Change-point detection and kernel methods

Damien Garreau

THÈSE DE DOCTORAT

de l'Université de recherche Paris Sciences Lettres
PSL Research University

Préparée à l'École Normale Supérieure

Change-point Detection and Kernel Methods
Détection de ruptures et méthodes à noyaux

École doctorale n°386

ÉCOLE DOCTORALE DE SCIENCES MATHÉMATIQUES DE PARIS CENTRE

Spécialité MATHÉMATIQUES

COMPOSITION DU JURY :

M. Stéphane ROBIN
INRA, Rapporteur

M. Bharath SRIPERUMBUDUR
Pennsylvania State University,
Rapporteur

M. Sylvain ARLOT
Université Paris Sud, Directeur de thèse

M. Gérard BIAU
Université Paris VI, Directeur de thèse

Mme. Céline LEVY-LEDUC
Agro ParisTech, Examinateur
Président du jury

M. Jean-Philippe VERT
ENS, Institut Curie, Mines ParisTech,
Examinateur

Soutenue par Damien Garreau
le 12 octobre 2017

Dirigée par **Sylvain Arlot et Gérard Biau**

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS

Je laisse Sisyphe au bas de la montagne ! On retrouve toujours son fardeau. Mais Sisyphe enseigne la fidélité supérieure qui nie les dieux et soulève les rochers. Lui aussi juge que tout est bien. Cet univers désormais sans maître ne lui paraît ni stérile ni futile. Chacun des grains de cette pierre, chaque éclat minéral de cette montagne pleine de nuit, à lui seul, forme un monde. La lutte elle-même vers les sommets suffit à remplir un cœur d'homme. Il faut imaginer Sisyphe heureux.

—Albert Camus, *Le mythe de Sisyphe*

To my family.

Acknowledgements

First and foremost, I want to thank my advisors. Sylvain, you have been a tremendous mentor during these years. I value more than I can express the scientific dialogue that we had, this thesis would not exist without your ideas. I also realize how much I have learned as a researcher thanks to your careful reading of my drafts and your constant quest for improvement in any result. From the bottom of my heart, thank you.

Gérard, while your involvement was less scientific, it was nonetheless critical. I am forever indebted to you for the confidence you entrusted me with and your political insights. I must add that, as everyone else, I am quite impressed — and maybe slightly worried! — by your superhuman average response time to any email.

I also want to thank members of the thesis committee, whose presence made the defense possible. Stéphane, Bharath, I would like to express my sincere gratitude for accepting to read my manuscript during summer time, I am sure that was not an easy task in spite of all my efforts. Your insightful comments were much appreciated. Céline, Jean-Philippe, I am incredibly honored that you accepted to be part of my jury, thank you.

Inria has been an incredible research environment throughout the years, mainly thanks to the incredibly talented people who work here. Francis, thank you for making me part of the Sierra team and for letting me stay after Sylvain left! I am also grateful to all the members and former members of the Willow-Sierra consortium for making these years so pleasant. Jean-Baptiste, Loic, Piotr, Nicolas, Rémy, Vincent, Christophe, Guilhem, Théophile, Alexandre, Vadim, Sesh, Rémi, Pierre, Pascal, Gauthier, Robert, Edouard, Yana, Maxime, Julia, Anastasia, Antoine, Vincent, Guillaume, Damien, Nino, Gül: discussing everyday with you was a pleasure, whatever the subject.

At this point I want to address very special thanks to the amazing people who shared an office with me during this PhD. Rémi, I miss our countless controversial debates. I sincerely hope that you did not record anything I said. Aymeric and Nicolas, the end of this PhD would have been unbearable without the mixture of hard work, sense of humor and optimism you brought to the office everyday. I feel extremely lucky to have spent so much time in your company.

Of course I would not have been able to complete this thesis without the unwavering support of my friends. Thibault, Sarah, Andrea, Fanny, Pierre-Yves, QBi, Chatoune, Rémi, Hatch, Raph, Julie, Tonio, it is always a pleasure to spend some time with you. Adrien, Baptiste, David, Florent, Guillaume, Laurent, Matthieu,

Nicolas, thank you for being there when I needed it. Michel, Züber, Marine, Jean-Noël, Marie-Christine, Antoine, Paul, it is a shame we do not see each other more often. Sophie, you are the best roommate one could dream of.

Apart from jury members, a number of people kindly proof-read part of this manuscript. Dominique, Antoine, Miro, Rémi and Andrea, I am incredibly grateful for your feedback, this manuscript would not be the same without you.

Thanks to Inria and DGA for providing the material support that made my PhD possible.

Let us not forget some essential contribution to this thesis: Laurent, Constance, Dominique, Elisabeth, Marion, Vincent, thank you for helping me to organize the "pot de thèse". Antoine, thank you for printing my manuscript.

Finally, I want to express my utmost gratitude to my family. Papa, Maman, Marion, merci pour votre soutien toutes ces années.

Contents

Contributions and thesis outline

This thesis is divided in 5 main chapters that can be read independently of the others to some extent. Chapter 1 is an autonomous introductory chapter, whereas Chapter 2 introduces notations and concepts needed for the understanding of Chapter 3. All experiments are gathered in Chapter 4, which can be read with only a superficial understanding of the algorithm and the theoretical results. Chapter 5 is a stand-alone chapter dedicated to the study of the median heuristic.

Chapter 1. In this first chapter, we introduce the change-point problem and kernel methods, which are the main topics of this manuscript. After a brief historical tour, we define the notion of consistency, which is our key concern in Chapter 3, and the kernel trick, which is the building block of kernel change-point detection.

Chapter 2. Next, we describe the kernel change-point detection procedure and explain the algorithm. We also present the framework under which we conduct our analysis of kernel change-point detection. Finally, we review some already known facts about kernel change-point detection.

Chapter 3. In Chapter 3, we state and prove our main results pertaining to kernel change-point detection. Namely, we show that kernel change-point detection has good theoretical properties for change-point estimation with independent data, under a boundedness assumption. We prove this result both for a linear penalty and a penalty function that originates from model selection. In the asymptotic setting, our result implies that kernel change-point detection estimates consistently all changes in the "kernel mean" of the distribution of data, at speed $\log(n)/n$ with respect to the sample size n. Since we make no assumptions on the minimal size of the true segments, this matches minimax lower bounds. The proof is based upon a concentration result for Hilbert-valued random variables. Under a weaker finite-variance assumption, we obtain some partial results. We also expose in much detail the different notions of distance between segmentations, and prove that they all coincide for sufficiently close segmentations.

This chapter is based upon the article Garreau and Arlot [2016], under submission to the *Electronic Journal of Statistics*.

Chapter 4. In this chapter, we first focus on practical issues associated to kernel change-point detection, namely the choice of the penalty constant when the penalty

function is linear and the choice of the kernel. We show how the dimension jump heuristic can be a reasonable choice for the penalty constant in simulations. We also compute a key quantity depending on the kernel that appears in our theoretical results, and show how this quantity is linked to the performance of KCP in practice. Some of the computations that we present are novel, up to the best of our knowledge. Finally, we demonstrate experimentally the consistency results proved in Chapter 3.

This chapter is based upon Garreau and Arlot [2016] and additional experiments.

Chapter 5. This final chapter is devoted to the median heuristic choice, a popular tool to set the bandwidth of radial basis function kernels. For large sample size, we show that the median heuristic behaves approximately as the median of a distribution that we describe completely in the setting of kernel two-sample test and kernel change-point detection. More precisely, we show that the median heuristic is asymptotically normal around this value. We illustrate these findings in a simple setting, where the underlying distributions are multivariate Gaussian.

This chapter is based upon Garreau [2017].

Notations

We recall here some notations used throughout the manuscript.

