\ell \in \{1, \ldots, D^\star\}$ is given by the realization of a random variable $S_\ell \in \{1, \ldots, 7\}$, each integer representing one of the 7 possible distributions. The variables S_ℓ are generated as follows: S_1 is uniformly chosen among $\{1, \ldots, 7\}$, and for every $\ell \in \{1, \ldots, D^\star - 1\}$, given S_ℓ, $S_{\ell+1}$ is uniformly chosen among $\{1, \ldots, 7\}\backslash\{S_\ell\}$. Figure 1a shows one sample generated according to this scenario.



(a) Scenario 1: changes in the pair (mean, variance).

(b) Scenario 2: constant mean and variance.

(c) Scenario 3: histogram-valued data (first three coordinates).

Figure 1: Examples of generated signals (blue plain curve) in the three scenarios. Red vertical dashed lines visualize the true change-points locations.

Scenario 2: Real-valued data with equal means and equal variances. The distribution of $X_i \in \mathbb{R}$ is randomly chosen among (1) $\mathcal{B}(0.5)$ (Bernoulli), (2) $\mathcal{N}(0.5, 0.25)$ (Gaussian) and (3) $\mathcal{E}(0.5)$ (Exponential). These three distributions have a mean 0.5 and a variance 0.25.

The distribution within segment $\ell \in \{1, \dots, D^\star\}$ is given by the realization of a random variable $S_\ell \in \{1, 2, 3\}$, similarly to what is done in Scenario 1 (replacing 7 by 3). Figure 1b shows one sample generated according to this scenario.

Scenario 3: Histogram-valued data. The observations X_i belong to the d-dimensional simplex with $d = 20$ (Example 4), that is, $X_i = (a_1, \dots, a_d) \in [0, 1]^d$ with $\sum_{j=1}^d a_j = 1$. For each $\ell \in \{1, \dots, D^\star\}$, we randomly generate d parameter values $p_1^\ell, \dots, p_d^\ell$ independently with uniform distribution over $[0, c_3]$ with $c_3 = 0.2$. Then within the ℓ-th segment of $\tau^\star$, X_i follows a Dirichlet distribution with parameter $(p_1^\ell, \dots, p_d^\ell)$. Figure 1c displays the first three coordinates of one sample generated according to this scenario.

6.2 Parameters of the procedure

For each sample, we apply our kernel change-point procedure (Algorithm 1) with the following choices for its parameters. We always take $D_{\max} = 100$.

For the first two scenarios, we consider three kernels:

(i) The linear kernel $k^{\text{lin}}(x, y) = xy$.

(ii) The Hermite kernel given by $k^{\text{H}}_{\sigma_H}(x, y)$ defined in Section 3.2. In Scenario 1, $\sigma_H = 1$. In Scenario 2, $\sigma_H = 0.1$.

(iii) The Gaussian kernel $k^{\text{G}}_{\sigma_G}$ defined in Section 3.2, with $\sigma_G = 0.1$.

For Scenario 3, we consider the χ^2 kernel $k_{\chi^2}(x, y)$ defined in Section 3.2, and the Gaussian kernel $k^{\text{G}}_{\sigma_G}$ with $\sigma_G = 1$.

In each scenario several candidate values have been explored for the bandwidth parameters of the above kernels. We have selected the ones with the most representative results.

For choosing the constants c_1, c_2 arising from Step 2 of Algorithm 1, we use the "slope heuristic" method, and more precisely a variant proposed by Lebarbier (2002, Section 4.3.2) for the calibration of two constants for change-point detection. We first perform a linear regression of $\widehat{\mathcal{R}}_n(\widehat{\tau}(D))$ against $\binom{n-1}{D-1}$ and D/n for $0.6 \times D_{\max} \leq D \leq D_{\max}$. Then, denoting by $\widehat{s}_1, \widehat{s}_2$ the coefficients obtained, we define $c_i = -\alpha \widehat{s}_i$ for $i = 1, 2$, with $\alpha = 2$. The slope heuristic has been justified theoretically in various settings (for instance by Arlot and Massart, 2009, for regressograms) and is supported by numerous experiments (Baudry et al., 2012). Note that we also considered other values of the constant $\alpha \in [0.8, 2.5]$; we only report here the results for $\alpha = 2$ because it corresponds to the classical advice when using the slope heuristics, and it is among the best choices for α according to our experiments.

6.3 Results

We now summarize the results of our experiments.

Distance between segmentations. In order to assess the quality of the segmentation $\widehat{\tau}$ as an estimator of the true segmentation $\tau^\star$, we consider two measures of distance between segmentations. For any $\tau, \tau' \in \mathcal{T}_n$, we define the Hausdorff distance between τ and τ' by

$$d_H(\tau, \tau') := \max\left\{ \max_{1 \leq i \leq D_\tau - 1} \min_{1 \leq j \leq D_{\tau'} - 1} \left|\tau_i - \tau'_j\right|, \max_{1 \leq j \leq D_{\tau'} - 1} \min_{1 \leq i \leq D_\tau - 1} \left|\tau_i - \tau'_j\right| \right\}$$

and the Frobenius distance between τ and τ' (see Lajugie et al., 2014) by

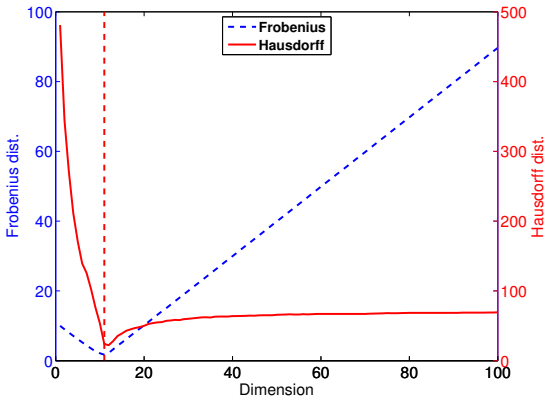
$$d_F(\tau, \tau') := \left\| M^\tau - M^{\tau'} \right\|_F = \sqrt{\sum_{1 \leq i,j \leq n} (M^\tau_{i,j} - M^{\tau'}_{i,j})^2},$$

$$\text{where} \qquad M^\tau_{i,j} = \frac{\mathbf{1}_{\{i \text{ and } j \text{ belong to the same segment of } \tau\}}}{\text{Card}(\text{segment of } \tau \text{ containing } i \text{ and } j)} \ .$$

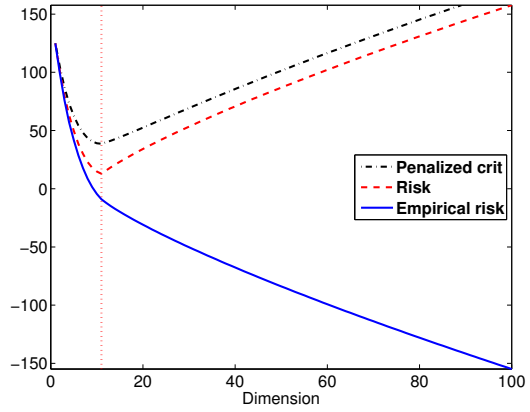
Note that $M^\tau = \Pi_\tau$ the projection matrix onto F_τ when $\mathcal{H} = \mathbb{R}$, that is, for the linear kernel on $\mathcal{X} = \mathbb{R}$. The Hausdorff distance is probably more classical in the change-point literature, but Figure 2a shows that the Frobenius distance is more informative for comparing $(\widehat{\tau}(D))_{D > D^\star}$. Indeed, when D is already a bit larger than $D^\star$, adding false change-points makes the segmentation worse without increasing much d_H; on the contrary, d_F^2 readily takes into account these additional false change-points.

Illustration of Algorithm 1. Figure 2 illustrates the typical behaviour of Algorithm 1 when k is well-suited to the change-point problem we consider. It summarizes results obtained in Scenario 1 with $k = k^{\mathrm{G}}$ the Gaussian kernel.

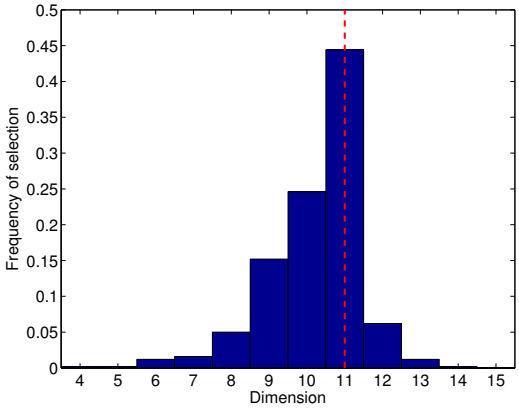
Figure 2a shows the expected distance between the true segmentation $\tau^\star$ and the segmentations $(\widehat{\tau}(D))_{1 \leq D \leq D_{\max}}$ produced at Step 1 of Algorithm 1. As expected, the distance is clearly minimal at $D = D^\star$, for both Hausdorff and Frobenius distances. Note that for each individual sample, $d(\widehat{\tau}(D), \tau^\star)$ behaves exactly as the expectation shown on Figure 2a, up to minor fluctuations. Moreover, the minimal value of the distance is small enough to suggest that $\widehat{\tau}(D^\star)$ is indeed close to $\tau^\star$. For instance, $\mathbb{E}[d_F(\widehat{\tau}(D^\star), \tau^\star)] \approx 1.71$, with a 95%



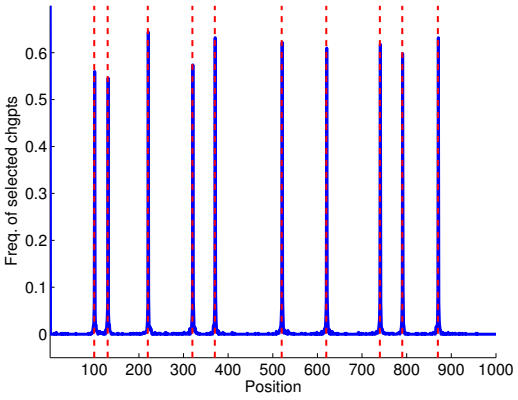
(a) Average distance (d_F or d_H) between $\widehat{\tau}(D)$ and $\tau^\star$, as a function of D.



