



**HAL**  
open science

## Kernel change-point detection

Sylvain Arlot, Alain Celisse, Zaid Harchaoui

► **To cite this version:**

| Sylvain Arlot, Alain Celisse, Zaid Harchaoui. Kernel change-point detection. 2012. hal-00671174v1

**HAL Id: hal-00671174**

**<https://hal.science/hal-00671174v1>**

Preprint submitted on 17 Feb 2012 (v1), last revised 14 Mar 2019 (v3)

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

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

# Kernel change-point detection

**Sylvain Arlot**

SYLVAIN.ARLOT@ENS.FR

*CNRS ; Sierra Project-Team  
Laboratoire d'Informatique de l'Ecole Normale Supérieure  
(CNRS/ENS/INRIA UMR 8548)  
45, rue d'Ulm, 75 230 Paris, France*

**Alain Celisse**

CELISSE@MATH.UNIV-LILLE1.FR

*Laboratoire de Mathématiques Painlevé  
MODAL Project-Team  
UMR 8524 CNRS-Université Lille 1  
59 655 Villeneuve d'Ascq Cedex, France*

**Zaïd Harchaoui**

ZAID.HARCHAOU@INRIA.FR

*LEAR project-team and LJK  
655, avenue de l'Europe  
38 334 Saint Ismier Cedex, France*

**Editor:** To be submitted

## Abstract

We tackle the change-point problem with data belonging to a general set. We propose a penalty for choosing the number of change-points in the kernel algorithm of [Harchaoui and Cappé \(2007\)](#). This penalty generalizes the one proposed for one dimensional signals by [Lebarbier \(2005\)](#). We prove it satisfies a non-asymptotic oracle inequality by showing new concentration results in Hilbert spaces. Experiments on synthetic and real data illustrate the accuracy of our method, showing it can detect changes in the whole distribution, even when the mean and variance are constant. Our algorithm can also deal with data of complex nature, such as the GIST descriptors which are commonly used for video temporal segmentation.

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

## 1. Introduction

A central topic in machine learning is finding the boundary between samples drawn from different probability distributions. This goal is at the intersection of supervised learning (such as binary classification, see [Vapnik, 1998](#); [Steinwart and Christmann, 2008](#)) and unsupervised learning (such as clustering, see [von Luxburg, 2009](#)). In the latter case, a major theoretical issue arises when considering real-world problems, namely the model selection issue which corresponds to selecting the number of clusters ([Ben-David et al., 2006](#); [von Luxburg, 2009](#)). This issue is still an open problem.

In this paper, we consider a related topic, the change-point problem ([Carlstein et al., 1994](#)). Let  $X_1, \dots, X_n$  be a sequence of independent random variables, whose distribution abruptly changes at given unknown instants (change-points). The change-point problem

consists in (i) estimating the change-point locations given their number, (ii) determining the number of change-points.

Given a positive semi-definite kernel  $k$  and its associated feature map  $\Phi$ , our approach is to solve the change-point regression problem *via* model selection with  $\Phi(X_1), \dots, \Phi(X_n) \in \mathcal{H}$  some Hilbert space, by extending the work of [Lebarbier \(2005\)](#) to the Hilbertian setting.

Unlike usual model selection approaches in the one dimensional setting focusing on changes in the mean or the variance ([Lavielle, 2005](#); [Lebarbier, 2005](#)), our approach can capture changes in higher-order moments of probability distributions, using the machinery of reproducing kernel Hilbert spaces. Another strength of the kernelized least-squares algorithm we propose is it can process time series with observations of any nature, as long as some positive-definite kernel can be defined on their support, including data belonging to some structured spaces such as the  $d$ -dimensional simplex. This is particularly appropriate for temporal segmentation of video streams (see [Section 6](#)) for automatic summarization of video archives. For multivariate signals in  $\mathbb{R}^d$ , other approaches were recently proposed, mainly dedicated to biological applications. [Picard et al. \(2011\)](#) focus on changes in the mean and make a Gaussian assumption on the signal. [Bleakley and Vert \(2011\)](#) propose a fused lasso based algorithm to perform segmentation of the mean as well. Our approach is more general since it is not limited to changes in the mean and does not rely on any distributional assumption on the intra-segment distributions.

*Without assuming the number of change-points is known*, our algorithm makes use of the efficient algorithm of [Harchaoui and Cappé \(2007\)](#). This is a significant improvement for practical application. Furthermore, we prove theoretical guarantees for our data-driven choice of the number of change-points, with a non-asymptotic oracle inequality ([Theorem 1](#)).

The main contributions of the paper are the following: (i) proposing a penalty extending the one of [Lebarbier \(2005\)](#) to the kernel change-point problem, which allows a data-driven choice of the number of change-points, (ii) proving it satisfies a non-asymptotic oracle inequality ([Theorem 1](#)), by developing new concentration results in Hilbert spaces, (iii) showing with experiments ([Section 6](#)) the resulting algorithm is promising in terms of applications, both for detecting changes in distribution that are not changes in the mean or the variance, and for analyzing data of complex nature such as video streams.

## 2. Model selection for the change-point problem: one-dimensional data

Let us start by summarizing how the change-point problem has been cast as a model selection problem in the case of one-dimensional data ([Lavielle, 2005](#); [Lebarbier, 2005](#)). Let  $0 \leq t_1 < \dots < t_n \leq 1$  be deterministic instants of observation,  $\mu^*$  some measurable function  $[0, 1] \rightarrow \mathcal{H} = \mathbb{R}$  and

$$\forall i \in \{1, \dots, n\}, \quad Y_i = \mu_i^* + \varepsilon_i, \quad \text{where} \quad \mu_i^* = \mu^*(t_i)$$

and  $\varepsilon_1, \dots, \varepsilon_n$  are independent and identically distributed random variables with  $\mathbb{E}[\varepsilon_i] = 0$  and  $\mathbb{E}[\varepsilon_i^2] = \sigma^2 > 0$ . The mean  $\mu^*(t_i)$  of the observations  $Y_i$  is assumed piecewise constant and the goal is to find *change-points*, that is the location of jumps in the mean. A classical approach is to solve a least-squares regression problem by estimating  $\mu^*$  with a piecewise constant function, with the number of change-points selected through a model selection

procedure (see [Yao, 1988](#); [Yao and Au, 1989](#); [Lavielle and Moulines, 2000](#); [Boysen et al., 2009](#), and Section 4.3).

Since  $\mu^*$  is only evaluated at  $t_1, \dots, t_n$ , it is considered as an element of  $\mathcal{H}^n$  with its Euclidean structure given by  $\|f - g\|^2 = \sum_{i=1}^n (f(t_i) - g(t_i))^2$  for every  $f, g \in \mathcal{H}^n$ . We also use the notation  $Y = (Y_1, \dots, Y_n)' \in \mathcal{H}^n$ . For every function  $f : [0, 1] \rightarrow \mathcal{H}$ , we define respectively its quadratic and empirical risk

$$\mathcal{R}(f) := \frac{1}{n} \|f - \mu^*\|^2 \quad \text{and} \quad \widehat{\mathcal{R}}_n(f) := \frac{1}{n} \|f - Y\|^2 . \quad (1)$$

Let  $\mathcal{M}_n$  be the set of segmentations of  $\{1, \dots, n\}$ , that is, the set of partitions of the form  $\{\{1, \dots, k_1\}, \{k_1+1, \dots, k_2\}, \dots, \{k_{D-1}-1, \dots, n\}\}$  with  $D \geq 1$  and  $1 \leq k_1 < \dots < k_{D-1} \leq n$ . For every  $m \in \mathcal{M}_n$ , let  $D_m = \text{Card}(m)$  and  $S_m$  be the set of functions  $\{t_1, \dots, t_n\} \rightarrow \mathcal{H}$  that are constant over  $(t_i)_{i \in \lambda}$  for every segment  $\lambda \in m$ . Then, the associated empirical risk minimizer, called regressogram, is defined by

$$\widehat{\mu}_m \in \operatorname{argmin}_{f \in S_m} \left\{ \widehat{\mathcal{R}}_n(f) \right\}, \quad \text{so that} \quad \forall \lambda \in m, \forall i \in \lambda, \quad \widehat{\mu}_m(t_i) = \frac{1}{\text{Card}(\lambda)} \sum_{j \in \lambda} Y_j .$$

The goal is to build a data-driven choice  $\widehat{m} \in \mathcal{M}_n$  such that the quadratic risk  $\mathcal{R}(\widehat{\mu}_{\widehat{m}})$  is minimal. Following [Birgé and Massart \(2001\)](#) and [Lebarbier \(2005\)](#), this model selection problem can be solved in a non-asymptotic manner by penalization:

$$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \|\widehat{\mu}_m - Y\|^2 + \text{pen}(m) \right\} , \quad (2)$$

$$\text{where} \quad \text{pen}(m) = \text{pen}_{\text{BM}}(m) := \frac{\sigma^2 D_m}{n} \left( c_1 \log \left( \frac{n}{D_m} \right) + c_2 \right) \quad \text{with} \quad c_1, c_2 > 0 . \quad (3)$$

If the noise variables  $\varepsilon_i$  are Gaussian, [Lebarbier \(2005\)](#) proved that Eq. (2) leads to an oracle inequality, that is, constants  $c_1, c_2, K_1, K_2 > 0$  exist such that

$$\mathbb{E} \left[ \frac{1}{n} \|\widehat{\mu}_{\widehat{m}} - \mu^*\|^2 \right] \leq K_1 \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \|\widehat{\mu}_m - \mu^*\|^2 + \text{pen}_{\text{BM}}(m) \right\} + \frac{K_2 \sigma^2}{n} . \quad (4)$$

The  $\log(n)$  term in the penalty is the unavoidable price for ignoring change-point locations ([Birgé and Massart, 2007](#)). Furthermore, extensive simulation experiments of [Lebarbier \(2005\)](#) suggested the values  $c_1 = 2$ ,  $c_2 = 5$  and an efficient data-driven way of estimating  $\sigma^2$ , called the slope heuristics.

### 3. Kernel change-point problem

Let us now describe how we generalize the approach of Section 2 to detecting changes in the probability distribution of the signals that belong to any set (not necessarily vector spaces).