Abbreviations

e.g.	*exempli gratia*
Eq.	Equation
et al.	*et alii*
etc.	*et cetera*
Fig.	Figure
i.e.	*id est*
i.i.d.	independent and identically distributed
KCP	kernel change-point detection
MMD	Maximum mean discrepancy
p.	page
p.s.d.	positive semi-definite
Prop.	Proposition
resp.	respectively
RKHS	reproducing kernel Hilbert space
s.t.	such that

General mathematical notations

$:=$	define equals
$\lvert\cdot\rvert$	absolute value
$\lVert\cdot\rVert$	norm of a vector
$\lVert\cdot\rVert_{\mathcal{H}}$	norm of a vector in $\mathcal{H}$
$\lVert\cdot\rVert_{\mathrm{F}}$	Frobenius norm of a matrix
$\arg\max$, $\arg\min$	argument of the maxima (resp. minima)
$\mathbb{C}$	set of complex numbers
conv	convex hull of a set
diag	diagonal of a matrix
diam	diameter of a set
erf	error function, defined by $\operatorname{erf}(x) := \frac{2}{\sqrt{\pi}} \int_0^x \mathrm{e}^{-t^2} \,\mathrm{d}\, t$
erfc	complementary error function $\operatorname{erfc}(x) := 1 - \operatorname{erf}(x)$

$\mathbb{E}$	expectation of a random variable
$\mathcal{F}$	Fourier transform
$\mathcal{H}$	Hilbert space
I_d	identity matrix of size $d \times d$
$\mathbb{1}_{\cdot \in A}$	indicator function of the set A
Med	sample median
med	theoretical median
$\mathrm{o}(\cdot), \mathrm{O}(\cdot)$	Landau notations
$\mathbb{N}$	non-negative integers
$\mathbb{N}^\star$	positive integers
$\mathbb{P}$	probability of an event
$\mathbb{R}$	set of real numbers
$\langle \cdot, \cdot \rangle$	scalar product
$\langle \cdot, \cdot \rangle_{\mathcal{H}}$	scalar product in $\mathcal{H}$
Span	span of a set of vectors
Tr	trace of a matrix
$x^\top$	transpose of the matrix x
Var	variance of a random variable
$\mathcal{X}$	input space

Other notations

A_τ	linear part in the decomposition of ψ_τ
C	penalty constant
$C_{\min}, C_{\max}$	minimal and maximal theoretical penalty constant
crit	a model selection criterion
$\underline{\Delta}, \overline{\Delta}$	smallest and largest jump size in the RKHS
D_τ	number of segments of τ
$D^\star$	number of segments of $\tau^\star$
$\widehat{D}$	number of segments of $\widehat{\tau}$
d	generic distance
d_{F}	Frobenius distance between segmentations
d_{H}	Hausdorff distance between segmentations
$\mathrm{d}_\infty, \mathrm{d}_\infty^{(1)}, \mathrm{d}_\infty^{(2)}, \mathrm{d}_\infty^{(3)}$	custom distances, defined in Section 3.2.2
$\widehat{\mathcal{R}}$	empirical risk, defined by Eq. (2.1)
k	a positive semi-definite kernel
K	the Gram matrix, defined by Eq. (1.10)
$\underline{\Lambda}, \overline{\Lambda}$	minimum and maximum segment size
L_τ	linear part in the decomposition of ψ_τ
M	kernel bound in Assumption 2.1
$\mu^\star$	true regression function, element of $\mathcal{H}^n$
$\widehat{\mu}$	estimated regression function, element of $\mathcal{H}^n$
ν	(positive) bandwidth

pen	generic penalty function
pen_ℓ	linear penalty, defined by Eq. (2.3)
pen_L	alternative penalty, defined by Eq. (2.4)
ψ_τ	defined by Eq. (3.11)
Q_τ	quadratic part in the decomposition of ψ_τ
$\mathcal{R}$	quadratic risk, defined by Eq. (2.12)
$\mathcal{T}_n$	set of all the segmentations of $\{1, \dots, n\}$
H	squared median heuristic, defined by Eq. (5.2)
τ	a segmentation
$\widehat{\tau}$	estimated segmentation
$\tau^\star$	true segmentation
V	variance bound in Assumption 2.2
v_1	speed of convergence in Theorem 3.1
v_1'	speed of convergence in Theorem 3.2
v_2	speed of convergence in Theorem 3.3

Chapter 1

Introduction

A large part of this manuscript is devoted to questions in regard to kernel change-point detection, a method introduced by Arlot et al. [2012]. The main idea underlying this procedure is to adapt a penalized least-squares change-point detection scheme to data belonging to a general set on which a positive semi-definite kernel is defined. Concepts from two distinct branches of statistics meet here: change-point detection and kernel methods. In this introduction, we aim to present both, with the goal of relating our work to the existing literature. A precise description of kernel change-point detection is delayed to Chapter 2.

We begin with a description of the change-point problem in Section 1.1, setting some vocabulary extensively used throughout this manuscript. Before giving an overview of the literature pertaining to change-point detection in Section 1.3, we expose some real-world situations where this problem naturally arises in Section 1.2. We then briefly present the powerful machinery of kernel methods in Section 1.4.

1.1 The change-point problem

Change-point detection is a long-standing question in mathematical statistics, which has attracted a lot of attention since the 30s. Our goal in this section is not to present the vast literature associated with this problem in an exhaustive fashion, but rather to present the main ideas that shaped the field, and to relate them to this thesis.

In the study of time series, it is natural to assume *stationarity*, that is, time-shift invariance of the data probability distribution. Indeed, suppose that we are recording data coming from a natural phenomenon during a time period where there is no trend or shock in the background, then there is no reason for the distribution parameters to change. This assumption often fails in practice, where the environment does endure potentially large changes, and a more reasonable assumption is to consider that the observed phenomenon is stationary only on smaller time-units — see Fig. 1-1 for a typical situation. The goal of change-point detection is to recover these segments as accurately as possible.

If the signal we consider is multi-dimensional, we assume that changes in the

distribution occur *simultaneously* in the different dimensions of the signal. It is not necessary that changes take place in all of them, though it should be clear that the problem becomes quite challenging if only a few coordinates undergo a change in a high-dimensional setting. We will not deal with the situation where each dimension is segmented differently, that is, *joint segmentation*, and we refer to Picard et al. [2011] and references therein for an introduction.

Let us specify some vocabulary and concepts that are going to be used throughout this thesis.

Off-line *vs* on-line setting. When observations $X_1, \ldots, X_n$ are obtained one at a time, we say that the setting is *on-line*, or *sequential*. The goal is to detect changes as quickly as possible, while keeping the number of false alarms as low as possible. We talk about *off-line* or *a posteriori* detection of changes when the data is obtained all at once. In this case there is no need for real-time processing of the data, and one can take advantage of this additional computation time. Note that on-line procedures can be applied in the off-line setting, just by running through the data-points as if they were being observed one by one. Less obvious is the possibility to use an off-line procedure in the on-line setting, handling batches of newly observed data-points with the off-line procedure. However, the batches' size is then a strict lower bound to the quickest detection time. Let us emphasize that this thesis is essentially concerned with off-line change-point detection. Nevertheless, we briefly account for the main ideas and essential results pertaining to on-line change-point detection in Section 1.3.1.

Single *vs* multiple change-point detection. There is quite a difference between detecting a *single* change-point and *multiple* change-points in the off-line setting. It is important to understand that the second problem is much harder, since the number of possible outcomes for the procedure jumps from $n-1$ to 2^{n-1} if the number of change-points is unknown [1], where n is the number of observations. This thesis is mainly concerned with the second problem. Note that methods designed for detecting multiple change-points can obviously be applied to the detection of a single change-point. The converse is also possible, even though there is no guarantee it will yield the desired results. For instance, one can proceed as follows: (i) search the entire dataset for a change-point, (ii) if a change-point is found, separate the dataset in two parts (before and after), and (iii) iterate. This generic procedure is called *binary segmentation*, see Vostrikova [1981].

Parametric *vs* non-parametric. It is common to make *parametric* assumptions, *i.e.*, to assume that the distribution of the observations belongs to a family of distributions that can be described using a finite number of parameters. Thus the changes occur in one (or more) of the parameters describing the distribution. These assumptions are important both for the design and the analysis of change-point procedures.

1. There are $\binom{n-1}{k-1}$ ways for a signal of size n to be segmented in k segments, hence a total of $\sum_{k=1}^{n-1}\binom{n-1}{k-1} = 2^{n-1}$ possibilities.

We are more interested in *non-parametric* settings, where little is known about the underlying distributions.

In brief, we are interested in the off-line detection of multiple changes in potentially multi-dimensional time-series with simultaneous changes, without parametric assumptions.