(b) Average risk $\mathcal{R}\big(\widehat{\mu}_{\widehat{\tau}(D)}\big)$, empirical risk $\widehat{\mathcal{R}}_n\big(\widehat{\mu}_{\widehat{\tau}(D)}\big)$ and penalized criterion as a function of D.



(c) Distribution of $\widehat{D}$.



(d) Probability, for each instant $i \in \{1, \ldots, n\}$, that $\widehat{\tau} = \widehat{\tau}(\widehat{D})$ puts a change-point at i.

Figure 2: Scenario 1: $\mathcal{X} = \mathbb{R}$, variable (mean, variance). Performance of Algorithm 1 with kernel k^{G}. The value $D^\star$ and the localization of the true change-points in $\tau^\star$ are materialized by vertical red lines.

error bar smaller than 0.11. The closeness between $\widehat{\tau}(D^\star)$ and $\tau^\star$ when $k = k^{\mathrm{G}}$ can also be visualized on Figure 7c in the supplementary material.

As a comparison, when $k = k^{\mathrm{lin}}$ in the same setting, $\widehat{\tau}(D^\star)$ is much further from $\tau^\star$ since $\mathbb{E}[d_F(\widehat{\tau}(D^\star), \tau^\star)] \approx 10.39 \pm 0.24$, and a permutation test shows that the difference is significant, with a p-value smaller than 10^{-13}. See also Figures 5 and 7a in the supplementary material.

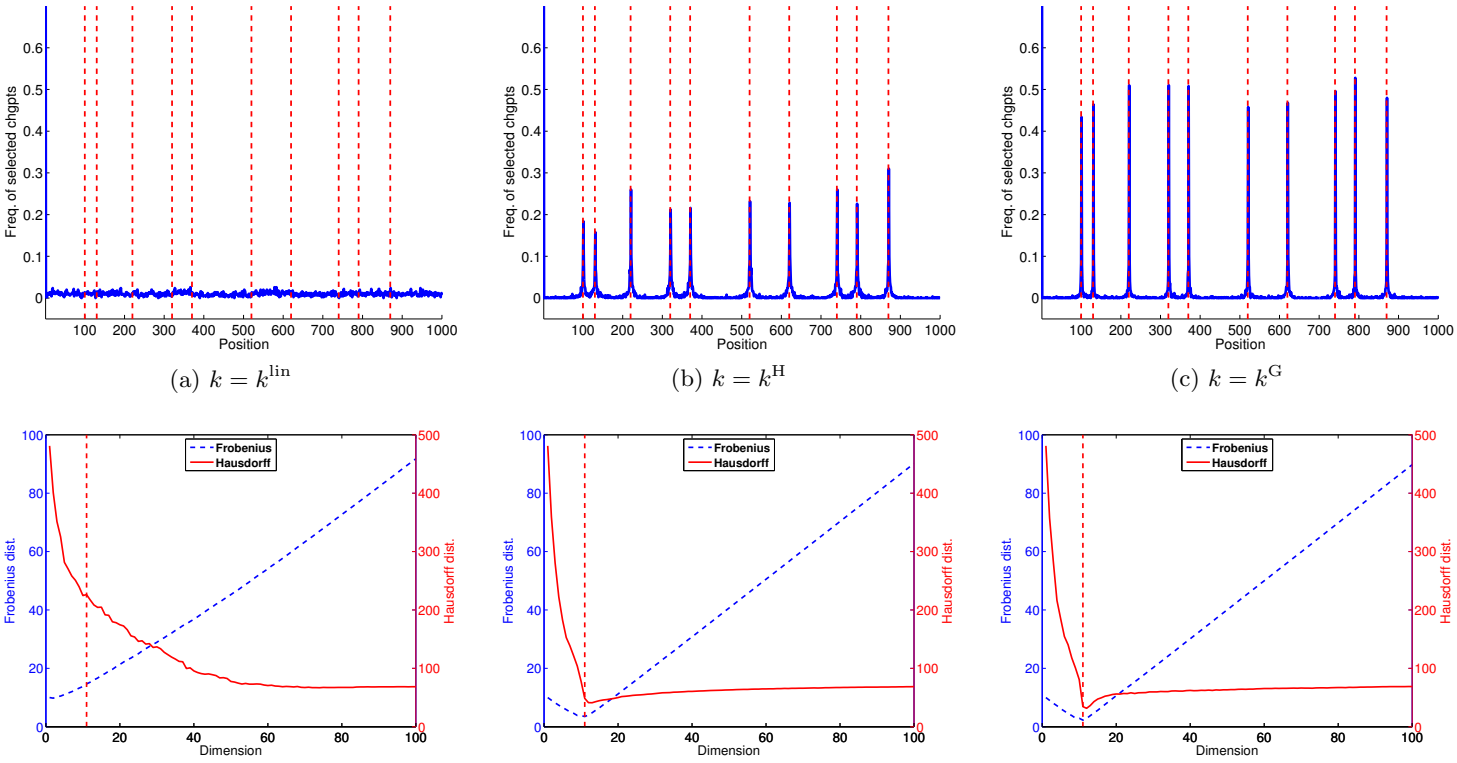
Step 2 of Algorithm 1 is illustrated by Figures 2b and 2c. The expectation of the penalized criterion is minimal at $D = D^\star$ (as well as for the risk of $\widehat{\mu}_{\widehat{\tau}(D)}$), and takes significantly larger values when $D \neq D^\star$ (Figure 2b). As a result, Algorithm 1 often selects a number of change-points $\widehat{D} - 1$ close to its true value $D^\star - 1$ (Figure 2c). Overall, this suggests that the model selection procedure used at Step 2 of Algorithm 1 works fairly well.

The overall performance of Algorithm 1 as a change-point detection procedure is illustrated by Figure 2d. Each true change-point has a probability larger than 0.5 to be recovered *exactly* by $\widehat{\tau}$. If one groups the positions i by blocks of six elements $\{6j, 6j+1, \ldots, 6j+5\}$, $j \geq 1$, the frequency of detection of a change-point by $\widehat{\tau}$ in each block containing a true change-point is between 79 and 89%. Importantly, such figures are obtained without overestimating much the number of change-points, according to Figure 2c. Figures 8a and 8b in the supplementary material show that more standard change-point detection algorithms—that is, Algorithm 1 with $k = k^{\mathrm{lin}}$ or k^{H}—have a slightly worse performance.

Comparison of three kernels in Scenario 2. Scenario 2 proposes a challenging change-point problem with real-valued data: the distribution of the X_i changes while the mean *and* the variance remain constant. The performance of Algorithm 1 with three kernels—k^{lin}, k^{H} and k^{G}—is shown on Figure 3. The linear kernel k^{lin} corresponds to the classical least-squares change-point algorithm (Lebarbier, 2005), which is designed to detect changes in the mean, hence it should fail in Scenario 2. Algorithm 1 with the Hermite kernel k^{H} is a natural "hand-made" extension of this classical approach, since it corresponds to applying the least-squares change-point algorithm to the feature vectors $(H_{j,h}(X_i))_{1 \leq j \leq 5}$. By construction, it should be able to detect changes in the first five moments on the X_i. On the contrary, taking $k = k^{\mathrm{G}}$ the Gaussian kernel fully relies on the versatility of Algorithm 1, which makes possible to consider (virtually) infinite-dimensional feature vectors $k^{\mathrm{G}}(X_i, \cdot)$. Since k^{G} is characteristic, it should be able to detect any change in the distribution of the X_i.

In order to compare these three kernels within Algorithm 1, let us first assume that the number of change-points is known, hence we can estimate $\tau^\star$ with $\widehat{\tau}(D^\star)$. Then, Figures 3a, 3b and 3c show that k^{lin}, k^{H} and k^{G} behave as expected: k^{lin} seems to put the change-points of $\widehat{\tau}(D^\star)$ uniformly at random over $\{1, \ldots, n\}$, while k^{H} and k^{G} are able to localize the true change-points with a rather large probability of success. The Gaussian kernel here shows a significantly better detection power, compared to k^{H}: the frequency of exact detection of the true change-points is between 43 and 53% with k^{G}, and between 16 and 31% with k^{H}. The same holds when considering blocks of size 6: k^{G} then detects 70 to 82% of the change-points, while k^{H} only detects 43 to 67% of them.

Figures 3d, 3e and 3f show that a similar comparison between k^{lin}, k^{H}, and k^{G} holds over the whole set of segmentations $(\widehat{\tau}(D))_{1 \leq D \leq D_{\max}}$ provided by Step 1 of Algorithm 1. With the linear kernel (Figure 3d), the Frobenius distance between $\widehat{\tau}(D)$ and $\tau^\star$ is almost minimal



(a) $k = k^{\mathrm{lin}}$ (b) $k = k^{\mathrm{H}}$ (c) $k = k^{\mathrm{G}}$

(d) $k = k^{\mathrm{lin}}$ (e) $k = k^{\mathrm{H}}$ (f) $k = k^{\mathrm{G}}$

Figure 3: Scenario 2: $\mathcal{X} = \mathbb{R}$, constant mean and variance. Performance of Algorithm 1 with three different kernels k. The value $D^\star$ and the localization of the true change-points in $\tau^\star$ are materialized by vertical red lines. **Top:** Probability, for each instant $i \in \{1, \ldots, n\}$, that $\widehat{\tau}(D^\star)$ puts a change-point at i. **Bottom:** Average distance (d_F or d_H) between $\widehat{\tau}(D)$ and $\tau^\star$, as a function of D.

for $D = 1$, which suggests that $\widehat{\tau}(D)$ is not far from random guessing for all D. The shape of the Hausdorff distance—first decreasing fastly, then almost constant—also supports this interpretation: A small number of purely random guesses do lead to a fast decrease of d_H; and for large dimensions, adding a new random guess does not move away $\widehat{\tau}(D)$ from $\tau^\star$ if $\widehat{\tau}(D)$ already contains all the worst possible candidate change-points (which are the furthest from the true change-points). The Hermite kernel does much better according to Figure 3e: the distance from $\widehat{\tau}(D)$ to $\tau^\star$ is minimal for D close to $D^\star$, and the minimal expected distance, $\inf_D \mathbb{E}[d_F(\widehat{\tau}(D), \tau^\star)] \approx 3.49 \pm 0.15$ (with confidence 95%), is much smaller than when $k = k^{\mathrm{lin}}$ (in which case $\inf_D \mathbb{E}[d_F(\widehat{\tau}(D), \tau^\star)] \approx 14.65 \pm 0.13$); this difference is significant (a permutation test yields a p-value smaller than 10^{-15}). Nevertheless, we obtain even better performance for $(\widehat{\tau}(D))_{1 \leq D \leq D_{\max}}$ with $k = k^{\mathrm{G}}$, for which the minimal distance to $\tau^\star$ is obtained at $D = D^\star$, with a minimal expected value equal to 2.27 ± 0.16; the difference between minimal expected distances with k^{H} and k^{G} is significant (p-value smaller than 10^{-15}).