#### 3.1 Problem

Let  $\mathcal{X}$  be some set and assume we observe independent random variables  $X_1, \dots, X_n \in \mathcal{X}$  at time  $t_1, \dots, t_n$  with a piecewise-constant probability distribution. The goal is to find

abrupt changes *in the distribution* of the time series  $X_1, \dots, X_n$ , whereas classical change-point estimation seeks for changes in the first moments of the distribution such as the mean or the variance (Korostelev and Korosteleva, 2011). Let  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$  be some positive definite kernel,  $\mathcal{H} = \mathcal{H}_k$  the associated reproducing kernel Hilbert space, and  $\Phi : \mathcal{X} \rightarrow \mathcal{H}$  the canonical feature map defined by  $\Phi(x) = k(x, \cdot)$  (see Schölkopf and Smola, 2001; Cucker and Zhou, 2007; Steinwart and Christmann, 2008, for a detailed presentation of reproducing kernel Hilbert spaces). Then, for every  $i \in \{1, \dots, n\}$  we define

$$Y_i = \Phi(X_i) \in \mathcal{H}$$

and  $\mu_i^* \in \mathcal{H}$  the mean element of the distribution of  $X_i$ , that is,

$$\forall g \in \mathcal{H}, \quad \langle \mu_i^*, g \rangle_{\mathcal{H}} = \mathbb{E}[g(X_i)] = \mathbb{E}[\langle Y_i, g \rangle_{\mathcal{H}}] .$$

Following Sriperumbudur et al. (2008, 2010), we can exploit the strong connection between the mean element  $\mu_i^*$  and the distribution of  $X_i$ . For instance with translation-invariant kernels satisfying a condition on their Fourier transform, equality of mean elements implies equality of probability distributions (Sriperumbudur et al., 2008). So, we can focus on detecting changes in the mean elements, assuming

$$\mu_1^* = \dots = \mu_{k_1^*}^*, \quad \mu_{k_1^*+1}^* = \dots = \mu_{k_2^*}^*, \quad \dots \quad \mu_{k_{D^*-1}^*+1}^* = \dots = \mu_n^*$$

for some  $1 \leq k_1^* < \dots < k_{D^*-1}^* \leq n$  (the true change-point indices). Moreover if we define  $\varepsilon_i := Y_i - \mu_i^*$  (for which we assume its ‘‘variance’’  $v_i = \mathbb{E}[\|\varepsilon_i\|_{\mathcal{H}}^2]$  is finite for every  $i$ ), the approach of Section 2 formally extends from  $\mathcal{H} = \mathbb{R}$  to any Hilbert space  $\mathcal{H}$ . The quadratic and empirical risks of  $f \in \mathcal{H}^n$  are then defined by Eq. (1) again with  $\|f - g\|^2 = \sum_{i=1}^n \|f_i - g_i\|_{\mathcal{H}}^2$  for every  $f, g \in \mathcal{H}^n$ .

The rest of the paper provides theoretical grounds for such an extension, showing in particular a penalty of the form (3) can still be used in the kernel setting with  $\sigma^2$  replaced by an upper bound on  $\max_i v_i$ . Since we aim at analyzing high-dimensional time series, we will provide an analysis from the non-asymptotic point of view, by proving an oracle inequality similar to Eq. (4). Note that such an extension is formally straightforward, but it still requires to solve some theoretical issues, since some key elements for proving Eq. (4) are no longer valid in the Hilbertian setting. These issues are detailed in the next subsection.

### 3.2 Related work and theoretical challenges

A kernelized version of the approach of Section 2 was proposed by Harchaoui and Cappé (2007), but assuming the number of change-points is known. Our algorithm is the same for every fixed number of change-points, but goes one step further, since *we do not assume the number of changes is known a priori*.

The penalty (3) and the proofs of Birgé and Massart (2001) and Lebarbier (2005) cannot be extended directly in our case because (i)  $Y_i = \Phi(X_i)$  are not *real* but *Hilbert space valued* random variables (with a possibly infinite-dimensional Hilbert space), (ii) Birgé and Massart’s approach heavily relies on the assumption that the noise  $\varepsilon_i$  is *Gaussian with a constant variance* which is questionable in our Hilbertian setting. Indeed, if data were Gaussian in the feature space, then any linear projection would follow a Gaussian

distribution, and kernel principal component analysis with usual kernels indicate this does not hold for most real data sets.

The key step in Birgé and Massart’s approach is to design a penalty  $\text{pen}(\cdot)$  such that

$$\forall m \in \mathcal{M}_n, \quad \text{pen}(m) \geq \text{pen}_{\text{id}}(m) := \frac{1}{n} \|\widehat{\mu}_m - \mu^*\|^2 - \frac{1}{n} \|\widehat{\mu}_m - Y\|^2 \quad (5)$$

with high probability, without taking  $\text{pen}(m)$  larger than necessary. The quantity  $\text{pen}_{\text{id}}(m)$  is called “ideal penalty” since using it in Eq. (2) would lead to minimizing the quadratic risk. For proving Eq. (5), Birgé and Massart (2001) use the concentration properties of functions of Gaussian variables.

In our non-Gaussian Hilbertian setting, two concentration inequalities could be used instead: (i) Pinelis-Sakhanenko’s inequality (Pinelis and Sakhanenko, 1986), (ii) Talagrand’s inequality (see Bousquet, 2002). The first one cannot be used as such since it is not a concentration but a deviation inequality, hence too loose for our purpose. The second one is not accurate enough in our setting because it yields too large deviation terms, see Remark 7 in the appendix.

#### 4. Oracle inequality for the kernel change-point problem

This section shows how the penalty (3) can be extended to the Hilbertian setting of Section 3, by proving an oracle inequality (Theorem 1).

##### 4.1 Assumptions

Without a Gaussian homoscedastic assumption, we need to assume the following. Let us recall  $v_i := \mathbb{E}[\|Y_i - \mu_i^*\|_{\mathcal{H}}^2] = \mathbb{E}[\|\varepsilon_i\|_{\mathcal{H}}^2]$ , for every  $i$ .

$$\text{Bounded data/kernel : } \exists M > 0, \quad \sup_{1 \leq i \leq n} \|Y_i\|_{\mathcal{H}}^2 = k(X_i, X_i) \leq M^2 \text{ a.s.} \quad (\mathbf{Db})$$

$$\text{Bounded variance : } \exists v_{\max} < +\infty, \quad \max_{1 \leq i \leq n} v_i \leq v_{\max} \quad (\mathbf{Vmax})$$

$$\text{Minimal variance : } \exists 0 < c_{\min} < +\infty, \quad \min_{1 \leq i \leq n} v_i \geq \frac{M^2}{c_{\min}} =: v_{\min} > 0 . \quad (\mathbf{Vmin})$$

Let us make a few remarks:

- **(Db)** implies **(Vmax)** with  $v_{\max} = M^2$  since  $v_i = \mathbb{E}[k(X_i, X_i)] - \|\mu_i^*\|_{\mathcal{H}}^2 \leq M^2$ .
- if  $k$  is translation invariant, that is,  $k(x, x') = k(x - x')$  (e.g., the Gaussian and Laplace kernels), then  $v_i = k(0) - \|\mu_i^*\|_{\mathcal{H}}^2$  so that **(Vmax)** and **(Vmin)** are assumptions on  $\|\mu_i^*\|_{\mathcal{H}}$ .
- if **(Db)** holds true,  $v_i = \text{tr}(\Sigma_i)$  where  $\Sigma_i$  is the covariance operator of  $\Phi(X_i)$ .
- if  $\mathcal{X} = \mathbb{R}^d$  and  $k(x, y) = \langle x, y \rangle$ ,  $v_i = \text{tr}(\Sigma_i)$  where  $\Sigma_i$  is the covariance matrix of  $\varepsilon_i$ .

## 4.2 Oracle inequality for change-point estimation

The following theorem shows an oracle inequality still holds for the kernel change-point problem with a penalty of the form (3) where  $\sigma^2$  is replaced by  $v_{\max}$ , up to numerical constants.

**Theorem 1** *Let us consider the kernel change-point problem described in Section 3. Assume (Db), (Vmin) and (Vmax) hold true. Then, some numerical constant  $L_1 > 0$  exists such that for every  $x > 0$ , an event of probability at least  $1 - e^{-x}$  exists on which, for any  $C \geq c_{\min}^2 L_1$  and any*

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ \hat{\mathcal{R}}_n(\hat{\mu}_m) + \operatorname{pen}(m) \right\} \text{ with } \operatorname{pen}(m) = \frac{C v_{\max} D_m}{n} \left[ 1 + \log \left( \frac{n}{D_m} \right) \right], \quad (6)$$

$$\mathcal{R}(\hat{\mu}_{\hat{m}}) \leq 2 \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}(\hat{\mu}_m) + 2 \operatorname{pen}(m) \right\} + \frac{C (\log 4 + x) v_{\max}}{n}. \quad (7)$$

A sketch of proof of Theorem 1 is given in Section 4.4 and a complete proof can be found in Appendix B.5. If  $\mathcal{X} = \mathbb{R}$ ,  $k(x, y) = xy$ , and  $\forall i, v_i = v_{\max} > 0$ , we recover (Theorem 1) an oracle inequality similar to the one of Lebarbier (2005).

Note that (Db) is a classical assumption in the machine learning literature on kernels. It holds true for instance with bounded kernels such as the Gaussian kernel (see Section 5.2). In particular, it avoids assuming data are Gaussian as in Birgé and Massart (2001). Assumptions (Vmin)–(Vmax) are a natural extension of homoscedastic setting of Birgé and Massart (2001), which would not be realistic in our Hilbertian setting. Note that a fully heteroscedastic setting might be considered, for instance following the ideas of Arlot and Celisse (2011), but with a more complex algorithm and no theoretical guarantees. We choose (Vmin)–(Vmax) as a compromise between these two extremes.

The constant 2 in front of the oracle inequality (7) can be chosen arbitrary close to 1, at the price of an increase of the numerical constant  $L_1$  (which appears in the penalty and in the remainder term through  $C$ ). Besides, the constant  $C$  suggested by the proof of Theorem 1 certainly is not tight, as in all similar non-asymptotic oracle inequalities.

Finally, let us mention a byproduct of the proof of Theorem 1 which is detailed in Appendix A: If some prior knowledge restricts the possible positions of change-points to a subset of  $\{t_1, \dots, t_n\}$  with  $\mathcal{O}(\log n)$  elements, then a smaller penalty can be used instead of (6), leading to an oracle inequality that is optimal in the homoscedastic case.

## 4.3 Discussion: change-point problem and oracle inequalities

Let us discuss the relationship between minimizing the risk (proving an oracle inequality like Eq. (4)) and the original change-point problem.

In the one-dimensional setting ( $\mathcal{X} = \mathcal{H} = \mathbb{R}$ ), an oracle inequality shows that  $\hat{\mu}_{\hat{m}}$  is close to the best piecewise-constant estimator of  $\mu^*$  in terms of quadratic risk. So, we can roughly expect that  $\hat{m}$  detects all jumps of size  $(\mu^*(t_{i+1}) - \mu^*(t_i))^2$  significantly larger than the noise-level  $\sigma^2/N$ , where  $N$  is the number of observations available around the jump. In the non-asymptotic point of view, it seems reasonable (and desirable) to aim only at detecting jumps for which enough observations are available, which explains why the procedure proposed by Lebarbier (2005) yields good results in terms of change-point estimation.