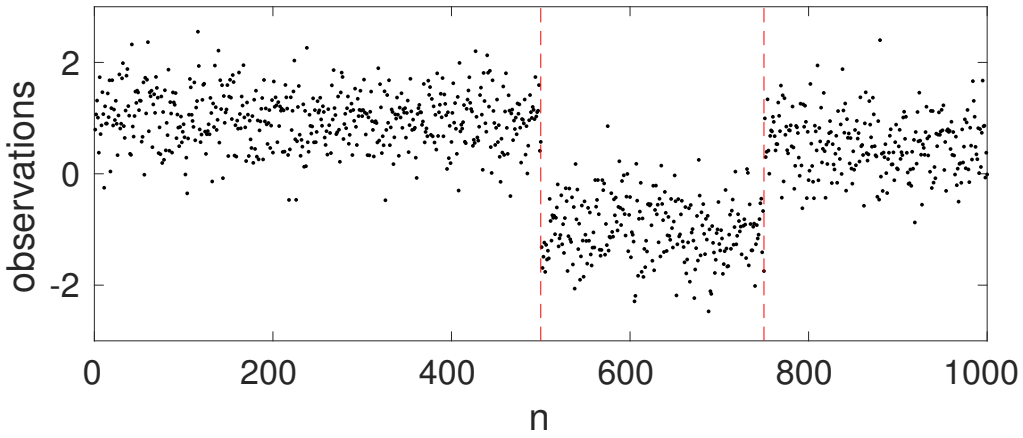

Figure 1-1 – Here we observe 10^3 observations of a Gaussian distribution (marked as black dots) whose mean abruptly changes from 1 to -1 and then to 0.5 at position 500 and 750 (indicated by broken vertical red lines) and whose variance stays constant ($\sigma = 0.5$). The goal of change-point detection is to recover these positions from the observations. We are here in an off-line setting, with multiple change-points, in a parametric model.

1.2 Examples

In this section we present some real situations where change-point problems arise. The historical example for the use of change-point detection comes from process quality control [Shewhart, 1931], that is, the monitoring of industrial processes which can be either *in control* or *out of control*. The goal is then the quickest detection of anomalous behaviors, with the least false-alarms (see Basseville and Nikiforov [1993, Chapter 1]). It is essentially an on-line problematic, that we do not develop here. As we have seen in Section 1.1, this thesis is focused on the off-line detection of changes, so we present in more details some examples coming from fields where off-line change-point detection is relevant.

Array Comparative Genomic Hybridization. During cellular division, the duplication of the genome can go astray and a large number of base pairs can be deleted or copied more than once. This phenomenon is called copy-number variation and occurs frequently in the life of a cancerous cell. Comparative Genomic Hybridization

(CGH) allows to estimate the copy-number variations in a particular genome with respect to normal. More precisely, for each locus of the genome with a precision of 5 to 10Mb,[2] CGH provides an estimate of the ratio between the copy-number of a test subject DNA and a reference. Note that before the development of CGH in the early 90s [Kallioniemi et al., 1992], it was extremely costly to obtain this global information.

For the sake of completeness, let us explain briefly how CGH data is obtained: the DNA of both samples is first extracted and colored with a fluorescent marker, generally green for the test sample and red for the reference. Both samples are then heated, which makes the DNA strands of the chromosomes separate. Next, they are dropped down on a microchip that contains ordered single DNA strands of the same genome as the reference. Locally, if there are more copies of the test DNA, it associates preferentially with the control DNA and green prevails — conversely, if there are less copies of the test DNA, red prevails. The final data is obtained by measuring the fluorescence intensities.

Array CGH (aCGH) is a technical refinement of CGH that allows to work at a much finer scale (5 to 10kb) and is now widely used instead of CGH. As CGH, it is not exempt from noise, which comes from experimental measurement imprecision and the log-ratio is generally assumed to be Gaussian. An example of aCGH data is depicted in Fig. 1-2.

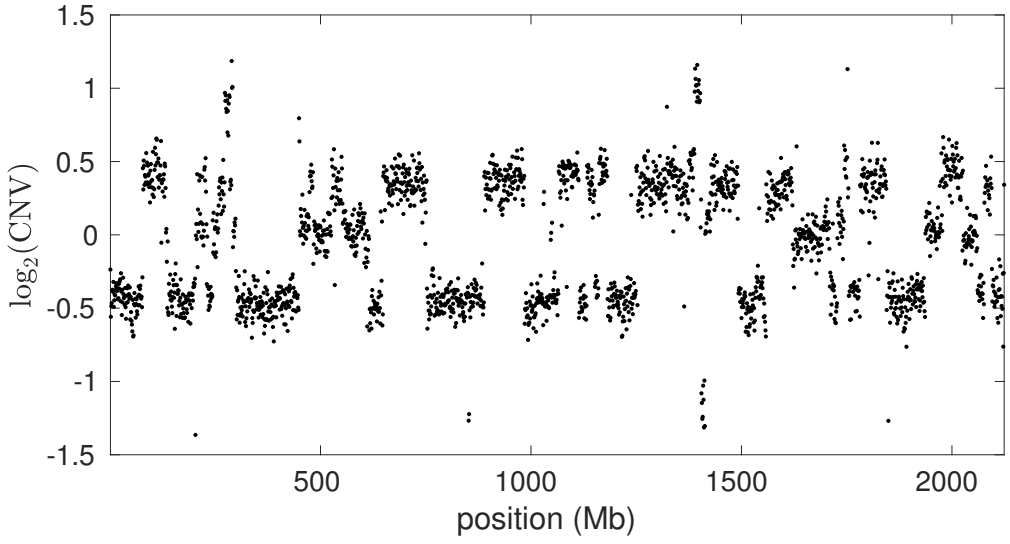

Figure 1-2 – Logarithm of the copy-number variation as a function of the position in the genome of cancer cell T47D [data from Snijders et al., 2001]. Positive (resp. negative) values correspond to genome positions where the copy number is higher (resp. lower) that normal. As can be seen on this example, the copy-number variation can be modeled well by a noisy piecewise constant function.

2. A base pair (bp) consists of an $A-T$ or $G-C$ pair; it is approximately 340pm long. A megabase (Mb) denotes 10^6 base pairs. The size of the Human genome is approximately $3,000$Mb.

Interestingly, several types of tumors show a consistent pattern of such genetic aberrations, *i.e.*, large connected portions of the genome are consistently over or under-replicated — in fact CGH was invented precisely to study the genetic anomalies of tumor cells. Hence, such patterns can serve as a signature of the tumor, and identifying precisely the genome segments on which the copy-number is consistently higher or lower than normal can be used to diagnose cancer [Bejjani and Shaffer, 2006]. When there are multiple tumors, it is also possible to use these signatures for identifying which tumor is metastasizing [Weiss et al., 2003]. These patterns are collections of geographic segments of the genome, and both during the identification and diagnosis stage, we are faced with an off-line multiple change-point detection problem.

Modeling of financial time series. Proposing relevant models for financial time series, in particular share prices, is a subject of the utmost importance both in quantitative finance and econometrics. For a given stock, portfolio, or stock market index, investors are primarily concerned with the *return*, that is, the net gain or loss generated by an investment strategy.

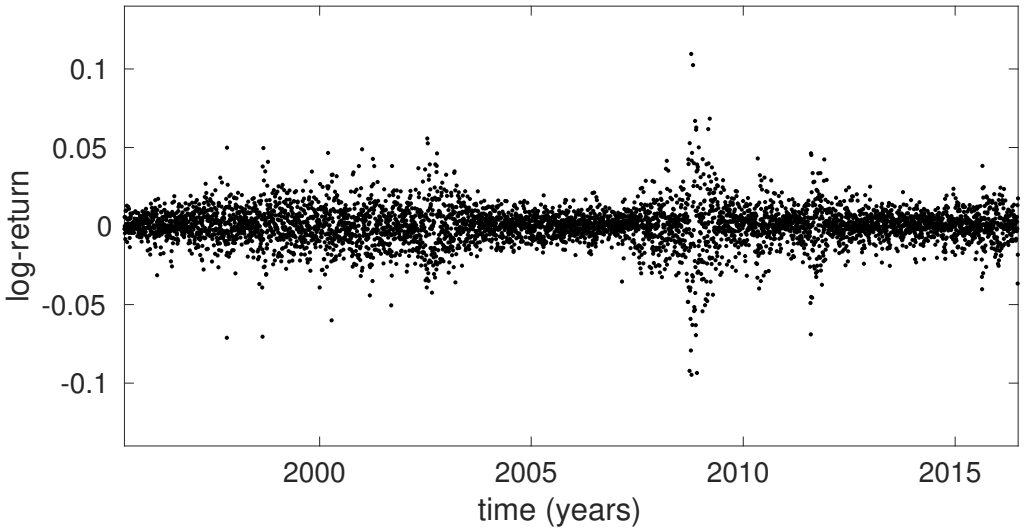

Figure 1-3 – Log-returns of the Standard and Poors 500 Price index, closing prices, from June 1, 1995 to June 1, 2017. The S&P500 is an index reflecting the market capitalization of the 500 largest listed companies in the USA. As one can see, the X_ts have a fairly stationary behavior between structural changes that often corresponds with major financial crises. The most visible in this graph is the 2008 financial crisis.