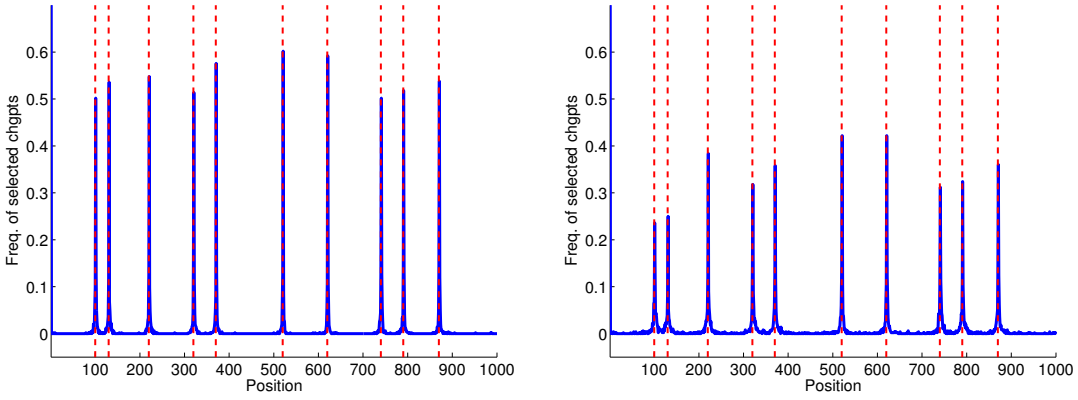
When $D = \widehat{D}$ is chosen by Algorithm 1, k^{G} still leads to the best performance in terms of recovering the exact change-points compared to k^{lin} and k^{H}, as illustrated by Figures 9a, 9b and 9c in the supplementary material. The (surprising) better behavior of k^{lin} compared to k^{H} from Figures 9a and 9b is an artifact only resulting from the choice of $\widehat{D}$ larger than $D^\star$ with k^{lin} (see Figure 10a). This overly large $\widehat{D}$ then induces many false change-points, which we would like to avoid.

Overall, the best performance in Scenario 2 is clearly obtained with k^{G}, while k^{lin} completely fails and k^{H} yields a decent but suboptimal procedure.

We can notice that other settings can lead to different behaviours. For instance, in Scenario 1, according to Figure 8a in the supplementary material, k^{lin} can detect fairly well the true change-points—as expected since the mean (almost) always changes in this scenario, see Table 1 in the supplementary material—, but this is at the price of a strong overestimation of the number of change-points (Figure 6a). In the same setting, k^{H} provides fairly good results (Figure 8b), while k^{G} remains the best choice (Figure 2d).

Since k^{G} is a characteristic kernel, these results suggest that Algorithm 1 with a characteristic kernel k might be more versatile than classical least-squares change-point algorithms and their extensions. A more detailed simulation experiment would nevertheless be needed to confirm this hypothesis. We also refer to Section 7.2 for a discussion on the choice of k for a given change-point problem.

Structured data. Figure 4 illustrates the performance of Algorithm 1 on some histogram-valued data (Scenario 3). Since a d-dimensional histogram is also an element of $\mathbb{R}^d$, we can analyze such data either with a kernel taking into account the histogram structure (such as k_{χ^2}) or with a usual kernel on $\mathbb{R}^d$ (such as k^{lin} or k^{G}; here, we consider k^{G}, which seems more reliable according to our experiments in Scenarios 1 and 2). Assuming that the number of change-points is known, taking $k = k_{\chi^2}$ yields quite good results according to Figure 4a, at least in comparison with $k = k^{\mathrm{G}}$ (Figure 4b). Similar results hold with a fully data-driven number of change-points, as shown by Figures 11a and 11b in the supplementary material. Hence, choosing a kernel such as k_{χ^2}, which takes into account the histogram structure of the X_i, can improve much the change-point detection performance, compared to taking a



(a) $k = k_{\chi^2}$ (b) $k = k^{\mathrm{G}}$

Figure 4: Scenario 3: histogram-valued data. Performance of Algorithm 1 with two different kernels k. Probability, for each instant $i \in \{1, \dots, n\}$, that $\widehat{\tau}(D^\star)$ puts a change-point at i. Vertical red lines show the true change-points locations.

kernel such as k^{G}, which ignores the structure of the X_i.

Let us emphasize that Scenario 3 is quite challenging— changes are hard to distinguish on Figure 1c—, which has been chosen on purpose. Preliminary experiments have been done with larger values of c_3—which makes the change-point problem easier, see Section 6.1—, leading to an almost perfect localization of all change-points by Algorithm 1 with $k = k_{\chi^2}$.

7 Conclusion

This paper proposes a kernel change-point algorithm (Algorithm 1), based upon a penalization procedure generalizing the one of Comte and Rozenholc (2004) and Lebarbier (2005) to RKHS-valued data. Such an extension significantly broadens the range of possible applications of the algorithm, since it can deal with complex or structured data, and it can detect changes in the full distribution of the data—not only the mean or the variance. The new theoretical tools developed in the paper—mostly, a concentration inequality for some function of RKHS-valued random variables (Proposition 1)—could be useful in other settings, such as clustering in reproducing kernel Hilbert spaces or functional data analysis. Let us now end the paper with three open questions about Algorithm 1.

7.1 Identification of the change-point locations

A natural question for a change-point algorithm is its consistency for estimating the true change-point locations $\tau^\star$. More precisely, let us assume that some $\tau^\star \in \mathcal{T}_n$ exists such that

$$P_{X_{\tau^\star_{\ell-1}+1}} = \cdots = P_{X_{\tau^\star_\ell}} \qquad \text{for } 1 \leq \ell \leq D_{\tau^\star}\,, \qquad P_{X_{\tau^\star_\ell}} \neq P_{X_{\tau^\star_\ell+1}} \qquad \text{for } 1 \leq \ell \leq D_{\tau^\star} - 1$$

and $D_{\tau^\star}$ is fixed as n tends to infinity (even if $\tau^\star$ necessarily depends on n). The goal is to prove that $d(\widehat{\tau}, \tau^\star)$ tends to zero almost surely as n tends to infinity, where d is some

distance on $\mathcal{T}_n$, for instance $n^{-1}d_F$ or $n^{-1}d_H$ as defined in Section 6.3. Many papers prove such a consistency result for other change-point algorithms in various settings (for instance, Yao, 1988; Lavielle and Moulines, 2000; Frick et al., 2014; Matteson and James, 2014). Answering this question for Algorithm 1 is beyond the scope of the paper. It will be proved in a forthcoming paper (Garreau and Arlot, 2016) that Algorithm 1 is indeed consistent under mild assumptions.

7.2 Choosing the kernel k

A major practical and theoretical question about Algorithm 1 is the choice of the kernel k. Fully answering this question is beyond the scope of the paper, but we can already provide a few guidelines, based upon the theoretical and experimental results that we already have, and review some previous works tackling a related question.

First, simulation experiments in Section 6 show that the performance can strongly vary with k. They suggest that using a characteristic kernel—such as the Gaussian kernel k^{G}—yields a more versatile procedure when the goal is to detect changes in the full distribution of the data. Nevertheless, for a given change-point problem, all characteristic kernels certainly are not equivalent. For instance, unshown experimental results suggest that k_h^{G} with a clearly bad choice of the bandwidth h—say, smaller than 10^{-4} or larger than 10^4 in settings similar to Scenario 1—leads to a poor performance of Algorithm 1, despite the fact that k_h^{G} is characteristic for any $h > 0$.

Furthermore, for a given setting, a non characteristic kernel can be a good choice: when the goal is to detect changes in the mean of $X_i \in \mathbb{R}^d$, k^{lin} is known to work very well (Lebarbier, 2005).

Second, our theoretical interpretation of Algorithm 1 in Section 4.2 suggests how the performance of Algorithm 1 depends on k, hence on which basis k should be chosen. Indeed, Algorithm 1 focuses on changes in the mean $\mu_1^\star, \dots, \mu_n^\star$ of the time series $Y_1, \dots, Y_n \in \mathcal{H}$. A change between P_{X_i} and $P_{X_{i+1}}$ should be detected more easily when

$$\|\mu_{i+1}^\star - \mu_i^\star\|_{\mathcal{H}}^2 = \mathbb{E}\big[k(X_{i+1}, X_{i+1})\big] - 2\mathbb{E}\big[k(X_{i+1}, X_i)\big] + \mathbb{E}\big[k(X_i, X_i)\big]$$

is larger, compared to the "noise level" $\max\{v_i, v_{i+1}\}$. When $P_{X_i} \neq P_{X_{i+1}}$, we know that $\|\mu_{i+1}^\star - \mu_i^\star\|_{\mathcal{H}}$ is positive for any characteristic kernel k, while it might be equal to zero when k is not characteristic. But the fact that k is characteristic or not is not sufficient to guess whether k will work well or not, according to the above heuristic.

The problem of choosing a kernel has been considered for many different tasks in the machine learning literature. Let us only mention here some references that are tackling this question in a framework close to change-point detection: choosing the best kernel for a two-sample or an homogeneity test.

For choosing the bandwidth h of a Gaussian kernel, a classical heuristic is to take h equal to some median of $(\|X_i - X_j\|_{\mathcal{H}})_{i<j}$ (see Gretton et al., 2012a, Section 8, and references therein). This idea can be used for change-point detection with Algorithm 1, as done in a preliminary version of the present paper (Arlot et al., 2012, Section 6.2).