In the kernelized version of this approach, a similar heuristics holds (as confirmed by our simulation experiments, see Section 6). However, both the size of a jump, now measured by  $\|\mu_{i+1}^* - \mu_i^*\|_{\mathcal{H}}^2$ , and the noise-level depend on the kernel  $k$ . So,  $k$  should be chosen in order to maximize the signal-to-noise ratio at every true change-point.

For instance, even when  $\mathcal{X} = \mathbb{R}$ , choosing an appropriate kernel  $k$  can lead to detect changes in the mean (with  $k(x, y) = xy$ ), but also in other features of the distribution (for instance with the Gaussian kernel, see the experiments of Section 6). Therefore, kernelizing Birgé and Massart's approach can also be useful in the one-dimensional case when we do not look for changes in the mean.

#### 4.4 Sketch of the proof of Theorem 1

The proof mostly follows the general approach of Birgé and Massart for proving an oracle inequality, that is, we prove new concentration inequalities (Propositions 2 and 3) that are needed to show the penalty defined by Eq. (6) satisfies Eq. (5) with a large probability. Note that our proof actually leads to a more general model selection result (Theorem 8 in the appendix) which admits corollaries of independent interest (see Appendix A).

##### 4.4.1 ELEMENTARY COMPUTATIONS

The proof starts by splitting the ideal penalty defined by Eq. (6) into two terms that will be concentrated separately. All statements that are not proved here are detailed in Appendix B.1.

Recall that for every  $m \in \mathcal{M}_n$ ,  $S_m$  is the vector space of functions  $\{t_1, \dots, t_n\} \rightarrow \mathcal{H}$  that are constant over each  $\lambda \in m$ , and all functions  $f : \{t_1, \dots, t_n\} \rightarrow \mathcal{H}$  are written as elements of  $\mathcal{H}^n$  by denoting  $f_i = f(t_i)$ . In particular,  $S_m$  is considered as a linear subspace of  $\mathcal{H}^n$ . For  $f, g \in \mathcal{H}^n$ , let  $\langle f, g \rangle := \sum_{i=1}^n \langle f_i, g_i \rangle_{\mathcal{H}}$  denote the canonical scalar product in  $\mathcal{H}^n$ . The associated regressogram estimator is uniquely defined by

$$\hat{\mu}_m = \Pi_m Y \quad \text{where} \quad \forall g \in \mathcal{H}^n, \quad \Pi_m g := \operatorname{argmin}_{f \in S_m} \left\{ \frac{1}{n} \|f - g\|_{\mathcal{H}}^2 \right\}$$

is the orthogonal projection of  $g$  onto  $S_m$ . We define also  $\mu_m^* := \Pi_m \mu^*$ , and remark that

$$\forall g \in \mathcal{H}^n, \forall \lambda \in m, \forall i \in \lambda, \quad (\Pi_m g)_i = \frac{1}{\operatorname{Card}(\lambda)} \sum_{j \in \lambda} g_j . \quad (8)$$

Then,

$$\operatorname{pen}_{\text{id}}(m) = \frac{2}{n} \|\Pi_m \varepsilon\|^2 - \frac{2}{n} \langle (I - \Pi_m) \mu^*, \varepsilon \rangle - \frac{1}{n} \|\varepsilon\|^2 . \quad (9)$$

The term  $n^{-1} \|\varepsilon\|^2$  does not depend on  $m$  so it can be removed from the ideal penalty. The expectations of the two other terms are given by

$$\mathbb{E} [\langle (I - \Pi_m) \mu^*, \varepsilon \rangle] = 0 \quad \text{and} \quad \mathbb{E} \left[ \|\Pi_m \varepsilon\|^2 \right] = \sum_{\lambda \in m} v_\lambda \quad \text{where} \quad v_\lambda := \frac{1}{\operatorname{Card}(\lambda)} \sum_{i \in \lambda} v_i \quad (10)$$

$$\text{so that} \quad \mathbb{E} \left[ \operatorname{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 \right] = \frac{2}{n} \sum_{\lambda \in m} v_\lambda . \quad (11)$$



Then, the key results we need for showing the penalty (6) satisfies Eq. (5) are concentration inequalities for  $\langle (I - \Pi_m)\mu^*, \varepsilon \rangle$  (Proposition 2) and for  $\|\Pi_m \varepsilon\|^2$  (Proposition 3).

#### 4.4.2 TWO NEW CONCENTRATION INEQUALITIES

First, for the linear term, we prove in Appendix B.2 the following result, mostly by applying Bernstein's inequality.

**Proposition 2 (Concentration of the linear term)** *Let  $m \in \mathcal{M}_n$  and  $\Pi_m$  be defined by Eq. (8). If (Db) holds true, then for every  $x > 0$ , with probability at least  $1 - 2e^{-x}$ ,*

$$\forall \theta > 0, \quad |\langle (I - \Pi_m)\mu^*, \varepsilon \rangle| \leq \theta \|\mu_m^* - \mu^*\|^2 + \left( \frac{v_{\max}}{2\theta} + \frac{4M^2}{3} \right) x. \quad (12)$$

Second, for the quadratic term, we prove in Appendix B.3 the following result, that relies on a combination of Bernstein and Pinelis-Sakhanenko inequalities. Note that directly using Talagrand's inequality (Bousquet, 2002) would lead to a less precise result in our setting, see Remark 7 in the appendix for details.

**Proposition 3 (Concentration of the quadratic term)** *Let  $m \in \mathcal{M}_n$  and  $\Pi_m$  be defined by Eq. (8). If (Db), (Vmin) and (Vmax) hold true, then, for every  $x > 0$ , with probability at least  $1 - 2e^{-x}$ ,*

$$\forall \theta \in (0, 1], \quad \left| \|\Pi_m \varepsilon\|^2 - \mathbb{E} \left[ \|\Pi_m \varepsilon\|^2 \right] \right| \leq \theta \mathbb{E} \left[ \|\Pi_m \varepsilon\|^2 \right] + \frac{49c_{\min}^2 v_{\max} x}{\theta}. \quad (13)$$

#### 4.4.3 CONCLUSION OF THE PROOF

The first step towards Eq. (5) is to get a uniform concentration inequality for the ideal penalty from the combination of Eq. (9), Eq. (11), Proposition 2 and Proposition 3: for every  $x \geq 0$ , an event  $\Omega_m(x)$  of probability at least  $1 - 4e^{-x}$  exists on which

$$\forall \theta \in (0, 1], \quad \left| \text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 - \frac{2}{n} \sum_{\lambda} v_{\lambda} \right| \leq \frac{4\theta}{n} \|\mu^* - \hat{\mu}_m\|^2 + r(x, \theta), \quad (14)$$

where  $r(x, \theta) := 213c_{\min}^2 v_{\max} x / (n\theta)$ . By definition (6) of  $\hat{m}$ , for every  $m \in \mathcal{M}_n$ ,

$$\frac{1}{n} \|\mu^* - \hat{\mu}_{\hat{m}}\|^2 + [\text{pen}(\hat{m}) - \text{pen}_{\text{id}}(\hat{m})] \leq \frac{1}{n} \|\mu^* - \hat{\mu}_m\|^2 + [\text{pen}(m) - \text{pen}_{\text{id}}(m)]. \quad (15)$$

Therefore, uniform bounds on the deviations of  $\text{pen}(m) - \text{pen}_{\text{id}}(m) - n^{-1} \|\varepsilon\|^2$  are sufficient to get an oracle inequality. Let  $(x_m)_{m \in \mathcal{M}_n} \in (0, +\infty)^{\mathcal{M}_n}$  to be chosen later, and define the event  $\Omega := \bigcap_{m \in \mathcal{M}_n} \Omega_m(x_m)$ . By the union bound,  $\mathbb{P}(\Omega) \geq 1 - 4 \sum_{m \in \mathcal{M}_n} e^{-x_m}$ . Then, combining Eq. (14) and (15), for every penalty such that  $\text{pen}(m) \geq 2n^{-1} \sum_{\lambda \in \mathcal{M}} v_{\lambda} + r(x_m, \theta)$  for every  $m \in \mathcal{M}_n$ , on  $\Omega$ , for every  $\theta \in (0, 1]$ ,

$$\frac{1 - 4\theta}{n} \|\mu^* - \hat{\mu}_{\hat{m}}\|^2 \leq \inf_{m \in \mathcal{M}} \left\{ \frac{1 + 4\theta}{n} \|\mu^* - \hat{\mu}_m\|^2 + \text{pen}(m) - \frac{2}{n} \sum_{\lambda \in \mathcal{M}} v_{\lambda} + r(x_m, \theta) \right\}.$$

This proves a general oracle inequality (stated as Theorem 8 in the appendix), which implies Theorem 1 by taking  $x_m = D_m(\log(2) + 1 + \log(\frac{n}{D_m})) + \log 4 + x$ . Indeed, taking  $\theta = 1/12$ , we get  $2 \sum_{\lambda \in m} v_\lambda + nr(x_m, \theta) \leq v_{\max} D_m C [1 + \log(\frac{n}{D_m})] + C_3$  for some constants  $C, C_3$ , and  $\hat{m}$  remains unchanged by removing  $C_3$  from the penalty. Finally, the probability of  $\Omega^c$  is upper bounded by

$$\sum_{1 \leq D \leq n} \text{Card} \{m \in \mathcal{M}_n / D_m = D\} e^{-D(\log(2)+1+\log(\frac{n}{D})) - x} \leq e^{-x} \sum_{D \geq 1} 2^{-D} = e^{-x} .$$

## 5. Kernel multiple change-point algorithm

This section summarizes the multiple change-point estimation algorithm suggested by Theorem 1, and gives some examples of kernels for *vectorial* and *non-vectorial* data.

### 5.1 Algorithm

**Input:** observations  $X_1, \dots, X_n \in \mathcal{X}$ , a positive definite kernel  $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ , some constants  $C \geq 1$ ,  $D_{\max} \leq n$  and  $v_{\max}$  such that **(Vmax)** holds true.