Let us focus on a specific financial asset, write s_t the value of this asset at time $t \in \mathbb{Z}$, and denote by X_t the logarithmic return of this asset, that is, $X_t := \log(s_{t+1}/s_t)$ (see Fig. 1-3 for a plot of log-returns on a real dataset). It is common to assume that X is a generalized autoregressive model (GARCH). There is tremendous literature on GARCH processes, we refer to Tsay [2005] for an introduction. It falls out of the

scope of this thesis to present this topic rigorously. To set ideas straight, let us just say that X can be decomposed as

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum_{i=1}^{q} \beta_i \sigma_{t-i}^2 \end{cases} \tag{1.1}$$

where the Z_ts are i.i.d. standard Gaussian random variables, p and q are positive integers, $\alpha_0 > 0$, and $(\alpha_i)_{1 \leq i \leq p}, (\beta_i)_{1 \leq i \leq q} \in \mathbb{R}_+$.

There is empirical evidence to show that the above model is satisfying for short periods of time. Furthermore, it is practical to estimate the parameters of model (1.1) when it is assumed that the observations constitute a stationary sequence of random variables (see Straumann [2005] for a comprehensive monograph on the estimation of GARCH models). But for longer time periods, this assumption is not realistic, in particular due to abrupt changes in the variance process during times of economic turmoil [Mikosch and Starica, 2004]. A simple way out is to consider that the observations are only stationary on short periods of time and to update the parameters estimates on each segment, hence the need for an off-line change-point detection procedure.

Let us mention that some applications in finance are also focused on the real-time problem, where the goal is to detect the apparition of structural breaks in "live" financial time series [Pepelyshev and Polunchenko, 2015].

Video processing. In a movie, a *shot* is the longest continuous sequence that originates from a single camera, and shot detection is the problem of detecting the beginning and ending of each shot. The goal is to produce small homogeneous movie parts, that can be easily used for indexing or more involved movie processing tasks.

Even though this information is sometimes available as meta-data in recent standards such as MPEG-7, there is still an ongoing research effort to develop reliable techniques for shot detection; we refer to Cotsaces et al. [2006] for a review. Once the movie transformed into a vector-valued time series, shot detection is essentially an off-line, non-parametric, multiple change-point detection problem. Nevertheless, let us emphasize two major differences with the setting that we introduced in Section 1.1: (i) there is no definitive feature extraction procedure for this task, and (ii) not all changes are abrupt — if a clear transition between shots (a *cut*) is most frequently employed, a gradual transition is also a possibility (a *dissolve*). An example of clear transitions between shots is pictured in Fig. 1-4.

An additional difficulty is the absence of prior information regarding the number of change-points. Indeed, if some movies contain hundreds of distinct shots, others have very few.[3] A relevant task in the latter case is scene detection, that is, segmenting the video according to the actions taking place. When appropriate features are used, this is once again an off-line multiple change-point problem [Allen et al., 2016].

In all these previous examples, change-point detection can be performed "by hand." We think that automatic change-point detection methods can help the practitioner

3. *Victoria* (2015), a 2 hours 18 minutes long movie, consists of a single continuous shot.

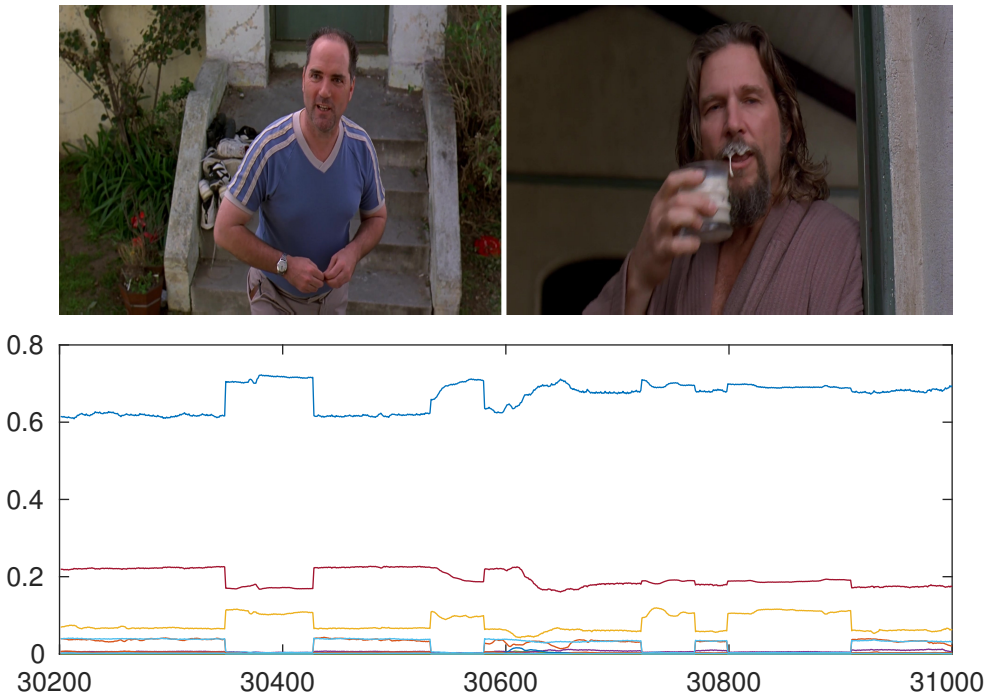

Figure 1-4 – An excerpt of the movie The Big Lebowski. Here we can see 800 frames — corresponding roughly to 30s — of the movie, and for each frame the corresponding color histogram (27 colors) in the bottom panel. In this scene, the Dude is talking to his landlord Marty while drinking a White Russian. The camera alternates between shots of the Dude on his doorway (upper right panel) and Marty in front of him (upper left panel). Abrupt changes in the color histogram of the frames match this alternation.

in many respects, as is the case with most statistical methods:

— to increase the reliability of the data analysis;

— to decrease the risk of unexplainable errors;

— to increase the speed of data processing by several orders of magnitude, allowing to handle much more data;

— to free some time for other tasks of more interest.

1.3 History

In this section, we present some important algorithms both for on-line and off-line change-point detection.

1.3.1 Sequential change-point detection

Historically, change-point problems have come from the sequential point of view. We present here some of the most important ideas in the on-line setting. This account is far from exhaustive. First, we omit Bayesian procedures, which are completely out of the scope of this manuscript. See Fearnhead and Liu [2007] and references therein for an introduction to this point of view on the on-line change-point problem. Second, we do not mention sequential testing procedures since we shall encounter them shortly hereafter in the context of off-line change-point detection. We refer to Basseville and Nikiforov [1993] for an extensive overview of on-line procedures and Tartakovsky et al. [2014] regarding sequential testing.

In the following, $X_1, \dots, X_n$ are sequential observations of a real-valued random process that undergoes a change in the mean and, for any $i \in \{1, \dots, n\}$, we define

$$\mu_i^\star := \mathbb{E}\left[X_i\right] .$$

Control charts. The most immediate idea for detecting changes in signal is certainly to set a threshold and to decide that there is a change if a certain statistic crosses the threshold. How does one choose the threshold value and the statistic? The first rigorous attempt to answer this question was made by Shewhart [1931]. Suppose for now that the true mean of the observations $\mu^\star$ is known, as well as the variance σ^2. Choose a batch size N and set

$$\bar{X}(K) = \frac{1}{N} \sum_{i=N(K-1)+1}^{NK} X_i \, .$$

Then the proposed algorithm is to detect a change as soon as

$$\left|\bar{X}(K) - \mu^\star\right| > \kappa \frac{\sigma}{\sqrt{N}} \, ,$$

where κ is a constant. Keep in mind that in the context of quality control in which control charts were introduced, $\mu^\star$ is a known value prescribed by the industrial necessities. Moreover, σ^2 can be easily estimated if it is unknown, for instance by setting

$$\widehat{\sigma}^2 := \frac{1}{m-1} \sum_{i=1}^{m} \left(X_i - \bar{X}\right)^2 \, ,$$

when there is a strong belief that no change-point occurred in the first m observations. Concentration of measure can help choosing κ depending on the hypothesis satisfied by the observations — for instance, a refinement of Chebyshev's inequality [Meidell, 1918; Camp, 1922] guarantees that, for unimodal distributions, $|X_i - \mu| > 3\sigma$ with probability smaller than 0.05 hence the $\kappa = 3$ chosen by Shewhart.