A procedure for choosing the best convex combination of a finite number of kernels has been proposed by Gretton et al. (2012b), with the goal of building a powerful two-sample

test. Another idea for combining several kernels, for instance the family $\{k_h^{\mathrm{G}} : h > 0\}$, has been studied by Sriperumbudur et al. (2009) for homogeneity and independence tests. Roughly, the idea is to replace the MMD test statistics—which depends on a kernel k—by its supremum over the considered family of kernels. Nevertheless, the extension of these two ideas to change-point detection with Algorithm 1 does not seem straightforward.

7.3 Heteroscedasticity of data in $\mathcal{H}$

A possible drawback of Algorithm 1 is that it does not take into account the fact that the variance v_i of $Y_i = \Phi(X_i)$ can change with i: in general, the Y_i are heteroscedastic. In the case of real-valued data and the linear kernel k^{lin}, Arlot and Celisse (2011) have shown that heteroscedastic data can make Algorithm 1 fail, and that this failure cannot be fixed by changing the penalty used at Step 2: all the segmentations $\widehat{\tau}(D)$ produced at Step 1 can be wrong.

We conjecture that, for the Gaussian kernel k_h^{G} at least, when the bandwidth h is well chosen, the variances of the Y_i stay within a reasonably small range of values for most non-degenerate distributions. Indeed, according to Eq. (8),

$$v_i = 1 - \mathbb{E}\left[\exp\left(\frac{-\|X_i - X_i'\|_{\mathcal{H}}^2}{2h^2}\right)\right] \in [0,1]$$

where X_i' is an independent copy of X_i. If X_i is not deterministic and if h is smaller than the typical order of magnitude of $\|X_i - X_i'\|_{\mathcal{H}}$, then, v_i cannot be much smaller than its maximal value 1. The median heuristic and our simulation experiments suggest that "good" values of h for change-point detection are small enough, but this remains to be proved.

When heteroscedasticity is a problem for Algorithm 1, which probably occurs for some kernels beyond k^{lin}, we can think of combining Algorithm 1 with the ideas of Arlot and Celisse (2011), that is, replacing the empirical risk and the penalized criterion in Steps 1 and 2 of Algorithm 1 by cross-validation estimators of the risk $\mathcal{R}(\widehat{\mu}_\tau)$.

Acknowledgments

The authors thank Damien Garreau for some discussions that lead to an improvement of the theoretical results (Proposition 1 and Theorem 2, which were stated with the additional assumption that $\min_i v_i \geq cM^2 > 0$ in a previous version of this paper (Arlot et al., 2012)). This work was mostly done while Sylvain Arlot was financed by CNRS and member of the Sierra team in the Departement d'Informatique de l'Ecole normale superieure (CNRS/ENS/INRIA UMR 8548), 45 rue d'Ulm, F-75230 Paris Cedex 05, France, and Zaid Harchaoui was a member of the LEAR team of Inria. Sylvain Arlot and Alain Celisse were also supported by Institut des Hautes Études Scientifiques (IHES, Le Bois-Marie, 35, route de Chartres, 91440 Bures-Sur-Yvette, France) at the end of the writing of this paper. Sylvain Arlot is also member of the Select project-team of Inria Saclay.
The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-09-JCJC-0027-01 (Detect project) and ANR-14-CE23-0003-01 (Macaron project), the GARGANTUA project funded by the Mastodons program of

CNRS, the LabEx Persyval-Lab (ANR-11-LABX-0025), the BeFast project funded by the PEPS Fascido program of CNRS, and the Moore-Sloan Data Science Environment at NYU.

A Additional proofs

A.1 Proofs of Section 4.1

A.1.1 Proof of Eq. (5)

Let $f \in F_\tau$ and $g \in \mathcal{H}^n$. For any $\ell \in [\![1, D_\tau]\!]$, we define $I_\ell^\tau := [\![\tau_{\ell-1}+1, \tau_\ell]\!]$ the ℓ-th interval of τ, $f_{I_\ell^\tau}$ the common value of $(f_i)_{i \in I_\ell^\tau}$ and

$$\overline{g}_{I_\ell^\tau} := \frac{1}{\mathrm{Card}(I_\ell^\tau)} \sum_{i \in I_\ell^\tau} g_i = \frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{i \in I_\ell^\tau} g_i \ . \tag{29}$$

Then,

$$\begin{aligned}
\|f-g\|^2 &= \sum_{\ell=1}^{D_\tau} \sum_{i \in I_\ell^\tau} \left[\left\| f_{I_\ell^\tau} - \overline{g}_{I_\ell^\tau} \right\|_{\mathcal{H}}^2 + \left\| g_i - \overline{g}_{I_\ell^\tau} \right\|_{\mathcal{H}}^2 + 2 \left\langle f_{I_\ell^\tau} - \overline{g}_{I_\ell^\tau}, \overline{g}_{I_\ell^\tau} - g_i \right\rangle_{\mathcal{H}} \right] \\
&= \sum_{\ell=1}^{D_\tau} \left[(\tau_\ell - \tau_{\ell-1}) \left\| f_{I_\ell^\tau} - \overline{g}_{I_\ell^\tau} \right\|_{\mathcal{H}}^2 \right] + \sum_{\ell=1}^{D_\tau} \sum_{i \in I_\ell^\tau} \left\| g_i - \overline{g}_{I_\ell^\tau} \right\|_{\mathcal{H}}^2 \ .
\end{aligned}$$

since $\sum_{i \in I_\ell^\tau} (\overline{g}_{I_\ell^\tau} - g_i) = 0$. So, $\|f-g\|^2$ is minimal over $f \in F_\tau$ if and only if $f_{I_\ell^\tau} = \overline{g}_{I_\ell^\tau}$ for every $\ell \in [\![1, D_\tau]\!]$. ■

A.1.2 Proof of Eq. (6)

We use the notations introduced in the proof of Eq. (5). Then,

$$\|Y - \widehat{\mu}_\tau\|^2 = \sum_{\ell=1}^{D_\tau} \sum_{i \in I_\ell^\tau} \|Y_i - \overline{Y}_{I_\ell^\tau}\|_{\mathcal{H}}^2 = \sum_{\ell=1}^{D_\tau} \sum_{i \in I_\ell^\tau} \left(\|Y_i\|_{\mathcal{H}}^2 - \|\overline{Y}_{I_\ell^\tau}\|_{\mathcal{H}}^2 \right)$$

where we used Eq. (5) for the first equality, and that

$$\sum_{i \in I_\ell^\tau} \left\langle Y_i, \overline{Y}_{I_\ell^\tau} \right\rangle_{\mathcal{H}} = \mathrm{Card}(I_\ell^\tau) \left\| \overline{Y}_{I_\ell^\tau} \right\|_{\mathcal{H}}^2$$

for the second equality. Therefore,

$$\begin{aligned}
\|Y - \widehat{\mu}_\tau\|^2 &= \sum_{i=1}^n \|Y_i\|_{\mathcal{H}}^2 - \sum_{\ell=1}^{D_\tau} \frac{1}{\tau_\ell - \tau_{\ell-1}} \left\| \sum_{i \in I_\ell^\tau} Y_i \right\|_{\mathcal{H}}^2 \\
&= \sum_{i=1}^n \|Y_i\|_{\mathcal{H}}^2 - \sum_{\ell=1}^{D_\tau} \frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{i,j \in I_\ell^\tau} \langle Y_i, Y_j \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^n k(X_i, X_i) - \sum_{\ell=1}^{D_\tau} \frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{i,j \in I_\ell^\tau} k(X_i, X_j) \ ,
\end{aligned}$$

which proves Eq. (6). ■

A.2 Concentration of the linear term: proof of Proposition 3

Let us define $\mu_\tau^\star = \Pi_\tau \mu^\star$ and

$$S_\tau = \langle \mu^\star - \mu_\tau^\star, \varepsilon \rangle = \sum_{i=1}^n Z_i \quad \text{with} \quad Z_i = \left\langle (\mu^\star - \mu_\tau^\star)_i, \varepsilon_i \right\rangle_{\mathcal{H}} \ ,$$

The Z_is are independent and centered, so Eq. (31)–(32) in Lemma 5 below (which requires assumption (**Db**)) show that the conditions of Bernstein's inequality are satisfied (see Proposition 6). Therefore for every $x \geq 0$, with probability at least $1 - 2\mathrm{e}^{-x}$,

$$\begin{aligned}\left| \sum_{i=1}^n Z_i \right| &\leq \sqrt{2 v_{\max} \left\| \mu^\star - \mu_\tau^\star \right\|^2 x} + \frac{4M^2 x}{3} \\ &\leq \theta \left\| \mu^\star - \mu_\tau^\star \right\|^2 + \left(\frac{v_{\max}}{2\theta} + \frac{4M^2}{3} \right) x\end{aligned}$$

for every $\theta > 0$, using $2ab \leq \theta a^2 + \theta^{-1} b^2$. ■

A key argument in the proof is the following lemma.

Lemma 5 *For every $m \in \mathcal{M}_n$, if* (**Db**) *holds true, the following holds true with probability one*:

$$\forall i \in \{1, \ldots, n\} \ , \quad \|\mu_i^\star\|_{\mathcal{H}} \leq M \, , \qquad \|\varepsilon_i\|_{\mathcal{H}} \leq 2M \tag{30}$$

$$\textit{and} \quad \|(\mu^\star - \mu_\tau^\star)_i\|_{\mathcal{H}} \leq 2M \quad \textit{so that} \quad |Z_i| \leq 4M^2 \ . \tag{31}$$

$$\textit{In addition,} \quad \sum_{i=1}^n \mathrm{Var}\left(Z_i\right) \leq v_{\max} \left\| \mu^\star - \mu_\tau^\star \right\|^2 \ . \tag{32}$$

Proof [of Lemma 5] First, remark that for every i,

$$v_i = \mathbb{E}\big[\|\varepsilon_i\|^2\big] = \mathbb{E}\big[k(X_i, X_i)\big] - \|\mu_i^\star\|_{\mathcal{H}}^2 \geq 0 \ ,$$

so that with (**Db**),

$$\|\mu_i^\star\|_{\mathcal{H}}^2 \leq \mathbb{E}\big[k(X_i, X_i)\big] \leq M^2 \ ,$$

which proves the first bound in Eq. (30). As a consequence, by the triangular inequality,

$$\|\varepsilon_i\|_{\mathcal{H}} \leq \|Y_i\|_{\mathcal{H}} + \|\mu_i^\star\|_{\mathcal{H}} \leq 2M \ ,$$

that is, the second inequality in Eq. (30) holds true.