1. Define  $\Phi(x) = k(x, \cdot) \in \mathcal{H}$ , for every  $x \in \mathcal{X}$  and  $Y = (\Phi(X_i))_{1 \leq i \leq n} \in \mathcal{H}^n$ .
2. Define  $\hat{\mu}_m \in \mathcal{H}^n$  such that  $\forall \lambda \in m, \forall i \in \lambda, (\hat{\mu}_m)_i = n^{-1} \sum_{j \in \lambda} \Phi(X_j)$ , for every  $m \in \mathcal{M}_n$ , where  $\mathcal{M}_n$  denotes the set of segmentations of  $\{1, \dots, n\}$ .
3. Compute  $\hat{m}_D \in \text{argmin}_{m \in \mathcal{M}_n, D_m = D} \{n^{-1} \|Y - \hat{\mu}_m\|^2\}$ , for every  $D \in \{1, \dots, D_{\max}\}$ .
4. Compute  $\hat{D} \in \text{argmin}_{D \in \{1, \dots, D_{\max}\}} \{n^{-1} \|Y - \hat{\mu}_m\|^2 + \frac{C v_{\max} D_m}{n} (\log(\frac{n}{D_m}) + 1)\}$ .

**Output:** segmentation  $\hat{m} = \hat{m}_{\hat{D}}$ .

The above algorithm can be seen as a kernelized version of the one proposed by [Lebarbier \(2005\)](#). Our main contributions are the theoretical guarantees of Section 4 and the experiments of Section 6.

**Computational complexity:** For each fixed  $D$ , step 3 is the dynamic programming algorithm proposed by [Harchaoui and Cappé \(2007\)](#); see also [\(Kay, 1993\)](#). Computing  $(\hat{m}_D)_{1 \leq D \leq D_{\max}}$  requires at most  $\mathcal{O}(D_{\max} n^2)$  times the cost of computing any  $k(X_i, X_j)$ .

**Setting  $v_{\max}$ :** If **(Db)** holds true, one can always take  $v_{\max} = M^2$ , but this bound might be loose. In most real-world applications, it is realistic to assume that  $0 < \underline{t} < \bar{t} < 1$  are known such that all the change-points belong to  $[\underline{t}, \bar{t}]$ , that is,  $0 < \underline{t} < t_1^*$  and  $t_{D^*-1}^* < \bar{t} < 1$ . Such “edge instants” can usually be inferred from real-world knowledge, as in Section 6.2. Then, the signal is stationary over  $[0, \underline{t}]$  and over  $[\bar{t}, 1]$ , and we propose to estimate  $v_{\max}$  by

$$\hat{v}_{\max} = \max \left\{ \text{tr} \left( \hat{\Sigma}_{0:\underline{t}} \right), \text{tr} \left( \hat{\Sigma}_{\bar{t}:1} \right) \right\} \quad (16)$$

where  $\hat{\Sigma}_{a:b}$  is the empirical covariance estimator of  $(\Phi(X_i))_{a \leq i \leq b}$ . We shall use this estimate in the experiments of Section 6.

## 5.2 Examples of kernels

The algorithm of Section 5.1 can be used with various sets  $\mathcal{X}$  (not necessarily vector spaces), and with several different kernels  $k$  for a given  $\mathcal{X}$ . In particular, our approach is flexible with respect to the nature of data. It can handle any type of data as long as positive-definite kernel similarity measure for such data is available. Instances of such data are simplicial data (histograms), texts, trees, among others (Shawe-Taylor and Cristianini, 2004). Some classical kernel choices are detailed below.

- when  $\mathcal{X} = \mathbb{R}$ ,  $k(x, y) = xy$  and we recover the algorithm by Lebarbier (2005) since  $\|\Phi(x) - \Phi(x')\|_{\mathcal{H}}^2 = (x - x')^2$ .
- when  $\mathcal{X} = \mathbb{R}^d$ ,  $k(x, y) = \langle x, y \rangle_{\mathbb{R}^d}$  yields its natural extension since  $\|\Phi(x) - \Phi(x')\|_{\mathcal{H}}^2 = \sum_{i=1}^d (x_i - x'_i)^2$  the squared Euclidean norm in  $\mathbb{R}^d$ .
- when  $\mathcal{X} = \mathbb{R}^d$ , other choices are the Gaussian kernel with bandwidth  $h > 0$ ,  $k_h^G(x, y) = \exp(-\|x - y\|^2 / (2h^2))$  and the Laplace kernel with bandwidth  $h > 0$ ,  $k_h^L(x, y) = \exp(-\|x - y\| / (2h^2))$ ; see Section 6 for experimental results with such kernels.
- when  $\mathcal{X} = \{(p_1, \dots, p_d) \in [0, 1]^d \text{ such that } p_1 + \dots + p_d = 1\}$  the set of  $d$ -dimensional histograms, the intersection kernel is  $k(p, q) = \sum_{i=1}^d \min(p_i, q_i)$  (Hein and Bousquet, 2004; Maji et al., 2008); see Section 6 for experimental results with such a kernel.

## 6. Simulation experiments

We now present experimental results on the performance of our approach, respectively on synthetic data and on real data.

### 6.1 Synthetic data

First, we study the statistical behaviour of our approach for estimating the change-point locations of synthetic time series with  $\mathcal{X} = \mathbb{R}$  and  $n = 1000$ . The 9 change-point locations are fixed and chosen so the segments have various lengths, see the middle part of Figure 1. The intra-segment distributions are chosen randomly among the first ten probability distributions considered by Marron and Wand (1992) with common mean and variance. Since they only differ by their higher-order moments, standard approaches aiming at detecting changes in the mean or in variance would fail in such a situation.

We take the Gaussian RBF kernel  $k(x, y) = \exp(-(x - y)^2 / (2h^2))$  with  $2h^2$  among 0.1, 1 and  $\text{median}_{1 \leq i, j \leq n} \{\|X_i - X_j\|^2\}$ , the latter being a *classical heuristic* in kernel-based methods. We use the strategy presented in Section 5 and estimate  $v_{\max}$  with  $\hat{v}_{\max} := \max\{\text{tr}(\hat{\Sigma}_{0:\underline{t}}), \text{tr}(\hat{\Sigma}_{\bar{t}:1})\}$  where  $\underline{t} = 0.05$ ,  $\bar{t} = 0.95$  and  $t_i = i/n$  for all  $i$ . In preliminary experiments, we tested other strategies such as kernel-based counterparts of estimates of the maximal intra-segment variance using over-segmentation/under-segmentation or the so-called slope heuristic (Birgé and Massart, 2007). Comparing the average quadratic risks over 50 replications, the heuristic choice of the bandwidth clearly leads to the best performance (Table 2 in Appendix). Yet, fixed values of the bandwidth still lead to satisfactory results. A more detailed account of the performance of our algorithm is given in Figure 1, where the bandwidth is chosen with the classical heuristic. The left part of Figure 1 shows our criterion is minimal (in expectation) for

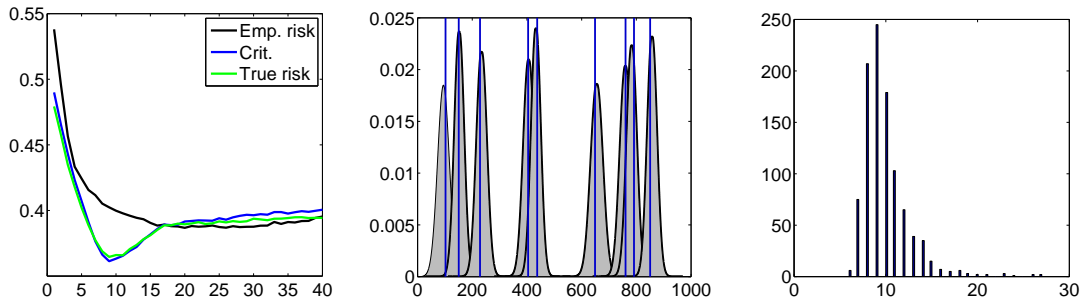


Figure 1: Synthetic data. Left: Expectations of the model selection criterion, empirical risk, and quadratic risk as a function of the number of candidate change-points. Middle: Pictorial representation of the frequency of detection of a change-point at each position; blue lines correspond to the true change-points. Right: Distribution of the estimated number of change-points  $\hat{D} - 1$ .

	Synthetic	Audio	Video
$\max_{\hat{t}_i \in \{\hat{t}_1, \dots, \hat{t}_{D-1}\}} \min_{t^* \in \{t_1^*, \dots, t_{D^*-1}^*\}}  \hat{t}_i - t^* $	$0.049 \pm 0.003$	$0.061 \pm 0.005$	$0.081 \pm 0.007$
$\max_{t^* \in \{t_1^*, \dots, t_{D^*-1}^*\}} \min_{\hat{t}_i \in \{\hat{t}_1, \dots, \hat{t}_{D-1}\}}  \hat{t}_i - t^* $	$0.053 \pm 0.006$	$0.079 \pm 0.006$	$0.093 \pm 0.007$

Table 1: Average Hausdorff distances between the estimated and true segmentation in the three experiments.

the same number of change-points as the quadratic risk, which equals the true number of change-points. On Figures 1–2, one can notice the empirical risk increases for large  $D$ . This phenomenon is due to the fact that we use heuristic rules for making the computation of  $\hat{m}_D$  faster, that do not always give the exact minimizer of the empirical risk for large values of  $D$ . Nevertheless, our algorithm is still accurate enough around the true and selected number of change-points. The right part of Figure 1 confirms the estimated number of breakpoints  $\hat{D} - 1$  is distributed around their true number. The middle part of Figure 1 represents the frequency of detection of a change-point at each location; for representation purposes, we fitted a mixture of gaussians centered around the true change-points, so their standard-deviations represent the accuracy of estimation of each true change-point. In particular, we observe the change-points are rather accurately detected, and that shorter segments are harder to detect accurately.

Table 1 provides results on the accuracy in estimating the change-point location in terms of Hausdorff distance between the set of estimated change-points  $\{\hat{t}_1, \dots, \hat{t}_{D-1}\}$  and the set of true change-points  $\{t_1^*, \dots, t_{D^*-1}^*\}$ , a common distance measure in the literature (Boysen et al., 2009).

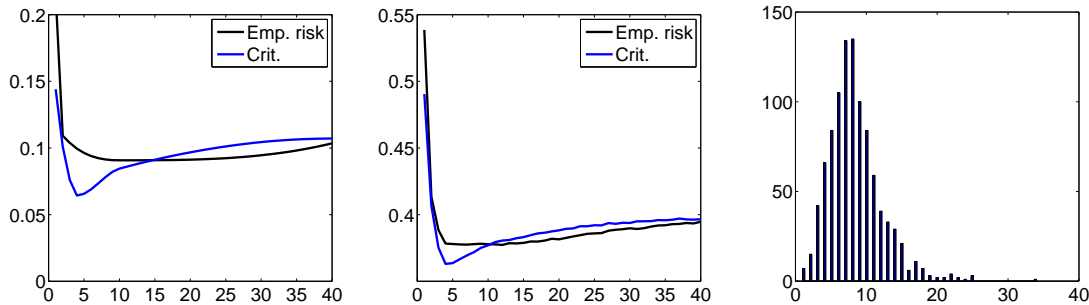


Figure 2: Real data experiment. Left and middle: our criterion and the empirical risk as a function of  $D$ , for one particular chunk (left: audio stream; middle: video stream). Right: Distribution of the estimated number of change-points  $\hat{D} - 1$  (video).