While numerous versions of Shewhart's control chart exist, they all share the same common idea: as the data is released, compute a statistic on each batch and take action if it crosses predefined thresholds [Page, 1954; Lai, 1974].

CUSUM. It was not before the 50s that optimal procedures were developed, after the seminal work of Wald [1947]. The breakthrough of the CUSUM algorithm [Page, 1954] is to take action not if the statistic crosses a threshold, but if it is far away from the historic minimum. In fact, this amounts to a control chart comparing against an *adaptive threshold*. Let us be more specific: as in the previous paragraph, we collect samples of size N. We then assign a score $Y(K)$ to the K-th sample, set $S_m = \sum_{K=1}^{m} Y(K)$, and take action if

$$S_m - \min_{0 \leq i < m} S_i \geq h \,, \tag{1.2}$$

where h is a positive constant. Even though it was not explicit in Page [1954] how to choose the scores $Y(K)$, it is common to use the log-likelihood ratio. That is, if we assume that the X_i are drawn according to a distribution p_θ with $\theta \in \Theta$ some parameter, equal to θ_0 before the break and θ_1 after, we set

$$Y(K) = \sum_{i=N(K-1)+1}^{NK} \log \frac{p_{\theta_1}(X_i)}{p_{\theta_0}(X_i)} \,.$$

Of course, Eq. (1.2) only allows to detect *positive* changes in the mean of the score. Using two CUSUM algorithms together, for instance, is a way to fix this problem [Page, 1954]. This line of work was later extended in a series of papers by Shiryaev [1961, 1963, 1965]. We refer to Lorden [1971]; Moustakides [1986] for optimality results regarding this algorithm.

Filtered derivative. The key idea behind the filtered derivative algorithm [Basseville, 1981] is simple: if there is no noise, then changes in the mean translate into sharp jumps in the absolute value of the discrete derivatives of the signal. More precisely, define the moving average

$$g_k = \sum_{i=0}^{N-1} \gamma_i \log \frac{p_{\theta_1}(X_{k-i})}{p_{\theta_0}(X_{k-i})} \,,$$

where the γ_i are positive weights. Then instead of deciding for a change whenever $g_k \geq h$ (a *finite moving average* control chart), we consider the discrete derivative $\nabla g_k := g_k - g_{k-1}$ and consider that there is a change-point if a sufficient number of discrete derivatives are above a threshold h, *i.e.*,

$$\sum_{i=0}^{N-1} \mathbb{1}_{\nabla g_{k-i} \geq h} \geq \eta \,.$$

The parameter η is typically small, e.g., $\eta = 2$. Similar ideas are used for edge-detection algorithm in image processing [Roberts, 1963].

We now turn to one of the main focuses of this manuscript, off-line change-point detection.

1.3.2 Off-line change-point detection

Even though on-line change-point methods predate off-line methods by more than fifteen years, the literature relating to the latter is no less spread-out. As in the on-line setting, we do not pretend to an exhaustive treatment of the off-line change-point literature and we refer to Brodsky and Darkhovsky [2013, Chapter 2] for a review of non-parametric off-line change-point detection methods. Rather, we want to present with a reasonable amount of detail two principles paramount to most of the off-line methods.

The first idea is to interpret the change-point detection problem as a statistical hypothesis test, deciding for a change-point if it is *statistically significant*. Another possibility is to cast the change-point problem as an estimation problem, where the change-points are parameters from the model that we wish to estimate. Among the estimation methods, we take a special interest in estimates obtained *via* the minimization of a least-squares criterion given that our object of interest, kernel change-point detection, is a natural extension of this line of thought.

Before presenting off-line methods in more detail, let us define a simplified *a posteriori* change-point setting. For any integer n, we call *segmentation* integers $1 \leq D \leq n$ and $\tau_0 := 0 < \tau_1 < \cdots < \tau_{D-1} < \tau_D = n$. We call *segments* the sets $\{1, \dots, \tau_1\}, \{\tau_1+1, \dots, \tau_2\}, \dots, \{\tau_{D-1}+1, \dots, n\}$. Given observations $X_1, \dots, X_n$, there is always a segmentation $\tau^\star$ with $D^\star$ segments such that the distribution of X_i is constant on the segments, but distinct for consecutive segments. In particular, the mean $\mu_i^\star = \mathbb{E}[X_i]$ is constant on these segments. This model is extended to a more general setting in Chapter 2. We may also make a parametric assumption on the observations, that is, assume that the X_i have a density f_θ, with $\theta = \theta_\ell$ on segment $\ell + 1$.

Hypothesis testing

A first idea for detecting a single change-point is to cast this problem as a hypothesis test. More specifically, one can test the null hypothesis

$$\mathcal{L}(X_1) = \mathcal{L}(X_2) = \cdots = \mathcal{L}(X_n) \ , \tag{H_0}$$

versus the alternative hypotheses

$$\exists t \in \big\{1, \dots, n-1\big\}, \qquad \mathcal{L}(X_1) = \cdots = \mathcal{L}(X_t) \neq \mathcal{L}(X_{t+1}) = \cdots = \mathcal{L}(X_n) \ . \tag{H_{alt}}$$

Note that (H_{alt}) is generally decomposed into the union of $n-1$ hypotheses

$$\mathcal{L}(X_1) = \cdots = \mathcal{L}(X_t) \neq \mathcal{L}(X_{t+1}) = \cdots = \mathcal{L}(X_n) \ . \tag{H_t}$$

In this section, we present some statistics for testing (H_0) versus (H_{alt}) existing in the literature, or variations thereof, in particular a one-sided version of (H_{alt}) when $\mathcal{L}(X_i)$ boils down to a single real parameter. For an exhaustive account of hypothesis testing in off-line change-point detection, we refer to Deshayes and Picard [1985] and

to James et al. [1987] for a comparison of the test powers — which we will not discuss below.

CUSUM and extensions. Let us restrict ourselves to the parametric framework defined in Section 1.3.2, that is, $\mathcal{L}(X_i) \sim f_\theta$ for some $\theta \in \Theta$. Keep in mind that θ can depend on i since the X_is are not identically distributed. Let us start with a very simple situation where the initial and final parameters are known, say θ_0 and θ_1. One of the simplest ideas for testing is then to build a test statistic from the likelihood function associated with the observations $X_1, \dots, X_n$. In Page [1957], the first off-line method that we know of, it is proposed to choose the (H_t) that maximizes the likelihood function of the hypothesis, that is, τ_1 is estimated by

$$\widehat{\tau}_1 \in \operatorname*{arg\,max}_{1 \le t \le n} \left\{ \sum_{j=1}^{t} \log f_{\theta_0}(x_j) + \sum_{j=t+1}^{n} \log f_{\theta_1}(x_j) \right\}.$$

In the case of a one-sided change in the mean of Gaussian observations with a known variance σ^2, Page [1957] recovers the cumulative sum from the CUSUM algorithm introduced at the beginning of Section 1.3.2. More precisely, we define

$$S_t := \sum_{j=1}^{t} (X_j - \theta_0 + \delta\sigma) \ ,$$

then reject (H_0) if $S_n - \max_{t<n} S_t < -h$, where h is a positive number, and in this case choose $\widehat{\tau}_1$ as the first t such that $S_n - \max_{t<n} S_t \ge 0$. This line of ideas is further studied in a series of three papers by Hinkley [1969, 1970, 1971], which obtain the asymptotic distribution of the estimate under Gaussian assumption.

If the initial parameter θ_0 is not known — but the variance σ^2 is still known —, Sen and Srivastava [1975a] following Gardner [1969] propose the statistic

$$U^\star = \frac{1}{n^2} \sum_{i=1}^{n-1} \left(\sum_{j=i}^{n-1} (X_{j+1} - \bar{X}) \right)^2 \ ,$$

and obtain the exact cumulative distribution function of $U^\star$, leading to power computations. Sen and Srivastava [1975a] claim that the power obtained is better for the test built with $U^\star$ than for the maximum likelihood statistic test.

This line of work was extended for unknown variance in Sen and Srivastava [1975b], following Chernoff and Zacks [1964], proposing among others the statistic $P_1 = U/V_1^{1/2}$, where

$$U = \sum_{i=1}^{n-1} i(X_{i+1} - \bar{X}) \qquad \text{and} \qquad V_1 = \frac{1}{2(n-1)} \sum_{i=1}^{n} (X_{i+1} - X_i)^2 \ .$$

The limiting distribution is not obtained in closed-form, but simulations suggest that P_1 has superior power.