Let us now define for every $i \in \{1, \ldots, n\}$, the integer $K(i) \in \{1, \ldots, D_\tau\}$ such that $I_{K(i)}^\tau = [\![\tau_{K(i)-1} + 1, \tau_{K(i)}]\!]$ is the unique interval of the segmentation τ such that $i \in I_{K(i)}^\tau$. Then,

$$(\mu^\star - \mu_\tau^\star)_i = \frac{1}{\tau_{K(i)} - \tau_{K(i)-1}} \sum_{j \in I_{K(i)}^\tau} (\mu_i^\star - \mu_j^\star),$$

so that the triangular inequality and Eq. (30) imply

$$\|(\mu^\star - \mu^\star_\tau)_i\|_{\mathcal{H}} \leq \sup_{j \in I^\tau_{K(i)}} \left\|\mu^\star_i - \mu^\star_j\right\|_{\mathcal{H}} \leq \sup_{1 \leq j,k \leq n} \left\|\mu^\star_k - \mu^\star_j\right\|_{\mathcal{H}} \leq 2 \sup_{1 \leq j \leq n} \left\|\mu^\star_j\right\|_{\mathcal{H}} \leq 2M \ ,$$

that is, the first part of Eq. (31) holds true. The second part of Eq. (31) directly follows from Cauchy-Schwarz's inequality. For proving Eq. (32), we remark that

$$\begin{aligned}
\mathbb{E}\left[Z_i^2\right] &= \mathbb{E}\Big[\left\langle (\mu^\star - \mu^\star_\tau)_i,\, \varepsilon_i \right\rangle^2_{\mathcal{H}}\Big] \\
&\leq \|(\mu^\star - \mu^\star_\tau)_i\|^2_{\mathcal{H}}\, \mathbb{E}\left[\|\varepsilon_i\|^2_{\mathcal{H}}\right] \qquad \text{by Cauchy-Schwarz's inequality} \\
&= \|(\mu^\star - \mu^\star_\tau)_i\|^2_{\mathcal{H}}\, v_i \leq \|(\mu^\star - \mu^\star_\tau)_i\|^2_{\mathcal{H}}\, v_{\max} \ ,
\end{aligned}$$

$$\text{so that} \quad \sum_{i=1}^n \mathrm{Var}\left(Z_i\right) \leq v_{\max} \left\|\mu^\star - \mu^\star_\tau\right\|^2 \ .$$

■

References

Nathalie Akakpo. Estimating a discrete distribution via histogram selection. *ESAIM: Probability and Statistics*, 15:1–29, 2011.

M. Z. Alaya, S. Gaïffas, and A. Guilloux. Learning the intensity of time events with change-points. *IEEE Transactions on Information Theory*, 61(9):5148–5171, 2015.

Elena Andreou and Eric Ghysels. Detecting multiple breaks in financial market volatility dynamics. *Journal of Applied Econometrics*, 17(5):579–600, 2002.

C. Ané. Detecting phylogenetic breakpoints and discordance from genome-wide alignments for species tree reconstruction. *Genome Biology and Evolution*, 3:246–258, 2011.

S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning*, 10:245–279, 2009.

Sylvain Arlot. *Contributions to statistical learning theory: estimator selection and change-point detection*. Habilitation à diriger des recherches, University Paris Diderot, December 2014. Available at `http://tel.archives-ouvertes.fr/tel-01094989`.

Sylvain Arlot and Alain Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Stat. Comput.*, 21(4):613–632, 2011.

Sylvain Arlot, Alain Celisse, and Zaïd Harchaoui. Kernel change-point detection, February 2012. arXiv:1202.3878v1.

Ivan E Auger and Charles E Lawrence. Algorithms for the optimal identification of segment neighborhoods. *Bulletin of mathematical biology*, 51(1):39–54, 1989.

Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with an unknown variance. *The Annals of Statistics*, pages 630–672, 2009.

Jean-Marc Bardet and Imen Kammoun. Detecting abrupt changes of the long-range dependence or the self-similarity of a gaussian process. *Comptes Rendus Mathématique*, 346 (13):789–794, 2008.

Jean-Marc Bardet, William Chakry Kengne, and Olivier Wintenberger. Multiple breaks detection in general causal time series using penalized quasi-likelihood. *Electron. J. Stat.*, 6:435–477 (electronic), 2012.

Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

I. Berkes, R. Gabrys, L. Horváth, and P. Kokoszka. Detecting changes in the mean of functional observations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(5):927–946, 2009.

Karine Bertin, Xavier Collilieux, Emilie Lebarbier, and Cristian Meza. Segmentation of multiple series using a lasso strategy, 2014. arXiv:1406.6627.

Gérard Biau, Kevin Bleakley, and David Mason. Long signal change-point detection, 2015. arXiv:1504.01702.

Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.

K. Bleakley and J.-Ph. Vert. The group fused lasso for multiple change-point detection, 2011. arXiv:1106.4199.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

L. Boysen, A. Kempe, V. Liebscher, A. Munk, and O. Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Stat.*, 37(1):157–183, 2009.

Boris E. Brodsky and Boris S. Darkhovsky. *Nonparametric methods in change-point problems*, volume 243 of *Mathematics and its Applications*. Kluwer Academic Publishers Group, Dordrecht, 1993.

Edward Carlstein, Hans-Georg Müller, and David Siegmund, editors. *Change-point problems*. IMS Lect. Notes, 1994.

A. Celisse, G. Marot, M. Pierre-Jean, and G. Rigaill. New efficient algorithms for kernel change-point detection. Private communication, 2016.

Jinyuan Chang, Bin Guo, and Qiwei Yao. Segmenting multiple time series by contemporaneous linear transformation, 2014. arXiv:1410.2323.

Hao Chen and Nancy Zhang. Graph-based change-point detection. *Ann. Statist.*, 43(1): 139–176, 2015.

Kacper P Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. In *Advances in Neural Information Processing Systems 27*, pages 3608–3616. Curran Associates, Inc., 2014.

Alice Cleynen and Émilie Lebarbier. Model selection for the segmentation of multiparameter exponential family distributions, 2014a. arXiv:1412.6697.

Alice Cleynen and Émilie Lebarbier. Segmentation of the poisson and negative binomial rate models: a penalized estimator. *ESAIM: Probability and Statistics*, 18:750–769, 2014b.

Xavier Collilieux, Émilie Lebarbier, and Stéphane Robin. A factor model approach for the joint segmentation with between-series correlation, 2015. arXiv:1505.05660.

Fabienne Comte and Yves Rozenholc. A new algorithm for fixed design regression and denoising. *Ann. Inst. Statist. Math.*, 56(3):449–473, 2004.

Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

R. Curtis, J. Xiang, A. Parikh, P. Kinnaird, and E. P. Xing. Enabling dynamic network analysis through visualization in TVNViewer. *BMC Bioinformatics*, 13(204), 2012.

Marco Cuturi, Kenji Fukumizu, and Jean-Philippe Vert. Semigroup kernels on measures. *J. Mach. Learn. Res.*, 6:1169–1198, 2005.

Aymeric Dieuleveut and Francis Bach. Non-parametric stochastic approximation with large step sizes, 2014. arXiv:1408.0361v1.

Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis*. Springer Series in Statistics. Springer, New York, 2006.

Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

Magalie Fromont, Béatrice Laurent, Matthieu Lerasle, and Patricia Reynaud-Bouret. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *JMLR W& CP (COLT 2012)*, volume 23, pages 23.1–23.23, 2012.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

Piotr Fryzlewicz and Suhasini Subba Rao. Multiple-change-point detection for auto-regressive conditional heteroscedastic processes. *Journal of the Royal Statistical Society: series B (statistical methodology)*, 76(5):903–924, 2014.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5:73–99, 2004a.

Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Kernel dimensionality reduction for supervised learning. In *Advances in Neural Information Processing Systems 16*, pages 81–88. MIT Press, 2004b.

Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496. Curran Associates, Inc., 2008.

Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. Private communication, 2016.

Vladimir J. Geneus, Jordan Cuevas, Eric Chicken, and J Pignatiello. A changepoint detection method for profile variance, 2014. arXiv:1408.7000.

Irene Gijbels, Peter Hall, and Aloïs Kneip. On the estimation of jump points in smooth curves. *Annals of the Institute of Statistical Mathematics*, 51(2):231–251, 1999.

Evarist Giné and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2016.

A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012a.

A. Gretton, D. Sejdinovic, Heiko S., S. Balakrishnan, M. Pontil, Kenji F., and B. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems 25*, pages 1205–1213. Curran Associates, Inc., 2012b.

Zaïd Harchaoui and Olivier Cappé. Retrospective change-point estimation with kernels. In *IEEE Workshop on Statistical Signal Processing*, 2007.

Zaïd Harchaoui, Francis Bach, and Éric Moulines. Testing for Homogeneity with Kernel Fisher Discriminant Analysis, April 2008. Available at `http://hal.archives-ouvertes.fr/hal-00270806/`.

Toby Hocking, Guillem Rigaill, Jean-Philippe Vert, and Francis Bach. Learning sparse penalties for change-point detection using max margin interval regression. In *Proc. The 30th Intern. Conf. on Mach. Learn.*, pages 172–180, 2013.

Christian Houdré and Patricia Reynaud-Bouret. Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic inequalities and applications*, volume 56 of *Progr. Probab.*, pages 55–69. Birkhäuser, Basel, 2003.

Steven M. Kay. *Fundamentals of statistical signal processing: detection theory*. Prentice-Hall, Inc., 1993.

L. L. Knowles and L.S. Kubatko. *Estimating species trees: practical and theoretical aspects*. Hobroken, N. J.: Wiley-Blackwell, 2010.