## 6.2 Real data: audio and video temporal segmentation

We now tackle the problem of temporal segmentation of the audio (resp. video) stream of entertainment TV shows into semantically homogeneous segments: trailer, audience applause, interview, music performance, and so on. We considered 50 chunks of audio (resp. video) streams delimited with two annotated changes at the border of this chunk. For each chunk, the true number of segments (given by manual annotation of semantically homogeneous parts of the TV show) is 5, and our goal is to recover these segments automatically, without knowing their number. Each chunk’s length is at most 30 minutes of the TV show and of 20 minutes on average.

**Audio part** We extracted every 10 ms the first 12 Mel Frequency Cepstral Coefficients (MFCC) of the audio track (Rabiner and Schäfer, 2007). MFCCs are commonly used features in speech recognition and audio processing. They provide a representation of the short-term power spectrum of a sound. We subsampled the signal when necessary to reduce the computing time of the dynamic programming part of our method. We used the Gaussian RBF kernel with a bandwidth automatically set using the classical heuristic rule as in Section 6.1. We present the performance of our approach on one particular audio chunk in Figure 2 (left). On this example, our approach selects the correct number of change-points in the time series on average (see Figure 3 in Appendix D) with a good accuracy (Table 1).

**Video part** We extracted 1024-dimensional GIST descriptors for each frame of the video track (Oliva and Torralba, 2001). GIST descriptors aggregate perceptual dimensions (naturalness, openness, roughness, expansion, ruggedness) that represent the dominant spatial structure of a scene. Again, we subsampled the signal when necessary to reduce the computing time. We used the so-called intersection kernel (Hein and Bousquet, 2004; Maji et al., 2008), which is appropriate for data belonging to  $d$ -dimensional simplices such as histograms-like GIST descriptors. Note that an attractive feature of the intersection kernel is that there is no hyperparameter (bandwidth) to tune. We present the performance of our approach on a particular video chunk in Figure 2 (middle and right). Here, the

good performance of our approach is less clear, as the average of the selected dimensions by our approach is 8.85 instead of 5. There are two explanations: (i) our estimate of  $v_{\max}$  is too rough, and over-segmentation is favored in the subsequent criterion, (ii) the GIST descriptors are too loose descriptors for this task.

## 7. Conclusion

We have proposed a penalty generalizing the one of [Lebarbier \(2005\)](#) to the kernel change-point problem, and showed it satisfies a non-asymptotic oracle inequality. Such an extension significantly broadens the possible applications of this penalization approach to the change-point problem. The theoretical tools developed for our method could also be used in other settings, such as clustering in general Hilbert spaces. As a future direction, we would like to investigate the kernel selection problem, which remains a major issue as in most machine learning problems (see the discussion of Section [4.3](#)).

## Acknowledgments

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

## References

- Sylvain Arlot and Alain Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Stat. Comput.*, 21(4):613–632, 2011.
- Shai Ben-David, Ulrike von Luxburg, and Dávid Pál. A sober look at clustering stability. In *COLT*, 2006.
- Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001.
- Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2):33–73, 2007.
- K. Bleakley and J.-Ph. Vert. The group fused lasso for multiple change-point detection. Technical report, ArXiv, 2011. arXiv:1106.4199.
- Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Math. Acad. Sci. Paris*, 334(6):495–500, 2002.
- Leif Boysen, Angela Kempe, Volkmar Liescher, Axel Munk, and Olaf Wittich. Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Stat.*, 37(1):157–183, 2009.
- Edward Carlstein, Hans-Georg Müller, and David Siegmund, editors. *Change-point problems*. IMS Lect. Notes, 1994.

- Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.
- Zaïd Harchaoui and O. Cappé. Retrospective change-point estimation with kernels. In *IEEE Workshop on Statistical Signal Processing*, 2007.
- M. Hein and O. Bousquet. Hilbertian metrics and positive-definite kernels on probability measures. In *AISTATS*, 2004.
- Steven M. Kay. *Fundamentals of statistical signal processing: detection theory*. Prentice-Hall, Inc., 1993.
- Alexander Korostelev and Olga Korosteleva. *Mathematical statistics. Asymptotic minimax theory*. Graduate Studies in Mathematics 119. American Mathematical Society (AMS), 2011.
- Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85(8):1501–1510, 2005.
- 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. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proces.*, 85:717–736, 2005.
- Subhransu Maji, Alexander C. Berg, and Jitendra Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.
- Colin L. Mallows. Some comments on  $C_p$ . *Technometrics*, 15:661–675, 1973.
- J.S. Marron and M.P. Wand. Exact mean integrated squared error. *Ann. Stat.*, 20(2):712–736, 1992.
- Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. ISBN 978-3-540-48497-4; 3-540-48497-3. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.
- Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42:145–175, May 2001.
- F. Picard, E. Lebarbier, M. Hoebeke, G. Rigaille, B. Thiam, and S. Robin. Joint segmentation, calling, and normalization of multiple cgh profiles. *Biostatistics*, 12(3):413 – 428, 2011. doi: 10.1093/biostatistics/kxq076.
- 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.
- Bernhard Schölkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.
- John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge Univ. Press, 2004.
- B. K. Sriperumbudur, K. Fukumizu, and G. R. Lanckriet. On the relation between universality, characteristic kernels and RKHS embedding of measures. In *Proc. of 13th International Conference on Artificial Intelligence and Statistics*, 2010.
- Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert R. G. Lanckriet, and Bernhard Schölkopf. Injective Hilbert space embeddings of probability measures. In *COLT*, 2008.
- I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Verlag, 2008.
- Vladimir Vapnik. *Statistical learning theory*. Wiley, 1998.
- Ulrike von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2(3), 2009.
- Y. Yao. Estimating the number of change-points via schwarz criterion. *Statistics and Probability Letters*, 6:181–189, 1988.
- Y.C. Yao and S. T. Au. Least-squares estimation of a step function. *Sankhya: The Indian Journal of Statistics*, 1989.



## Appendix A. Oracle inequality with a small collection of segmentations

Let us state a result which slightly differs from the primary goal of the paper (Theorem 1) but is a byproduct and can be of independent interest. Assume a subset  $\mathcal{M}'_n$  of the set  $\mathcal{M}_n$  of all segmentations of  $\{1, \dots, n\}$  is given such that

$$\exists \alpha_{\mathcal{M}'_n} > 0, \quad \text{Card}(\mathcal{M}'_n) \leq n^{\alpha_{\mathcal{M}'_n}} . \quad (\mathbf{Pol})$$

In particular, this setting corresponds to the situation where some prior knowledge restricts possible change-point locations to a subset of  $\{t_1, \dots, t_n\}$  with  $\mathcal{O}(\log n)$  elements. Let us now consider the model selection procedure defined by

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}'_n} \left\{ \frac{1}{n} \|\hat{\mu}_m - Y\|^2 + \operatorname{pen}(m) \right\} . \quad (17)$$

Then, Mallows' heuristics (Mallows, 1973) states that  $\operatorname{pen}(m) \approx \mathbb{E}[\operatorname{pen}_{\text{id}}(m)]$  leads to an oracle inequality. Making this informal argument rigorous, we obtain the following theorem, where assumption (Vmin) is replaced by a weakest assumption:

$$\exists 0 < c_{\min} < +\infty, \forall m \in \mathcal{M}'_n, \forall \lambda \in m, \quad v_\lambda := \frac{1}{\operatorname{Card}(\lambda)} \sum_{i \in \lambda} v_i \geq \frac{M^2}{c_{\min}} = v_{\min} . \quad (\mathbf{Vmin}')$$

**Theorem 4** *If (Db), (Vmin'), (Vmax) hold true and if  $\hat{m}$  satisfies Eq. (17) with*

$$\forall m \in \mathcal{M}'_n, \quad \operatorname{pen}(m) \geq \frac{2}{n} \sum_{\lambda \in m} v_\lambda , \quad (18)$$

*then, for every  $x \geq 0$ , an event of probability at least  $1 - e^{-x}$  exists on which, for every  $\theta \in (0, 1/8)$ ,*

$$\begin{aligned} \mathcal{R}(\hat{\mu}_{\hat{m}}) &\leq \frac{1}{1 - 4\theta} \inf_{m \in \mathcal{M}'_n} \left\{ (1 + 4\theta) \mathcal{R}(\hat{\mu}_m) + \operatorname{pen}(m) - \frac{2}{n} \sum_{\lambda \in m} v_\lambda \right\} \\ &\quad + [x + \log(4 \operatorname{Card}(\mathcal{M}'_n))] \frac{426 c_{\min}^2 v_{\max}}{n\theta} . \end{aligned}$$

Theorem 4 is proved in Section B.6. Note that Theorem 4 holds for every  $\mathcal{M}'_n$  (even  $\mathcal{M}'_n = \mathcal{M}_n$ ), but the remainder term is only reasonably small if assumption (Pol) holds true.

Since (Vmax) implies  $2 \sum_{\lambda \in m} v_\lambda \leq 2D_m v_{\max}$ , we get a formula for the penalty if an upper bound on  $v_{\max}$  is known or can be estimated. The corresponding procedure satisfies the following oracle inequality.

**Corollary 5** *In the framework of Theorem 4, let us assume some constant  $A > 0$  exists such that*

$$\operatorname{pen}(m) = \frac{2D_m A}{n} \geq \frac{2D_m v_{\max}}{n} . \quad (19)$$

Then, for every  $x \geq 0$ , an event of probability at least  $1 - e^{-x}$  exists on which, for every  $\theta \in (0, 1/8)$ ,

$$\begin{aligned} \mathcal{R}(\widehat{\mu}_{\widehat{m}}) &\leq \frac{A}{v_{\min}} \left( 1 + \frac{10}{\log(n)} \right) \inf_{m \in \mathcal{M}'_n} \{ \mathcal{R}(\widehat{\mu}_m) \} \\ &\quad + 426Ac_{\min}^3 \frac{\log(n)(x + \log(4\text{Card}(\mathcal{M}'_n)))}{n} . \end{aligned} \quad (20)$$

Corollary 5 is proved in Section B.7. If **(Db)** holds true for some known constant  $M$  (for instance,  $M = 1$  with the Gaussian and Laplace kernels), one can take  $A = M^2 \geq v_{\max}$  in the penalty (19).