Likelihood ratio test. Another possibility as a test statistic is called the likelihood ratio. The idea is to use as a statistic (the log of) the ratio between the likelihood under (H_{alt}) and the likelihood under (H_0). Suppose that σ^2 is known, and set

$$Q_k = \sum_{j=1}^{k} \left(X_j - \bar{X}_{1:k}\right)^2 + \sum_{j=k+1}^{n} \left(X_j - \bar{X}_{(k+1):n}\right)^2 ,$$

where $\bar{X}_{a:b}$ denotes the sample mean of the $b-a$ observations $X_{a+1}, \ldots, X_b$. Then the likelihood ratio test statistic for Gaussian observations of unknown mean before and after the change is given by $Q_n - Q_{k^\star}$; whenever σ^2 is unknown, this test statistic becomes equivalent to $\left(Q_n - Q_{k^\star}\right)/Q_{k^\star}$ [Hinkley, 1970; Hawkins, 1977]. The null distribution is given by Worsley [1979] when the variance is unknown, and this statistic was also adapted by Worsley [1986] in the exponential setting. It was later generalized in the multi-dimensional setting [Srivastava and Worsley, 1986; Arias-Castro et al., 2011]. We will encounter Q_k again, or rather a generalization of Q_k to the multiple change-point setting, later in this section.

Estimation procedures

In this section, we turn to an estimation-formulated version of the change-point problem. Rather than testing the possibility for each i to be a change-point, such methods aim to propose an estimator $\widehat{\tau}_n = \widehat{\tau}_n(X_1, \ldots, X_n)$ of the true segmentation $\tau^\star$. Assessing the quality of such an estimator is one of the central themes of Chapter 2, we thus recall some of the theoretical results associated with the methods presented in this section. Before being able to present these results, we want to be more precise on the meaning of *quality* in this context.

A first definition of consistency. As is often the case in statistics, a question of crucial importance to the practitioner is the adequacy between the estimator $\widehat{\tau}_n$ and $\tau^\star$, especially when the sample size n grows to infinity. We call *consistency* the asymptotic adequacy between $\widehat{\tau}_n$ and $\tau^\star$. In general, consistency results take the form

$$\text{with high probability,} \qquad \mathrm{d}\left(\widehat{\tau}_n, \tau^\star\right) \to 0 ,$$

where $\mathrm{d}\left(\cdot, \cdot\right)$ is a measure of similarity between segmentations — see Section 3.2.2 for a more involved treatment of this notion. Quite often, it is possible to give a quantitative version of the previous display, that is,

$$\text{with high probability,} \qquad \mathrm{d}\left(\widehat{\tau}_n, \tau^\star\right) \leq r_n ,$$

with $r_n \to 0$. We call r_n a *rate* of convergence. We will talk about *almost sure* consistency if, in the above statements, "with high probability" is replaced by "almost surely".

In the on-line setting, the meaning of $n \to \infty$ is rather obvious: the algorithm collects more and more data sequentially. However, in the off-line setting, there

are several ways to grow the sample size to infinity. The simplest asymptotic setting consists in observing for each n a sequence $X_{n,1}, \ldots, X_{n,n}$ of random variables, with the distribution of the $X_{n,i}$ being constant on the *segments* $\{1, \ldots, \tau^\star_{n,1}\}, \{\tau^\star_{n,1} + 1, \ldots, \tau^\star_{n,2}\}, \ldots, \{\tau^\star_{n,D^\star-1}+1, \ldots, n\}$, and the segments depending on n in the following fashion:

$$\exists \alpha_1, \ldots, \alpha_{D^\star-1} \in (0,1), \quad \forall 1 \leq i \leq D^\star - 1, \qquad \frac{\tau^\star_{n,i}}{n} \underset{n\to\infty}{\longrightarrow} \alpha_i \,. \tag{1.3}$$

We will often drop the n subscripts whenever the dependency in n is clear. An illustration of the asymptotic setting described here is given in Fig. 1-5.

Note that, in the setting (1.3), normalized segment sizes are bounded away from 0. In this case, it is known that the best possible rate achievable is $1/n$ [Korostelev, 1988; Korostelev and Tsybakov, 2012]. Namely,

$$\sup_{1\leq i\leq n} \left|\widehat{\tau}_{n,i} - \tau^\star_{n,i}\right| = \mathrm{O}_{\mathbb{P}}\left(\frac{1}{n}\right) .$$

Whenever this assumption is not satisfied, this rate degrades to $\log(n)/n$ [Brunel, 2014].

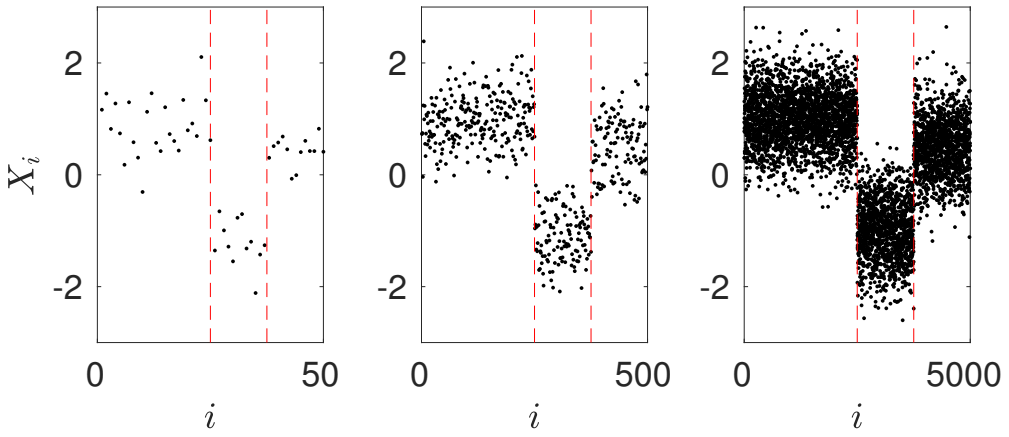

Figure 1-5 – Three samples $X_1, \ldots, X_n$ with several values of n. The distribution of the X_i is Gaussian with mean 1 (resp. -1, 0.5) on the first (resp. second, third) segment and standard deviation $\sigma = 0.5$. The increasing sample size leads to more observations on the segments, whose normalized sizes converge to constant numbers. *Left panel*: $n = 50$, *Middle panel*: $n = 500$, *Right panel*: $n = 5000$.

Maximum likelihood. Suppose that we are in a parametric setting and that we know the true number of change-points $D^\star$. It is then possible to generalize the approach of Hinkley [1970] and to write down the maximum likelihood estimator

of $\tau^\star$ as

$$\widehat{\tau} \in \underset{\substack{1\leq\tau_1<\cdots<\tau_{D-1}<n \\ \theta_1,\ldots,\theta_{D^\star}\in\Theta}}{\arg\max} \left\{ \sum_{\ell=1}^{D^\star} \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \log f_{\theta_\ell}(x_i) \right\}. \tag{1.4}$$

It is shown to be consistent in probability in He and Severini [2010] under a compactness hypothesis and technical assumptions on the behavior of the log-likelihood function.

Least-squares. Let us now assume furthermore that the observations are Gaussian with known variance σ^2, that is,

$$f_\theta(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\theta-t)^2}{2\sigma^2}\right),$$

with $\theta \in \mathbb{R}$. Then simple algebra shows that (1.4) becomes

$$\widehat{\tau}(D) \in \underset{\tau \text{ s.t. } D_\tau = D}{\arg\min} \left\{ \widehat{\mathcal{R}}_n(\tau) \right\}, \tag{1.5}$$

with

$$\widehat{\mathcal{R}}_n(\tau) := \frac{1}{n} \sum_{\ell=1}^{D} \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \left(X_i - \bar{X}_{\tau_{\ell-1}:\tau_\ell}\right)^2.$$

We call $\widehat{\mathcal{R}}_n(\tau)$ the sum of squares criterion or *least-squares criterion*, which is also an estimate of the variance assuming that the true segmentation is τ. It is very convenient to use the least-squares criterion rather than the likelihood function, since we do not need to know the distribution of the observations.