Irena Koprinska and Sergio Carrato. Temporal video segmentation: A survey. *Signal processing: Image communication*, 16(5):477–500, 2001.

Alexander Korostelev and Olga Korosteleva. *Mathematical statistics. Asymptotic minimax theory.* Graduate Studies in Mathematics 119. American Mathematical Society (AMS), 2011.

G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. *Science*, 311:88–90, 2006.

Rémi Lajugie, Sylvain Arlot, and Francis Bach. Large-margin metric learning for constrained partitioning problems. In *International Conference on Machine Learning (ICML)*, volume 32, pages 297–305, 2014. See also arXiv:1303.1280.

Marc Lavielle and Eric Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21(1):33–59, 2000.

Émilie Lebarbier. *Quelques approches pour la détection de ruptures à horizon fini.* PhD thesis, Université Paris-Sud, July 2002.

Émilie Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proces.*, 85:717–736, 2005.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*, volume 23 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 1991.

Shuang Li, Yao Xie, Hanjun Dai, and Le Song. M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems 28*, pages 3348–3356. Curran Associates, Inc., 2015.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

David S Matteson and Nicholas A James. A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109 (505):334–345, 2014.

Ian McCulloh. *Detecting Changes in a Dynamic Social Network.* PhD thesis, Institute for Software Research, School of Computer Science, Carnegie Mellon University, 2009. CMU-ISR-09-104.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.

Youngser Park, Heng Wang, Tobias Nöbauer, Alipasha Vaziri, and Carey E Priebe. Anomaly detection on whole-brain functional imaging of neuronal activity using graph scan statistics. *Neuron*, 2(3,000):4–000, 2015.

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

F. Picard, E. Lebarbier, E. Budinska, and S. Robin. Joint segmentation of multivariate gaussian processes using mixed linear models. *Computational Statistics & Data Analysis*, 55(2):1160 – 1170, 2011.

I.F. Pinelis and A.I. Sakhanenko. Remarks on inequalities for large deviation probabilities. *Theory Probab. Appl.*, 30:143–148, 1986.

Lawrence R. Rabiner and Ronald W. Schäfer. Introduction to digital signal processing. *Foundations and Trends in Information Retrieval*, 1(1–2):1–194, 2007.

M. Sauvé. Histogram selection in non gaussian regression. *ESAIM: Probability and Statistics*, 13:70–86, 2009.

B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

B. Schölkopf, K. Tsuda, and J.-Ph. Vert. *Kernel Methods in Computational Biology*. MIT Press, 2004.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5):2263–2291, 2013.

Olimjon Sharipov, Johannes Tewes, and Martin Wendler. Sequential block bootstrap in a hilbert space with application to change point analysis, 2014. arXiv:1412.0446.

John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, 2004.

Nino Shervashidze. *Scalable graph kernels*. PhD thesis, Universität Tübingen, 2012. Available at `http://hdl.handle.net/10900/49731`.

Yong Sheng Soh and Venkat Chandrasekaran. High-dimensional change-point estimation: Combining filtering with convex optimization, 2014. arXiv:1412.3731.

B. Sriperumbudur, K. Fukumizu, and G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *J. Mach. Learn. Res.*, 12:2389–2410, 2011.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Gert R. G. Lanckriet, and Bernhard Schölkopf. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems*, volume 21. NIPS Foundation (http://books.nips.cc), 2009.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Information Science and Statistics. Springer, New York, 2008.

Michel Talagrand. New concentration inequalities in product spaces. *Invent. Math.*, 126 (3):505–563, 1996.

Alexander Tartakovsky, Igor Nikiforov, and Basseville Michèle. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*, volume 136 of *Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC, Boca Raton, FL, 2014.

Heng Wang, Minh Tang, Yu-Seop Park, and Carey E Priebe. Locality statistics for anomaly detection in time series of graphs. *Signal Processing, IEEE Transactions on*, 62(3):703–717, 2014.

Chung-Hsien Wu and Chia-Hsin Hsieh. Multiple change-point audio segmentation and classification using an MDL-based Gaussian model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(2):647–657, 2006.

Y. Yao. Estimating the number of change-points via Schwarz criterion. *Statistics and Probability Letters*, 6:181–189, 1988.

Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In *Advances in Neural Information Processing Systems 26*, pages 755–763. Curran Associates, Inc., 2013.

N. R. Zhang and D. O. Siegmund. Modified Bayes Information Crietrion with Application to the Analysis of Comparative Genomic Hybridization Data. *Biometrics*, 63:22–32, 2007.

C. Zou, G. Yin, L. Feng, and Z. Wang. Nonparametric maximum likelihood approach to multiple change-point problems. *The Annals of Statistics*, 42(3):970–1002, 2014.

B Supplementary material

B.1 Classical concentration inequalities

This section collects a few results that are used throughout the paper.

B.1.1 Bernstein's inequality

Proposition 6 (Bernstein's inequality, as stated by Massart 2007, Proposition 2.9) *Let $X_1, \dots, X_n$ be independent real-valued random variables. Assume that some positive constants v and c exist such that, for every $k \geq 2$*

$$\sum_{i=1}^{n} \mathbb{E}\Big[|X_i|^k\Big] \leq \frac{k!}{2} v c^{k-2} \ . \tag{33}$$

Then, for every $x > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n} \big(X_i - \mathbb{E}[X_i]\big) > \sqrt{2vx} + cx\right) \leq \mathrm{e}^{-x} \ .$$

In particular, if for every $i \in \{1, \dots, n\}$, $|X_i| \leq 3c$ almost surely, Eq. (33) holds true with $v = \sum_{i=1}^{n} \mathrm{Var}(X_i)$.

B.1.2 Pinelis-Sakhanenko's inequality

Proposition 7 (Pinelis and Sakhanenko (1986), Corollary 1) *Let $X_1, \dots, X_n$ be independent and identically distributed random variables with values in some Hilbert space $\mathcal{H}$. Assume the X_i are centered and that constants $\sigma^2, c > 0$ exist such that for every $p \geq 2$,*

$$\sum_{i=1}^{n} \mathbb{E}\left[\|X_i\|_{\mathcal{H}}^p\right] \leq \frac{p!}{2} \sigma^2 c^{p-2} \ ,$$

Then, for every $x > 0$,

$$\mathbb{P}\left[\left\|\sum_{i=1}^{n} X_i\right\|_{\mathcal{H}} > x\right] \leq 2 \exp\left[-\frac{x^2}{2\left(\sigma^2 + cx\right)}\right] \ .$$

B.1.3 Talagrand's inequality

The following proposition is a refined version of Talagrand's concentration inequality (Talagrand, 1996), as it is stated by Boucheron et al. (2013, Corollary 12.12).

Proposition 8 (Corollary 12.12 of Boucheron et al. (2013)) *Let $X_1, \dots, X_n$ be independent vector-valued random variables and let*

$$Z = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} X_{i,f} \, .$$

Assume that for all $i \in \{1, \dots, n\}$ and $f \in \mathcal{F}$, $\mathbb{E}[X_{i,f}] = 0$ and $|X_{i,f}| \leq 1$. Define

$$\sigma^2 = \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \mathbb{E}\left[X_{i,f}^2\right] \qquad \text{and} \qquad v = 2\mathbb{E}[Z] + \sigma^2 \,.$$

Then, for all $x \geq 0$,

$$\mathbb{P}\Big(Z \geq \mathbb{E}[Z] + \sqrt{2vx} + \frac{x}{3}\Big) \leq \mathrm{e}^{-x} \tag{34}$$

$$\mathbb{P}\Big(Z \leq \mathbb{E}[Z] - \sqrt{2vx} - \frac{x}{8}\Big) \leq \mathrm{e}^{-x} \,. \tag{35}$$

B.2 Proof of Proposition 4

The first step is to write $\|\Pi_\tau \varepsilon\|$ of the form of Z in Proposition 8 for some well-chosen $(X_{i,f})_{1 \leq i \leq n,\, f \in \mathcal{G}_\tau}$. With for every $1 \leq K \leq D_\tau$, $\overline{f}_K = 1/(\tau_K - \tau_{K-1}) \sum_{i=\tau_{K-1}+1}^{\tau_K} f_i$, it comes

$$\begin{aligned}
\|\Pi_\tau \varepsilon\| &= \sup_{f \in \mathcal{H}^n,\, \|f\| \leq 1} \big|\langle f,\, \Pi_\tau \varepsilon \rangle\big| \\
&= \sup_{f \in \mathcal{H}^n,\, \|f\| \leq 1} \big|\langle \Pi_\tau f,\, \varepsilon \rangle\big| \\
&= \sup_{f \in \mathcal{H}^n,\, \sum_{K=1}^{D_\tau} (\tau_K - \tau_{K-1}) \|\overline{f}_K\|^2 \leq 1} \left| \sum_{K=1}^{D_\tau} \sum_{i=\tau_{K-1}+1}^{\tau_K} \big\langle \overline{f}_K,\, \varepsilon_i \big\rangle_{\mathcal{H}} \right| \\
&= \sup_{f \in \mathcal{G}_\tau} \sum_{i=1}^{n} \overline{X}_{i,f}
\end{aligned}$$

where $\mathcal{G}_\tau$ is some countable dense subset of

$$\left\{ f \in \mathcal{H}^n,\, \sum_{K=1}^{D_\tau} (\tau_K - \tau_{K-1}) \left\|\overline{f}_K\right\|_{\mathcal{H}}^2 \leq 1 \right\}$$

(such a set $\mathcal{G}_\tau$ exists since $\mathcal{H}$ is separable), and for every $i \in \{1, \dots, n\}$ and $f \in \mathcal{G}_\tau$,

$$\overline{X}_{i,f} = \left\langle \overline{f}_{K(i)},\, \varepsilon_i \right\rangle_{\mathcal{H}}$$

where we recall that $K(i)$ is defined in the proof of Lemma 5.