If  $A = v_{\max}$ , one recovers the leading constant  $v_{\max}/v_{\min}$  in front of the oracle inequality, which is the price for ignoring the variations of noise along the signal. In particular, when

$$\forall 1 \leq i \leq n, \quad v_i = v_{\max} > 0 , \quad (\mathbf{Vc})$$

**(Vmin')** holds true with  $v_{\min} = v_{\max}$  and the leading constant in the oracle inequality (20) is one at first order. If assumption **(Pol)** holds true, the remainder term is of order at most  $(\log(n))^2/n$  so that (20) is an “optimal” oracle inequality similar to the one proved when  $\mathcal{H} = \mathbb{R}$  by [Birgé and Massart \(2007\)](#) in the Gaussian regression setting.

The reason why penalties in Eq. (6) and (19) are different is that Eq. (19) only yields a good penalty when **(Pol)** holds true, so not for change-point detection as in Theorem 1. Indeed, Eq. (5) holds for  $\text{pen}(m) \approx \mathbb{E}[\text{pen}_{\text{id}}(m)]$  when the collection of models is “small” (that is, if **(Pol)** holds true), but not with a collection as large as  $\mathcal{M}_n$ . Eq. (6) shows which additional terms are necessary to get Eq. (5) with a “large” collection of models like  $\mathcal{M}_n$ .

## Appendix B. Proofs

This section gathers the proofs of all results stated previously in the paper.

### B.1 Proof of the statements of Section 4.4.1

**Proof of Eq. (8)** Let  $f \in S_m$ . For every  $\lambda \in m$ , let us define  $f_\lambda$  as the common value of  $(f_i)_{i \in \lambda}$ , and

$$\bar{g}_\lambda := \frac{1}{\text{Card}(\lambda)} \sum_{i \in \lambda} g_i .$$

Then,

$$\begin{aligned} \|f - g\|^2 &= \sum_{\lambda \in m} \sum_{i \in \lambda} \|f_\lambda - g_i\|_{\mathcal{H}}^2 \\ &= \sum_{\lambda \in m} \sum_{i \in \lambda} \left[ \|f_\lambda - \bar{g}_\lambda\|_{\mathcal{H}}^2 + \|g_i - \bar{g}_\lambda\|_{\mathcal{H}}^2 + 2 \langle f_\lambda - \bar{g}_\lambda, \bar{g}_\lambda - g_i \rangle_{\mathcal{H}} \right] \\ &= \sum_{\lambda \in m} \left[ \text{Card}(\lambda) \|f_\lambda - \bar{g}_\lambda\|_{\mathcal{H}}^2 \right] + \sum_{\lambda \in m} \sum_{i \in \lambda} \|g_i - \bar{g}_\lambda\|_{\mathcal{H}}^2 + 2 \sum_{\lambda \in m} \left\langle f_\lambda - \bar{g}_\lambda, \sum_{i \in \lambda} (\bar{g}_\lambda - g_i) \right\rangle_{\mathcal{H}} \\ &= \sum_{\lambda \in m} \left[ \text{Card}(\lambda) \|f_\lambda - \bar{g}_\lambda\|_{\mathcal{H}}^2 \right] + \sum_{\lambda \in m} \sum_{i \in \lambda} \|g_i - \bar{g}_\lambda\|_{\mathcal{H}}^2 . \end{aligned}$$

Then,  $\|f - g\|^2$  is minimal if and only if  $f_\lambda = \bar{g}_\lambda$  for every  $\lambda \in m$ . ■

For proving Eq. (9), we compute the empirical and quadratic risks of  $\hat{\mu}_m$ :

$$\frac{1}{n} \|Y - \hat{\mu}_m\|^2 = \frac{1}{n} \|\mu^* - \mu_m^*\|^2 + \frac{1}{n} \|\varepsilon\|^2 - \frac{1}{n} \|\Pi_m \varepsilon\|_{\mathcal{H}}^2 + \frac{2}{n} \langle (I - \Pi_m) \mu^*, \varepsilon \rangle \quad (21)$$

$$\frac{1}{n} \|\mu^* - \hat{\mu}_m\|^2 = \frac{1}{n} \|\mu^* - \mu_m^*\|^2 + \frac{1}{n} \|\Pi_m \varepsilon\|^2 \quad (22)$$

The term  $n^{-1} \|\mu^* - \mu_m^*\|^2$  is called approximation error, or bias.

**Proof of Eq. (21)**

$$\begin{aligned} \|Y - \hat{\mu}_m\|^2 &= \|Y - \Pi_m Y\|^2 \\ &= \|\mu^* - \Pi_m \mu^*\|^2 + \|\varepsilon - \Pi_m \varepsilon\|^2 + 2 \langle \mu^* - \Pi_m \mu^*, \varepsilon - \Pi_m \varepsilon \rangle \\ &= \|\mu^* - \mu_m^*\|^2 + \|\varepsilon\|^2 - \|\Pi_m \varepsilon\|^2 + 2 \langle (I - \Pi_m) \mu^*, \varepsilon \rangle \end{aligned}$$

since  $\Pi_m$  is an orthogonal projection. ■

**Proof of Eq. (22)**

$$\begin{aligned} \|\mu^* - \hat{\mu}_m\|^2 &= \|\mu^* - \mu_m^*\|^2 + 2 \langle \mu^* - \mu_m^*, \Pi_m \varepsilon \rangle + \|\Pi_m \varepsilon\|^2 \\ &= \|\mu^* - \mu_m^*\|^2 + \|\Pi_m \varepsilon\|^2 \end{aligned}$$

since  $\Pi_m$  is an orthogonal projection. ■

**Proof of Eq. (9)** Eq. (9) follows from Eq. (21)–(22) and from the definition (5) of the ideal penalty. ■

For proving Eq. (10), we will use that

$$\forall i, j \in \{1, \dots, n\}, \quad \mathbb{E} [\langle \varepsilon_i, \varepsilon_j \rangle_{\mathcal{H}}] = v_i \mathbf{1}_{i=j} = \mathbf{1}_{i=j} \left( \mathbb{E} [k(X_i, X_i)] - \|\mu_i^*\|_{\mathcal{H}}^2 \right). \quad (23)$$

**Proof of Eq. (23)** For every  $i, j \in \{1, \dots, n\}$ ,

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

**Proof of Eq. (10)** The first equality comes from the fact that  $\mathbb{E} [\langle f, \varepsilon \rangle] = 0$  for every (deterministic)  $f \in \mathcal{H}^n$ , by definition of  $\varepsilon = Y - \mu^*$ . For the second equality, Eq. (8) implies

$$\begin{aligned} \|\Pi_m \varepsilon\|^2 &= \sum_{\lambda \in m} \left[ n_\lambda \left\| \frac{1}{n_\lambda} \sum_{i \in \lambda} \varepsilon_i \right\|_{\mathcal{H}}^2 \right] = \sum_{\lambda \in m} \left[ \frac{1}{n_\lambda} \left\| \sum_{i \in \lambda} \varepsilon_i \right\|_{\mathcal{H}}^2 \right] \\ &= \sum_{\lambda \in m} \left[ \frac{1}{n_\lambda} \sum_{i, j \in \lambda} \langle \varepsilon_i, \varepsilon_j \rangle_{\mathcal{H}} \right] \end{aligned} \quad (24)$$

where  $\forall \lambda \in m$ ,  $n_\lambda := \text{Card}(\lambda)$ . Now, using Eq. (23), we get

$$\mathbb{E} \left[ \|\Pi_m \varepsilon\|^2 \right] = \sum_{\lambda \in m} \left[ \frac{1}{n_\lambda} \sum_{i \in \lambda} v_i \right] = \sum_{\lambda \in m} v_\lambda .$$

■

Eq. (11) follows from Eq. (9)–(10).

## B.2 Proof of Proposition 2

Let us note

$$S_m = \langle \mu^* - \mu_m^*, \varepsilon \rangle = \sum_{i=1}^n Z_i \quad \text{with} \quad Z_i = \langle (\mu^* - \mu_m^*)_i, \varepsilon_i \rangle_{\mathcal{H}} .$$

The  $Z_i$  are independent and centered, so Eq. (26)–(27) in Lemma 6 below (which requires assumption (Db)) show the conditions of Bernstein's inequality are satisfied (see Theorem 9). Therefore, for every  $x \geq 0$ , with probability at least  $1 - 2e^{-x}$ ,

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

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

$$\forall i \in \{1, \dots, n\} \quad \|\mu_i^*\|_{\mathcal{H}} \leq M, \quad \|\varepsilon_i\|_{\mathcal{H}} \leq 2M \quad (25)$$

$$\text{and} \quad \|(\mu^* - \mu_m^*)_i\|_{\mathcal{H}} \leq 2M \quad \text{so that} \quad |Z_i| \leq 4M^2 . \quad (26)$$

$$\text{In addition,} \quad \sum_{i=1}^n \text{Var}(Z_i) \leq v_{\max} \|\mu^* - \mu_m^*\|^2 . \quad (27)$$

**Proof** [of Lemma 6] First, remark that for every  $i$ ,

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

so that with (Db),

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

which proves the first bound in Eq. (25). As a consequence, by the triangular inequality,

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

that is, the second inequality in Eq. (25) holds true.

Let us now define, for every  $i \in \{1, \dots, n\}$ ,  $\lambda(i)$  as the unique element of  $m$  such that  $i \in \lambda(i)$ . Then,

$$(\mu^* - \mu_m^*)_i = \frac{1}{\text{Card}(\lambda(i))} \sum_{j \in \lambda(i)} (\mu_i^* - \mu_j^*)$$

so that the triangular inequality and Eq. (25) imply

$$\|(\mu^* - \mu_m^*)_i\|_{\mathcal{H}} \leq \sup_{j \in \lambda(i)} \|\mu_i^* - \mu_j^*\|_{\mathcal{H}} \leq \sup_{1 \leq j, k \leq n} \|\mu_k^* - \mu_j^*\|_{\mathcal{H}} \leq 2 \sup_{1 \leq j \leq n} \|\mu_j^*\|_{\mathcal{H}} \leq 2M ,$$

that is, the first part of Eq. (26) holds true. The second part of Eq. (26) directly follows from Cauchy-Schwarz inequality. For proving Eq. (27), we remark that

$$\begin{aligned} \mathbb{E} [Z_i^2] &= \mathbb{E} \left[ \langle (\mu^* - \mu_m^*)_i, \varepsilon_i \rangle_{\mathcal{H}}^2 \right] \\ &\leq \|(\mu^* - \mu_m^*)_i\|_{\mathcal{H}}^2 \mathbb{E} \left[ \|\varepsilon_i\|_{\mathcal{H}}^2 \right] \quad \text{by Cauchy-Schwarz inequality} \\ &\leq \|(\mu^* - \mu_m^*)_i\|_{\mathcal{H}}^2 v_{\max} \quad \text{by (Vmax)} , \end{aligned}$$

$$\text{so that } \sum_{i=1}^n \text{Var}(Z_i) \leq v_{\max} \|\mu^* - \mu_m^*\|^2 .$$

■

### B.3 Proof of Proposition 3

This proof is inspired from [Sauvé \(2009\)](#), where a similar concentration inequality was needed for real-valued data, in the context of regression with piecewise polynomial estimators. As in our setting, Talagrand's inequality was not precise enough in the setting of [Sauvé \(2009\)](#).