Fisher [1958] is the first to apply the least-squares criterion for a change-point problem to the best of our knowledge — note that his approach does not come from likelihood maximization but rather from variance minimization. Yao and Au [1989] prove that $\widehat{\tau}(D)$ is consistent in probability in the asymptotic setting (1.3) and under mild assumptions — namely the continuity of the cumulative distribution function of the observations and a moment hypothesis. These assumptions are weakened further in Bai and Perron [1998] and the minimax convergence rate of $1/n$ is obtained. The least-squares estimation procedure (1.5) was also shown to be consistent in the case of dependent processes (ARMA) with a single change-point in Bai [1994], a work later extended for weak dependent disturbance processes (mixingales) by Bai and Perron [1998]. Regarding multiple break-points, Lavielle [1999]; Lavielle and Moulines [2000] show the consistency of the least-squares estimate when $D^\star$ is known for a large class of dependent processes.

More precisely, define $\varepsilon_i := X_i - \mu_i^\star$ and set $S_{a:b} := \sum_{i=a}^{b} \varepsilon_i$, then the main assumption of Lavielle and Moulines [2000] on the dependency structure of the noise ε is, for some $\phi > 0$,

$$\exists C > 0, \quad \forall 1 \leq i, j \leq n, \qquad \mathbb{E}\left[S_{i:j}^2\right] \leq C\,|j-i+1|^\phi. \tag{$H(\phi)$}$$

This assumption is satisfied $\phi = 1$ for stationary processes such that the autocovariance function $\gamma(s) = \mathbb{E}[\varepsilon_{t+s}\varepsilon_t]$ satisfies $\sum_{s\geq 0} |\gamma(s)| < \infty$, and a variety of linear processes, e.g., any ARMA process. Assuming $(H(\phi))$ with $\phi \in [1, 2)$, $\widehat{\tau}(D^\star)$ is consistent with convergence rate $n^{\phi-2}$ [Lavielle and Moulines, 2000].

It is interesting to see that, at first sight, the minimization problem (1.5) seems rather daunting. Indeed, as we noticed before, the total number of segmentations with D segments is $\binom{n-1}{D-1}$ — too large a number for optimizing directly. But it turns out that (1.5) can be solved exactly in $\mathrm{O}(Dn^2)$ thanks to dynamic programming [Bellman, 1961]. We will discuss further this algorithmic question in Section 2.4.

Unknown number of change-points. Whenever the number of change-points is not known, the problem becomes far more compelling. Indeed, minimizing directly (1.5) without constraints on D_τ systematically outputs the segmentation consisting of n segments of unit size, which is definitely not insightful: the number of segments has to be chosen in another way. The idea of Yao [1988] is to consider the choice of $\widehat{D}$ as a model selection problem and to choose the number of change-points according to Schwartz criterion [Schwarz, 1978]. More precisely, given a $\widehat{\tau}(D)$ that minimizes $\widehat{\mathcal{R}}_n(\tau)$ for each D, Yao [1988] then chooses

$$\widehat{D}^{\mathrm{Yao}} \in \operatorname*{arg\,min}_{1\leq D\leq D_{\max}} \left\{ \frac{n}{2} \log \widehat{\mathcal{R}}_n(\widehat{\tau}(D)) + D \log(n) \right\},$$

where $D_{\max}$ is a user-defined upper bound on the number of segments, and set $\widehat{\tau}^{\mathrm{Yao}} := \widehat{\tau}(\widehat{D}^{\mathrm{Yao}})$. For homoscedastic [4] independent Gaussian observations and under Assumption (1.3), Yao [1988] shows that $\widehat{\tau}^{\mathrm{Yao}}$ is consistent in probability.

This result was extended for identically distributed error terms ε_i by Yao and Au [1989], with however a more restrictive condition on the growth of β_n. More precisely, let us assume, in addition to Assumption (1.3), that (i) the cumulative distribution function of the ε_i is continuous, (ii) $\mathbb{E}[\varepsilon_i^{2m}] < +\infty$ with $m \geq 3$. Let us also define

$$\widehat{D}^{\mathrm{YAu}} \in \operatorname*{arg\,min}_{1\leq D\leq D_{\max}} \left\{ n \log \widehat{\mathcal{R}}_n(\widehat{\tau}(D)) + D\beta_n \right\}, \tag{1.6}$$

with β_n satisfying $\beta_n n^{-2/m} \to \infty$ and $\beta_n n^{-1} \to 0$. In particular, $\beta_n \to \infty$. Then, $\widehat{\tau}^{\mathrm{YAu}} := \widehat{\tau}(\widehat{D}^{\mathrm{YAu}})$ is consistent in probability. Lee [1995, 1997] uses this estimator for observations belonging to an independent exponential family, that is, the $X_i \in \mathbb{R}$ are independent and have density with respect to the Lebesgue measure

$$f_\theta(x) = \exp\left(\theta x + \phi(\theta) + s(x)\right) .$$

The consistency is also proven for $\beta_n \gtrsim \log(n)$.

Let us mention that it is also possible to minimize the least-squares criterion for a given number of segments D, and then to use a testing procedure to decide if more

4. A sequence of random variables is *homoscedastic* if all the random variables in the sequence have the same variance. If not, we say that the random variables are *heteroscedastic*.

segments ought to be added, rather than using a criterion similar to (1.6) [see, e.g., Bai and Perron, 1998].

Penalized least-squares criterion

But the scheme that interests us the most is the minimization of the sum of the least-squares criterion and an additional term that increases with the number of segments. This term prevents us from choosing τ with too many segments. On the other side, if it is chosen too large, we risk selecting a segmentation with too few segments. More precisely, let us define $\mathcal{T}_n$ as the set of all segmentations of $\{1, \ldots, n\}$. In our setting, we call *penalty function* any mapping $\text{pen} : \mathcal{T}_n \to \mathbb{R}_+$. If $D_{\tau^1} \leq D_{\tau^2}$ implies $\text{pen}(\tau^1) \leq \text{pen}(\tau^2)$, we will say that the penalty is *non-decreasing*. We can now define the penalized least-squares procedure

$$\widehat{\tau}^{\text{pen}} \in \underset{\tau \in \mathcal{T}_n}{\arg\min} \left\{ \widehat{\mathcal{R}}_n(\tau) + \text{pen}(\tau) \right\}. \tag{1.7}$$

Let us consider a penalty function proportional to the number of segments, that is,

$$\text{pen}_\ell(\tau) := \frac{\beta_n D_\tau}{n}, \tag{1.8}$$

with β_n any sequence such that $\beta_n \to \infty$ and $\beta_n / n \to 0$. Then, under $(H(\phi))$ for any $\phi \in [1, 2)$, the estimator $\widehat{\tau}^{\text{pen}_{\text{LM}}}$ is consistent in probability with rate $1/n$ [Lavielle and Moulines, 2000; Lavielle, 2005]. It is remarkable to notice that the rate of convergence obtained is also minimax.

A linear penalty is not the only possibility. When the variance of the noise σ^2 is known, following ideas coming from model selection, Lebarbier [2002] proposes to replace the $\beta_n D_\tau$ term in (1.8) with

$$\text{pen}_{\text{L}}(\tau) := \frac{D_\tau \sigma^2}{n} \left(c_1 \log \frac{n}{D_\tau} + c_2 \right), \tag{1.9}$$

where c_1 and c_2 are positive constants. Calibrating these constants is of course a key question; Lebarbier [2002] argues in favor of setting $c_2 / c_1 = 2.5$ and using a *slope heuristic* [Baudry et al., 2012] for choosing the constant in front of the penalty shape. If the variance of the observations is unknown, Lebarbier [2002] also advocates to replace σ^2 by a Hall estimate [Hall et al., 1990]. In the case of heteroscedastic noise, a very different approach is to use cross-validation instead of penalization [Arlot and Celisse, 2011].

Using a result from Birgé and Massart [2001], it is proved in Lebarbier [2005] that the estimator of $\mu^\star$ satisfies an oracle inequality. It is another way to look at the quality of a change-point method, that is, to look at the quadratic risk between the estimator of $\mu^\star$ naturally associated with $\widehat{\tau}^{\text{Leb}}$ and the true piecewise constant function $\mu^\star$. Before stating this result more precisely, we introduce some notations. To each segmentation τ we associate X_τ, defined as the empirical mean of X on each

segment of τ. Namely,

$$\forall \tau_{\ell-1}+1 \leq i \leq \tau_\ell, \quad (X_\tau)_i := \frac{1}{|\tau_\ell - \tau_{\ell-1}|} \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} X_i \,.$$

Let us also define $\widehat{\mu} = X_{\widehat{\tau}}$ and $\|x\|^2 = \frac{1}{n}\sum_{i=1}^n x_i^2$. Then there exists some constants $C(c_1, c_2)$ and $C'(c_1, c_2)$ such that

$$\mathbb{E}\left[\|\mu^\star - \widehat{\mu}\|^2\right] \leq C \inf_{\tau \in \mathcal{T}_n} \left\{ \|\mu^\star - X_\tau\|^2 + \mathrm{pen}\,(\tau) \right\} + C' \frac{\sigma^2}{n} \,.$$

Informally, the previous display means that, in terms of quadratic risk, $\widehat{\tau}$ is as good as the best of the τ up to a $\log(n)$ factor. Note that this result does not give any information about the quality of the segmentation with respect to $\tau^\star$.