Let us now check that the assumptions of Proposition 8 are satisfied: $(\overline{X}_{1,f})_{f \in \mathcal{G}_\tau}, \dots, (\overline{X}_{n,f})_{f \in \mathcal{G}_\tau}$ are independent since $\varepsilon_1, \dots, \varepsilon_n$ are assumed independent. For every $i \in \{1, \dots, n\}$ and $f \in \mathcal{G}_\tau$,

$$\mathbb{E}\left[\overline{X}_{i,f}\right] = \mathbb{E}\left[\left\langle \overline{f}_{K(i)},\, \varepsilon_i \right\rangle_{\mathcal{H}}\right] = 0$$

since $\overline{f}_{K(i)} \in \mathcal{H}$ is deterministic, and for every $f \in \mathcal{G}_\tau$,

$$\left|\overline{X}_{i,f}\right| = \left|\left\langle \overline{f}_{K(i)},\, \varepsilon_i \right\rangle_{\mathcal{H}}\right| \leq \left\|\overline{f}_{K(i)}\right\|_{\mathcal{H}} \|\varepsilon_i\|_{\mathcal{H}} \leq \frac{2M}{\sqrt{\tau_{K(i)} - \tau_{K(i)-1}}} \leq 2M$$

by Cauchy-Schwarz's inequality, assumption (**Db**) and Lemma 5. So, we can apply Proposition 8 to

$$Z = \frac{1}{2M} \|\Pi_\tau \varepsilon\| = \sup_{f \in \mathcal{G}_\tau} \sum_{i=1}^{n} X_{i,f}$$

where $X_{i,f} := (2M)^{-1} \overline{X}_{i,f}$.

Before writing the resulting concentration inequality, let us first compute (and bound) the quantity denoted by σ^2 in the statement of Proposition 8. For every $f \in \mathcal{G}_\tau$,

$$\begin{aligned}
4M^2 \sum_{i=1}^{n} \mathbb{E}\left[X_{i,f}^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[\left\langle \overline{f}_{K(i)}, \varepsilon_i \right\rangle_{\mathcal{H}}^2\right] &\leq \sum_{i=1}^{n} \left[\left\|\overline{f}_{K(i)}\right\|_{\mathcal{H}}^2 \mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{H}}^2\right]\right] \\
&= \sum_{K=1}^{D_\tau} \left[\left\|\overline{f}_K\right\|_{\mathcal{H}}^2 \sum_{i=\tau_{K-1}+1}^{\tau_K} v_i\right] \\
&= \sum_{K=1}^{D_\tau} \left[\left(\tau_K - \tau_{K-1}\right) \left\|\overline{f}_K\right\|_{\mathcal{H}}^2 v_K^\tau\right]
\end{aligned}$$

by Cauchy-Schwarz's inequality. So, by definition of $\mathcal{G}_\tau$ and σ^2,

$$\sigma^2 \leq \frac{1}{4M^2} \max_{1 \leq K \leq D_\tau} v_K^\tau \,.$$

We can now write what Proposition 8 proves about the concentration of $\|\Pi_\tau \varepsilon\|$: for every $x \geq 0$, with probability at least $1 - \mathrm{e}^{-x}$,

$$\|\Pi_\tau \varepsilon\| - \mathbb{E}\big[\|\Pi_\tau \varepsilon\|\big] \leq 2M\sqrt{2vx} + \frac{2Mx}{3} \leq \sqrt{2x\left(4M\mathbb{E}\big[\|\Pi_\tau \varepsilon\|\big] + \max_{1 \leq K \leq D_\tau} v_K^\tau\right)} + \frac{2Mx}{3} \quad ,$$

and similarly, with probability at least $1 - \mathrm{e}^{-x}$,

$$\|\Pi_\tau \varepsilon\| - \mathbb{E}\big[\|\Pi_\tau \varepsilon\|\big] \geq -\sqrt{2x\left(4M\mathbb{E}\big[\|\Pi_\tau \varepsilon\|\big] + \max_{1 \leq K \leq D_\tau} v_K^\tau\right)} - \frac{Mx}{4} \quad .$$

So, using a union bound, we have just proved Eq. (22). ■

B.3 Second method for choosing c_1, c_2 in Algorithm 1

We describe an alternative to the slope heuristics for choosing c_1, c_2 in Algorithm 1.

When prior information guarantee that the "variance" is almost constant and that no change occurs in some parts of the observed time series—say, at the start and at the end—, we can estimate this "variance" within each of these parts and take $c_1 = c_2$ equal to

$$\widehat{c}_{var} := 2\max\big(\widehat{v}_s, \widehat{v}_e\big), \tag{36}$$

where

$$\widehat{v}_s := \frac{1}{|I_s| - 1} \sum_{i \in I_s} \left[k(X_i, X_i) + \frac{1}{|I_s|^2} \sum_{j,\ell \in I_s} k(X_j, X_\ell) - \frac{2}{|I_s|} \sum_{j \in I_s} k(X_i, X_j)\right]$$

denotes the empirical variance of the start $(X_i)_{i \in I_s}$ of the time series, and $\widehat{v}_e$ is defined similarly from the end $(X_i)_{i \in I_e}$ of the time series. The fact that an estimate of the variance multiplied by 2 is a good choice for $c_1 = c_2$ is justified by the numerical experiments made by Lebarbier (2005) in the case of the linear kernel and one-dimensional data. This strategy was used successfully in the real-data experiments of an earlier version of the present paper (Arlot et al., 2012, Section 6.2).

Distribution	Mean	Variance
$\mathcal{B}(10, 0.2)$	2	1.6
$\mathcal{NB}(3, 0.7)$	$9/7 \approx 1.29$	$90/49 \approx 1.84$
$\mathcal{H}(10, 5, 2)$	1	$4/9 \approx 0.44$
$\mathcal{N}(2.5, 0.25)$	2.5	0.25
$\gamma(0.5, 5)$	2.5	12.5
$\mathcal{W}(5, 2)$	$\frac{5\sqrt{\pi}}{2} \approx 4.43$	$25(1 - \frac{\pi}{4}) \approx 5.37$
$\mathcal{P}ar(1.5, 3)$	$9/4 = 2.25$	$27/16 \approx 1.69$

Table 1: Scenario 1, mean and variance for the seven distributions considered.

B.4 Additional details about simulation experiments

Data generation process

Table 1 provides the values of the mean and variance of the seven distribution considered in Scenario 1. It shows that the pair (mean, variance) changes at every change-point in Scenario 1, but the mean sometimes stays constant.

Further results

This section gathers some additional results concerning the experiments of Section 6.

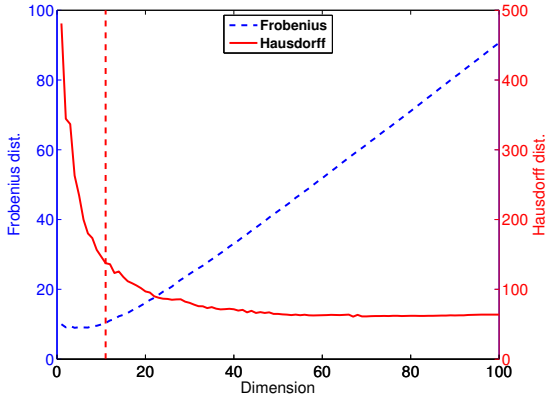
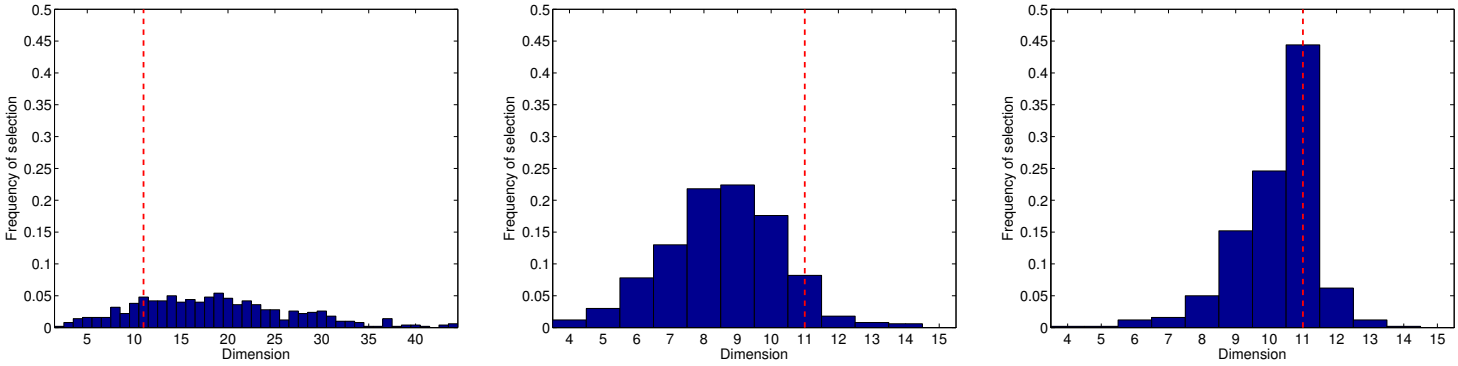
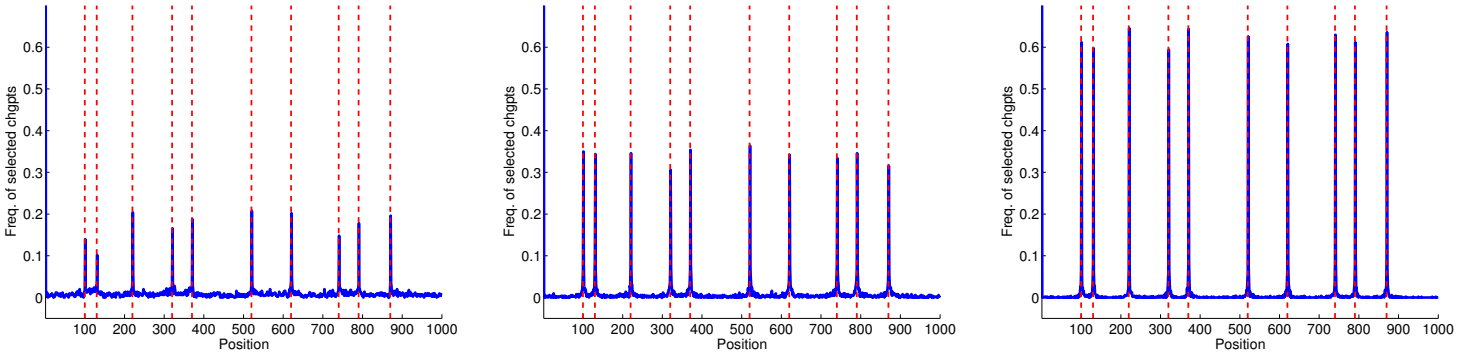


Figure 5: Scenario 1: $\mathcal{X} = \mathbb{R}$, variable (mean, variance). Performance of Algorithm 1 with kernel k^{lin}. Average distance (d_F or d_H) between $\widehat{\tau}(D)$ and $\tau^\star$, as a function of D.