Let us define

$$T_m := \|\Pi_m \varepsilon\|^2 = \sum_{\lambda \in m} T_\lambda \quad \text{with} \quad T_\lambda := \frac{1}{n_\lambda} \left\| \sum_{j \in \lambda} \varepsilon_j \right\|_{\mathcal{H}}^2 ,$$

according to Eq. (24). Now, remark that  $(T_\lambda)_{\lambda \in m}$  is a sequence of independent real-valued random variables, so we can get a concentration inequality for  $T_m$  via Bernstein's inequality, as long as  $T_\lambda$  satisfies some moment conditions (see Theorem 9). The rest of the proof will consist in showing such moment bounds, by using Pinelis-Sakhanenko deviation inequality (Proposition 10).

First, we showed in the proof of Eq. (10) that for every  $\lambda \in m$ ,  $\mathbb{E}[T_\lambda] = v_\lambda$ . Second, for every  $q \geq 2$ ,

$$\mathbb{E} [T_\lambda^q] = \frac{1}{n_\lambda^q} \mathbb{E} \left[ \left\| \sum_{k \in \lambda} \varepsilon_k \right\|_{\mathcal{H}}^{2q} \right] = \frac{1}{n_\lambda^q} \int_0^{2n_\lambda M} 2qx^{2q-1} \mathbb{P} \left[ \left\| \sum_{k \in \lambda} \varepsilon_k \right\|_{\mathcal{H}} \geq x \right] dx ,$$

since for every  $k$ ,  $\|\varepsilon_k\|_{\mathcal{H}} \leq 2M$  almost surely by Lemma 6, using (Db). Using again that  $\|\varepsilon_k\|_{\mathcal{H}} \leq 2M$  a.s., we get that for every  $p \geq 2$  and  $\lambda \in m$ ,

$$\sum_{k \in \lambda} \mathbb{E} [\|\varepsilon_k\|_{\mathcal{H}}^p] \leq (2M)^{p-2} \sum_{k \in \lambda} v_k \leq \frac{p!}{2} \left( \sum_{k \in \lambda} v_k \right) \left( \frac{2M}{3} \right)^{p-2},$$

that is, the assumption of Pinelis-Sakhanenko deviation inequality (see Proposition 10) holds true with  $c = 2M/3$  and  $\sigma^2 = \sum_{k \in \lambda} v_k$ . Therefore, using (Vmin), we get

$$\begin{aligned} \mathbb{E} [T_{\lambda}^q] &\leq \frac{1}{n_{\lambda}^q} \int_0^{2n_{\lambda}M} 2qx^{2q-1} 2 \exp \left[ -\frac{x^2}{2(n_{\lambda}v_{\lambda} + \frac{2Mx}{3})} \right] dx \\ &\leq \frac{4q}{n_{\lambda}^q} \int_0^{2n_{\lambda}M} x^{2q-1} \exp \left[ -\frac{x^2}{2n_{\lambda}v_{\lambda} \left(1 + \frac{4c_{\min}}{3}\right)} \right] dx \\ &\leq 2 \times (q!) \left[ 2v_{\lambda} \left(1 + \frac{4c_{\min}}{3}\right) \right]^q, \end{aligned}$$

since for every  $q \geq 1$ ,

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

Finally summing over  $\lambda \in m$ , it comes (using in particular that  $c_{\min} \geq 1$ )

$$\begin{aligned} \sum_{\lambda \in m} \mathbb{E} [T_{\lambda}^q] &\leq \frac{q!}{2} \times 4 \sum_{\lambda \in m} \left[ 2v_{\lambda} \left(1 + \frac{4c_{\min}}{3}\right) \right]^q \\ &\leq \frac{q!}{2} \times 4 \sum_{\lambda \in m} \left[ \frac{14c_{\min}v_{\lambda}}{3} \right]^q \\ &\leq \frac{q!}{2} \sum_{\lambda \in m} (87.5 c_{\min}^2 v_{\max} v_{\lambda}) [5c_{\min}v_{\max}]^{q-2}, \end{aligned}$$

that is, condition (35) of Bernstein's inequality holds with

$$v = 87.5 v_{\max} c_{\min}^2 \sum_{\lambda \in m} v_{\lambda} \quad \text{and} \quad c = 5c_{\min}v_{\max} .$$

Therefore, Bernstein inequality (see Theorem 9) shows that for every  $x > 0$ , with probability at least  $1 - 2e^{-x}$ ,

$$\begin{aligned} |T_m - \mathbb{E} [T_m]| &\leq \sqrt{175v_{\max}c_{\min}^2 \sum_{\lambda \in m} v_{\lambda}x + 5v_{\max}c_{\min}x} \\ &\leq \theta \sum_{\lambda} v_{\lambda} + \left( \frac{44c_{\min}^2}{\theta} + 5c_{\min} \right) v_{\max}x \\ &\leq \theta \sum_{\lambda} v_{\lambda} + \frac{49c_{\min}^2 v_{\max}x}{\theta} \end{aligned}$$

for every  $\theta \in (0, 1]$ , using also that  $c_{\min} \geq 1$ . ■

**Remark 7** *Let us emphasize that the classical approach for proving concentration results on  $\|\Pi_m \varepsilon\|$  when  $\varepsilon$  is bounded would not yield a result as precise as Proposition 3. Using for instance Talagrand's inequality (see [Bousquet, 2002](#)), we get*

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

and remark that for every  $f \in \mathcal{H}^n$ , the variables  $\langle f_i, (\Pi_m \varepsilon)_i \rangle_{\mathcal{H}}$  are real-valued, independent, centered and bounded. Then, instead of a main deviation term  $\theta v$  with  $v = \mathbb{E}[\|\Pi_m \varepsilon\|^2]$  as in Eq. (13), we would have  $v$  of order  $\sum_{i=1}^n \sup_{f \in \mathcal{H}^n, \|f\|=1} \mathbb{E}[\langle f_i, (\Pi_m \varepsilon)_i \rangle_{\mathcal{H}}^2]$  which is much larger than  $\mathbb{E}[\|\Pi_m \varepsilon\|^2] = \sup_{f \in \mathcal{H}^n, \|f\|=1} \sum_{i=1}^n \mathbb{E}[\langle f_i, (\Pi_m \varepsilon)_i \rangle_{\mathcal{H}}^2]$ . In the remainder of the proof, we do need a main deviation term  $\sqrt{2v}x$  with  $v$  proportional to  $\mathbb{E}[\|\Pi_m \varepsilon\|^2]$ , which is why we have to prove a result like Proposition 3.

#### B.4 Proof of a general model selection theorem

As sketched in Section 4.4, we first prove a general oracle inequality from which Theorems 1 and 4 are corollaries.

**Theorem 8** *Let  $\mathcal{M}'_n \subset \mathcal{M}_n$  and  $\hat{m}$  be some model selection procedure satisfying*

$$\hat{m} \in \operatorname{argmin}_{m \in \mathcal{M}'_n} \left\{ \frac{1}{n} \|\hat{\mu}_m - Y\|^2 + \operatorname{pen}(m) \right\}. \quad (28)$$

Assume that **(Db)**, **(Vmin)**, and **(Vmax)** hold true. Let  $(x_m)_{m \in \mathcal{M}'_n}$  be any collection of nonnegative numbers and assume that

$$\forall m \in \mathcal{M}_n, \quad \operatorname{pen}(m) \geq \frac{2}{n} \sum_{\lambda \in m} v_\lambda + r(x_m, \theta), \quad (29)$$

with  $r(x_m, \theta) := 213c_{\min}^2 v_{\max} x(\theta n)^{-1}$ . Then, an event  $\Omega_{(x_m)}$  exists such that  $\mathbb{P}(\Omega_{(x_m)}) \geq 1 - 4 \sum_{m \in \mathcal{M}'_n} e^{-x_m}$  and, on  $\Omega_{(x_m)}$ , for every  $\theta \in (0, 1/8)$ ,

$$\frac{1}{n} \|\mu^\star - \hat{\mu}_{\hat{m}}\|^2 \leq \frac{1}{1-4\theta} \inf_{m \in \mathcal{M}} \left\{ (1+4\theta) \frac{1}{n} \|\mu^\star - \hat{\mu}_m\|^2 + \operatorname{pen}(m) - \frac{2}{n} \sum_{\lambda \in m} v_\lambda + r(x_m, \theta) \right\}.$$

**Proof** [of Theorem 8] The first step is to combine Eq. (9), Eq. (11), Proposition 2 and Proposition 3. We get that for every  $x \geq 0$ , an event  $\Omega_m(x)$  of probability at least  $1 - 4e^{-x}$  exists on which, for every  $\theta > 0$ ,

$$\begin{aligned} \left| \operatorname{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 - \frac{2}{n} \sum_{\lambda} v_\lambda \right| &\leq \frac{2\theta}{n} \mathbb{E} \left[ \|\mu^\star - \hat{\mu}_m\|^2 \right] + \frac{98c_{\min}^2 v_{\max} x}{\theta n} + 2 \left( \frac{v_{\max}}{2\theta} + \frac{4M^2}{3} \right) \frac{x}{n} \\ &\leq \frac{2\theta}{n} \mathbb{E} \left[ \|\mu^\star - \hat{\mu}_m\|^2 \right] + \left[ \frac{98c_{\min}^2 + 1}{\theta} + \frac{8c_{\min}}{3} \right] \frac{v_{\max} x}{n}, \end{aligned}$$

where we used **(Vmin)**.

Using again Proposition 3 in combination with Eq. (22), we get that on  $\Omega_m(x)$ ,

$$\forall \theta \in (0, 1), \quad \mathbb{E} \left[ \|\mu^* - \hat{\mu}_m\|^2 \right] \leq (1 - \theta)^{-1} \left[ \|\mu^* - \hat{\mu}_m\|^2 + \theta^{-1} 49c_{\min}^2 v_{\max} x \right]. \quad (30)$$

Therefore, on  $\Omega_m(x)$ , for every  $\theta \in (0, 1/8)$ ,

$$\begin{aligned} & \left| \text{pen}_{\text{id}}(m) + \frac{1}{n} \|\varepsilon\|^2 - \frac{2}{n} \sum_{\lambda} v_{\lambda} \right| \quad (31) \\ & \leq \frac{2\theta}{(1 - \theta)n} \|\mu^* - \hat{\mu}_m\|^2 + \left[ \frac{2(1 + (1 - \theta)^{-1}) 49c_{\min}^2 + 1}{\theta} + \frac{8c_{\min}}{3} \right] \frac{v_{\max} x}{n} \\ & \leq \frac{4\theta}{n} \|\mu^* - \hat{\mu}_m\|^2 + \left[ 210c_{\min}^2 + \frac{4c_{\min}}{3} + 1 \right] \frac{v_{\max} x}{\theta n} \\ & = \frac{4\theta}{n} \|\mu^* - \hat{\mu}_m\|^2 + r_0(x, \theta), \quad (32) \end{aligned}$$

where

$$r_0(x, \theta) := \left[ 210c_{\min}^2 + \frac{4c_{\min}}{3} + 1 \right] \frac{v_{\max} x}{\theta n} \leq \frac{213c_{\min}^2 v_{\max} x}{\theta n} = r(x, \theta).$$

Then, let  $\Omega_{(x_m)} := \bigcap_{m \in \mathcal{M}_n} \Omega_m(x_m)$ . By the union bound,  $\mathbb{P}(\Omega) \geq 1 - 4 \sum_{m \in \mathcal{M}_n} e^{-x_m}$ . Now, by definition (28) of  $\hat{m}$ , for every  $m \in \mathcal{M}'_n$ ,

$$\frac{1}{n} \|\mu^* - \hat{\mu}_{\hat{m}}\|^2 + [\text{pen}(\hat{m}) - \text{pen}_{\text{id}}(\hat{m})] \leq \frac{1}{n} \|\mu^* - \hat{\mu}_m\|^2 + [\text{pen}(m) - \text{pen}_{\text{id}}(m)]. \quad (33)$$

Therefore, on  $\Omega_{(x_m)}$ , combining Eq. (32), (33) and the condition satisfied by  $\text{pen}(m)$ , we get the result for all  $\theta \in (0, 1/8)$ .  $\blacksquare$

## B.5 Proof of Theorem 1

We apply Theorem 8 with  $\mathcal{M}'_n = \mathcal{M}_n$  and  $x_m = D_m(\log(2) + 1 + \log(\frac{n}{D_m})) + \log 4 + x$ . Indeed, the probability of  $\Omega_{(x_m)}^c$  then is upper bounded by

$$\begin{aligned} 4 \sum_{m \in \mathcal{M}_n} e^{-x_m} &= \sum_{1 \leq D \leq n} \text{Card} \{m \in \mathcal{M}_n / D_m = D\} \exp \left[ -D \left( \log(2) + 1 + \log \left( \frac{n}{D} \right) \right) - x \right] \\ &= \sum_{1 \leq D \leq n} \binom{n-1}{D-1} \exp \left[ -D \left( \log(2) + 1 + \log \left( \frac{n}{D} \right) \right) - x \right] \\ &\leq e^{-x} \sum_{1 \leq D \leq n} \exp(-D \log(2)) \leq e^{-x} \sum_{D \geq 1} 2^{-D} = e^{-x}. \end{aligned}$$



Furthermore, we get for every  $\theta \in (0, 1/8)$  that

$$\begin{aligned} \frac{2}{n} \sum_{\lambda \in m} v_\lambda + r(x_m, \theta) &\leq \frac{2v_{\max} D_m}{n} + \frac{213c_{\min}^2 v_{\max} x_m}{\theta n} \\ &\leq \left[ 2 + \theta^{-1} 213c_{\min}^2 \left( \log(2) + 1 + \log\left(\frac{n}{D_m}\right) \right) \right] \frac{D_m v_{\max}}{n} \\ &\quad + 213c_{\min}^2 (\log 4 + x) \frac{v_{\max}}{\theta n} \\ &\leq \frac{v_{\max} D_m}{n} \left[ C_1 + C_2 \log\left(\frac{n}{D_m}\right) \right] + \frac{C_3}{n} \end{aligned}$$

with

$$\begin{aligned} C_1 &= C_1(\theta) = 361\theta^{-1} c_{\min}^2 \\ C_2 &= C_2(\theta) = 213\theta^{-1} c_{\min}^2 \\ C_3 &= C_3(x, \theta) = C_2(\theta) v_{\max} (\log 4 + x) . \end{aligned}$$

Note that  $C_3(x, \theta)$  is an additive term independent from  $m$ , so it can be safely removed from the penalty.

Finally, taking  $\theta = 1/12$  yields the result as long as  $C_1/c_{\min}^2$  and  $C_2/c_{\min}^2$  are larger than some numerical constant  $L_1 = 4332$ .  $\blacksquare$

## B.6 Proof of Theorem 4

First note that Theorem 8 does not rely on **(Vmin)** but only uses that **(Vmin')** holds true. Then, let us take  $x_m = x + \log(4\text{Card}(\mathcal{M}'_n))$  for every  $m \in \mathcal{M}'_n$  with  $x \geq 0$  in Theorem 8. First, we get

$$\mathbb{P}(\Omega_{(x_m)}) \geq 1 - 4 \left( \sum_{m \in \mathcal{M}'_n} e^{-\log(4\text{Card}(\mathcal{M}'_n))} \right) e^{-x} = 1 - e^{-x} .$$

Second, the condition (29) can be reduced to Eq. (18) since the term  $r(x, \theta)$  no longer depends on  $m$ . Therefore, it can be removed without changing the penalization procedure.  $\blacksquare$

## B.7 Proof of Corollary 5

We start from Theorem 4, denoting by  $\Omega$  the event on which the oracle inequality holds true.

First, assumption **(Vmax)** guarantees the penalty defined by Eq. (19) satisfies Eq. (18). Then, using assumption **(Vmin')**, we get

$$\begin{aligned} 2D_m A - 2 \sum_{\lambda \in m} v_\lambda &\leq 2D_m (A - v_{\min}) = \left( \frac{A}{v_{\min}} - 1 \right) 2D_m v_{\min} \\ &\leq \left( \frac{A}{v_{\min}} - 1 \right) 2 \sum_{\lambda \in m} v_\lambda . \end{aligned}$$

Therefore, on  $\Omega$ , since Eq. (30) holds true with  $x$  replaced by  $x + \log(4\text{Card}(\mathcal{M}'_n))$ , we get

$$2D_m A - 2 \sum_{\lambda \in m} v_\lambda \leq (1 - \theta)^{-1} \left( \frac{A}{v_{\min}} - 1 \right) \times \left[ \|\mu^* - \hat{\mu}_m\|^2 + \theta^{-1} 49c_{\min}^2 v_{\max} (x + \log(4\text{Card}(\mathcal{M}'_n))) \right]. \quad (34)$$

So, on  $\Omega$ , for every  $\theta \in (0, 1/8)$ ,

$$\frac{1}{n} \|\mu^* - \hat{\mu}_{\hat{m}}\|^2 \leq \frac{1}{1 - 4\theta} \left( 1 + 4\theta + (1 - \theta)^{-1} \left( \frac{A}{v_{\min}} - 1 \right) \right) \inf_{m \in \mathcal{M}'_n} \left\{ \frac{1}{n} \|\mu^* - \hat{\mu}_m\|^2 \right\} + \frac{c_{\min}^2 v_{\max}}{n\theta} (x + \log(4\text{Card}(\mathcal{M}'_n))) \left[ 426 + \frac{49}{(1 - \theta)(1 - 4\theta)} \left( \frac{A}{v_{\min}} - 1 \right) \right].$$

We get the result by taking  $\theta = \theta_n = (\log(n))^{-1}$  since for  $n$  larger than some numerical constant,

$$\frac{1}{1 - 4\theta_n} \left( 1 + 4\theta_n + (1 - \theta_n)^{-1} \left( \frac{A}{v_{\min}} - 1 \right) \right) \leq \frac{A}{v_{\min}} \left( 1 + \frac{10}{\log(n)} \right) \left[ 426 + \frac{49}{(1 - \theta_n)(1 - 4\theta_n)} \left( \frac{A}{v_{\min}} - 1 \right) \right] \leq \max \left\{ 426, \frac{49}{(1 - \theta_n)(1 - 4\theta_n)} \right\} \frac{A}{v_{\min}} \leq \frac{426A}{v_{\min}}$$

and  $\frac{c_{\min}^2 v_{\max} A}{v_{\min}} = \frac{c_{\min}^3 v_{\max} A}{M^2} \leq A c_{\min}^3$

where we used **(Vmin)**,  $A \geq v_{\max} \geq v_{\min}$ , and  $v_{\max} \leq M^2$ . ■

## Appendix C. Some useful results

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

**Theorem 9 (Bernstein's inequality, see Proposition 2.9 in (Massart, 2007))**

Let  $X_1, \dots, X_n$  be independent real valued random variables. Assume there exist positive constants  $v$  and  $c$  satisfying for every  $k \geq 2$

$$\sum_{i=1}^n \mathbb{E} \left[ |X_i|^k \right] \leq \frac{k!}{2} v c^{k-2}. \quad (35)$$

Then for every  $x > 0$ ,

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

In particular, if for every  $i$ ,  $|X_i| \leq 3c$  almost surely, Eq. (35) holds true with  $v = \sum_{i=1}^n \text{Var}(X_i)$ .

**Proposition 10 (Pinelis and Sakhnenko (1986), Corollary 1)** Let  $X_1, \dots, X_n$  be  $n$  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  $p \geq 2$ ,

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

Then, for every  $x > 0$ ,

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

## Appendix D. Additional simulation results

This section gathers a some additional results concerning the experiments of Section 6.1.

Kernel bandwidth	Risk ratio
$h = 0.1$	$3.56 \pm 0.17$
$h = 1.0$	$3.06 \pm 0.15$
adaptive $h$	$1.61 \pm 0.15$

Table 2: Synthetic data. Risk ratio  $\mathbb{E}[\mathcal{R}(\hat{\mu}_{\hat{m}}) / \inf_{m \in \mathcal{M}_n} \{\mathcal{R}(\hat{\mu}_m)\}]$  for three bandwidth choices.

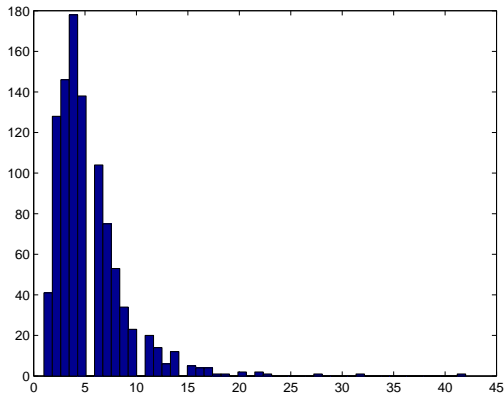


Figure 3: Real data experiment, audio stream: Distribution of the estimated number of change-points  $\hat{D} - 1$ .