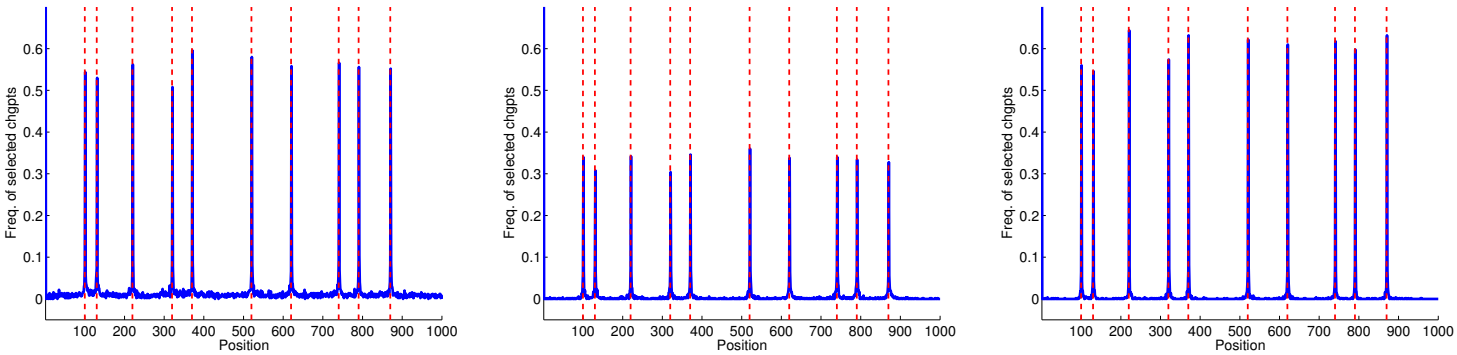
(a) $k = k^{\text{lin}}$ (b) $k = k^{\text{H}}$ (c) $k = k^{\text{G}}$

Figure 6: Scenario 1: $\mathcal{X} = \mathbb{R}$, variable (mean, variance). Algorithm 1 with three different kernels k. Distribution of $\widehat{D}$. (Figure 6c is a copy of Figure 2c, that we repeat here for making comparisons easier.)



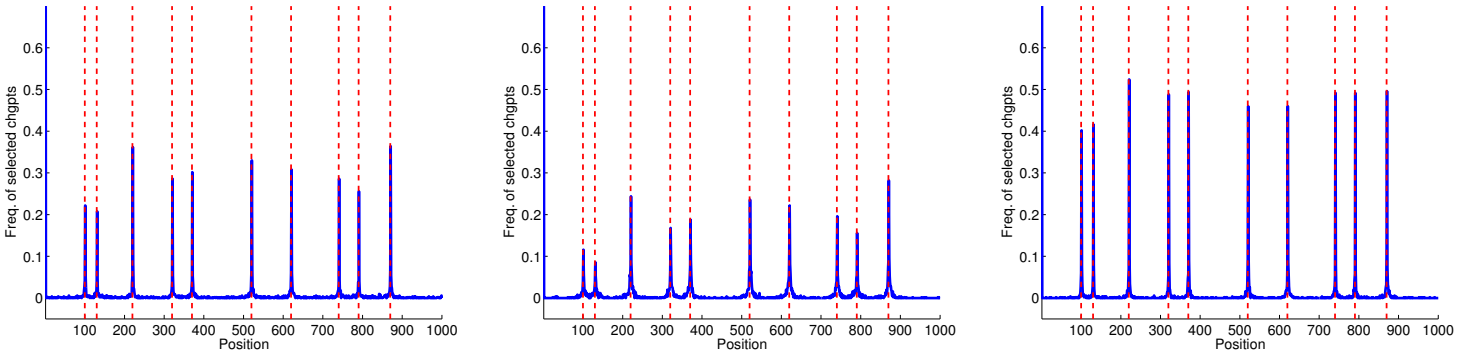
(a) $k = k^{\mathrm{lin}}$ (b) $k = k^{\mathrm{H}}$ (c) $k = k^{\mathrm{G}}$

Figure 7: Scenario 1: $\mathcal{X} = \mathbb{R}$, variable (mean, variance). Performance of Algorithm 1 with three different kernels. Probability, for each instant $i \in \{1, \ldots, n\}$, that $\widehat{\tau}(D^\star)$ puts a change-point at i.



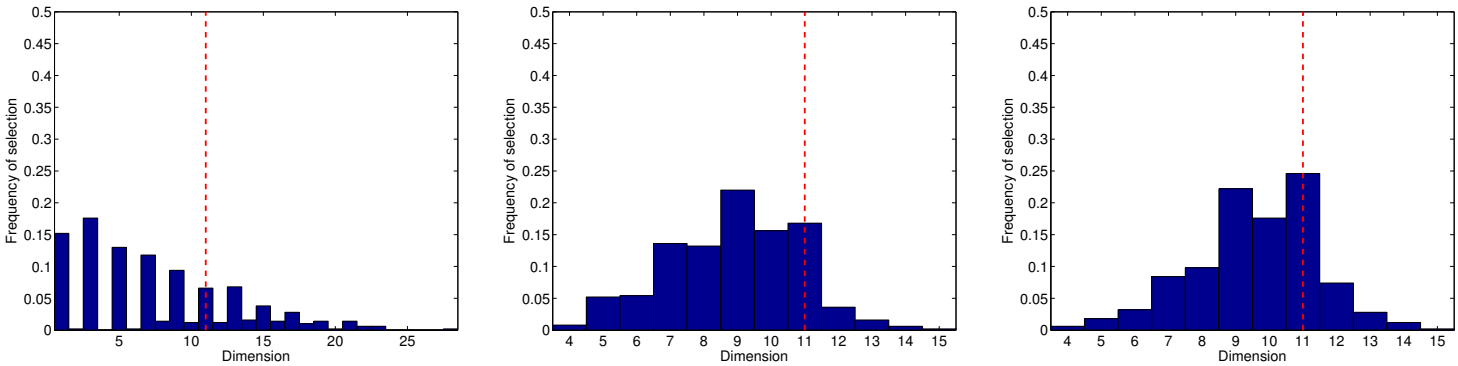
(a) $k = k^{\mathrm{lin}}$ (b) $k = k^{\mathrm{H}}$ (c) $k = k^{\mathrm{G}}$

Figure 8: Scenario 1: $\mathcal{X} = \mathbb{R}$, variable (mean, variance). Performance of Algorithm 1 with three different kernels. Probability, for each instant $i \in \{1, \ldots, n\}$, that $\widehat{\tau} = \widehat{\tau}(\widehat{D})$ puts a change-point at i.
For $k = k^{\mathrm{lin}}$, notice the high 'baseline' level of (wrong) detection of change-points, which is due to a frequent overestimation of the number of change-points, see Figure 6a.
(Figure 8c is a copy of Figure 2d, that we repeat here for making comparisons easier.)



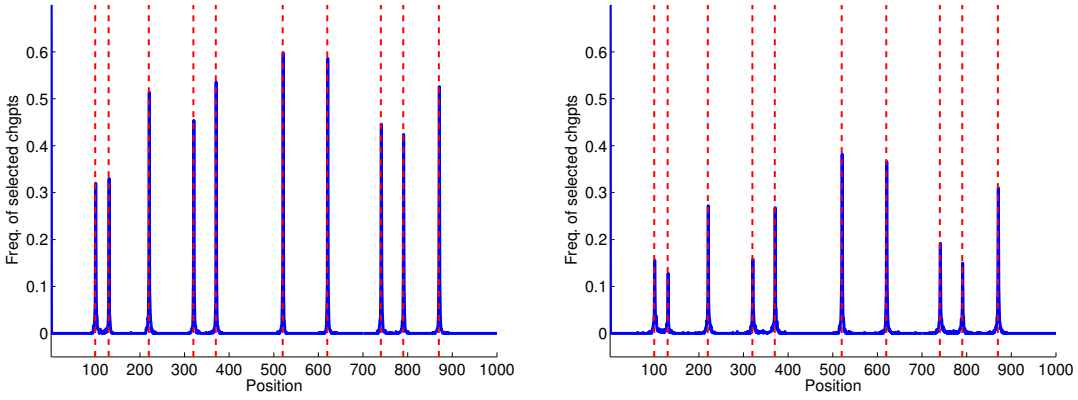
(a) $k = k^{\mathrm{lin}}$ (b) $k = k^{\mathrm{H}}$ (c) $k = k^{\mathrm{G}}$

Figure 9: Scenario 2: $\mathcal{X} = \mathbb{R}$, constant mean and variance. Performance of Algorithm 1 with three different kernels k. Probability, for each instant $i \in \{1, \ldots, n\}$, that $\widehat{\tau} = \widehat{\tau}(\widehat{D})$ puts a change-point at i.



(a) $k = k^{\mathrm{lin}}$ (b) $k = k^{\mathrm{H}}$ (c) $k = k^{\mathrm{G}}$

Figure 10: Scenario 2: $\mathcal{X} = \mathbb{R}$, constant mean and variance. Algorithm 1 with three different kernels k. Distribution of $\widehat{D}$.



(a) $k = k_{\chi^2}$ (b) $k = k^{\mathrm{G}}$

Figure 11: Scenario 3: histogram-valued data. Performance of Algorithm 1 with two different kernels. Probability, for each instant $i \in \{1, \ldots, n\}$, that $\widehat{\tau} = \widehat{\tau}(\widehat{D})$ puts a change-point at i.