Minimizing the least-squares criterion with a linear penalty (1.7) can be seen as the following optimization problem:

$$\underset{u \in \mathbb{R}^n}{\text{Minimize}} \, \|X-u\|^2 \quad \text{subject to} \quad \|(u_{i+1}-u_i)_{1\leq i<n}\|_0 = D^\star \,,$$

where $\|x\|_0$ is the number of non-zero components of x. Harchaoui and Lévy-Leduc [2010] propose to relax this ℓ_0 constraint into a ℓ_1 constraint, that is, to solve

$$\underset{u \in \mathbb{R}^n}{\text{Minimize}} \, \|X-u\|^2 \quad \text{subject to} \quad \|(u_{i+1}-u_i)_{1\leq i<n}\|_1 \leq D^\star \underline{\Delta} \,,$$

where $\underline{\Delta} := \max \left|\mu^\star_{i+1} - \mu^\star_i\right|$ and $\|x\|_1 := \sum_{i=1}^n |x_i|$ is the ℓ_1 norm. It turns out that the previous display exactly corresponds to the Least Absolute Shrinkage eStimatOr (LASSO) in least-squares regression [Tibshirani, 1996]. A major feature of this approach is to decrease the computational cost from $\mathrm{O}\left(D_{\max} n^2\right)$ to $\mathrm{O}\left(D_{\max} n \log(n)\right)$ — see Section 2.4 for more details on the implementation of penalized least-squares methods. Still, the estimated segmentation is consistent: for any segmentation τ, denote by $\underline{\Lambda}_\tau$ the normalized size of its smallest segment. Then, assuming sub-Gaussian noise [5], $\underline{\Lambda}_{\tau^\star} \geq (\log(n))^2/n^2$, and $\underline{\Delta} \geq (\log(n))^{1/4}$, Harchaoui and Lévy-Leduc [2010] prove, in particular, that the change-point locations are consistent conditionally to $\widehat{D} = D^\star$ with high probability. Note that the convergence rate $(\log(n))^2/n$ is optimal up to a logarithmic factor.

This approach has been successfully generalized to the multi-dimensional setting [Bleakley and Vert, 2011]; in this case it can be shown that the problem is equivalent to the group LASSO [Bakin, 1999; Lin and Zhang, 2006]. Bleakley and Vert [2011] show that their procedure is consistent in probability for a single change-point, provided that the noise level is smaller than a threshold depending essentially on the size and location of the jump.

5. We say that a random variable S is sub-Gaussian if there exists $C, v > 0$ such that $\mathbb{P}\left(|S| > t\right) \leq C\,\mathrm{e}^{-vt^2}$. Informally, the tails of the distribution of S decay at least as fast as the tails of a Gaussian random variable.

This concludes our introduction to change-point detection. We will complete our survey of the literature in Section 2.1.

1.4 Kernel methods

We now continue this general introduction with a brief presentation of kernel methods, the missing brick in the construction of kernel change-point detection — abbreviated KCP from now on. We do not intend to cover all of this vast topic, and we refer to Vert et al. [2004] for an introduction and to the monograph of Schölkopf and Smola [2002] for an extensive overview. In the following short introduction, we rather focus on essential tools for the understanding of the manuscript.

1.4.1 Positive semi-definite kernels

A fundamental idea when dealing with data is that of *similarity measure*, that is, a real-valued function whose values quantify how close two objects are. In most cases, the data has a d-dimensional representation and the Euclidean structure of $\mathbb{R}^d$ naturally provides a satisfactory notion of similarity between observations inherited from the scalar product. For instance, the dot product between two data-points A and B lying in the unit sphere reduces to the cosine of the angle $\widehat{AOB}$, a quantity that is close to 1 if A and B are in the same neighborhood and smaller otherwise.

Consider data living in a space $\mathcal{X}$ such that the Euclidean metric does not reflect the structure of the data or which is not equipped with a scalar product. Can we replace the inner product by a mapping $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that $k(\cdot, \cdot)$ and $\langle \cdot, \cdot \rangle$ benefit from similar properties? Can we define a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ which would mimic the behavior of a scalar product on $\mathcal{X}$ if it existed?

When the answer is affirmative, we shall say that k is a *positive semi-definite kernel*, a notion that dates back to Mercer [1909], building on ideas from Hilbert [1904]. We now give a precise definition.

Definition 1.1. Consider a non-empty set $\mathcal{X}$. Given a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $m \in \mathbb{N}^\star$ and $x_1, \dots, x_m \in \mathcal{X}$, then the matrix

$$\mathrm{K} := \left(k(x_i, x_j)\right)_{1 \leq i,j \leq m} \in \mathbb{R}^{m \times m} \tag{1.10}$$

is called the *Gram matrix* of k with respect to $x_1, \dots, x_m$. Any function k that gives rise to a positive semi-definite Gram matrix for any $x_1, \dots, x_m \in \mathcal{X}$ is called a *positive semi-definite kernel*. Equivalently, we ask for k to be a symmetric function such that, for any $m \in \mathbb{N}^\star$, $x_1, \dots, x_m \in \mathcal{X}$ and $\lambda_1, \dots, \lambda_m \in \mathbb{R}$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j k(x_i, x_j) \geq 0 \,. \tag{1.11}$$

In what follows, we use this concept extensively, hence we will abbreviate "positive semi-definite kernel" into "p.s.d. kernel", and often into "kernel". Note that it is

possible to replace "positive semi-definite" by "positive definite" in Definition 1.1 — or equivalently "non-negative" by "positive" in (1.11) —, we then speak of *positive definite* kernel. We will not use this slightly more restrictive notion. Neither are we concerned with extensions of positive semi-definite kernels, such as kernels with values in $\mathbb{C}$.

1.4.2 Examples of kernels

Let us give a few examples of kernels together with proofs of their positive semi-definiteness in the simplest cases.

Linear kernel. For data belonging to $\mathcal{X} = \mathbb{R}^d$, the most basic example is the *linear kernel* $k_\ell(x, y) := \langle x, y \rangle$. See Fig. 1-6 for a plot of $k_\ell(x, y)$ when $d = 1$. It is routine to verify Eq. (1.11) since

$$\sum_{i,j=1}^{m} \lambda_i \lambda_j k_\ell(x_i, x_j) = \left\| \sum_{i=1}^{m} \lambda_i x_i \right\|^2 \geq 0 \,,$$

by linearity of the dot product.

Polynomial kernel. A natural extension of the linear kernel is the *polynomial kernel*, defined as

$$k_{\mathrm{P}}(x, y) := (\langle x, y \rangle + c)^\alpha \,,$$

for $x, y \in \mathcal{X} = \mathbb{R}^d$, $c \geq 0$ and $\alpha \in \mathbb{N}^\star$. See Fig. 1-6 for a plot of $k_{\mathrm{P}}(x, y)$ when $d = 1$. Let us show that k_{P} is positive semi-definite when $\alpha = 2$. In this case, according to the multinomial theorem, $k_{\mathrm{P}}(x, y)$ can be written

$$\sum_{i=1}^{d} \left(x_i^2\right) \left(y_i^2\right) + \sum_{i=2}^{d} \sum_{j=1}^{i-1} \left(\sqrt{2} x_i x_j\right) \left(\sqrt{2} y_i y_j\right) + \sum_{i=1}^{d} \left(\sqrt{2c} x_i\right) \left(\sqrt{2c} y_i\right) + c^2 \,.$$

Therefore, $k_{\mathrm{P}}(x, y) = \langle \Phi(x), \Phi(y) \rangle$ with $\Phi(x)$ defined as

$$\left(x_n^2, \ldots, x_1^2, \sqrt{2} x_n x_{n-1}, \ldots, \sqrt{2} x_n x_1, \sqrt{2} x_{n-1} x_{n-2}, \ldots, \sqrt{2c} x_n \ldots, c\right)^\top \in \mathbb{R}^{\frac{(n+1)(n+2)}{2}} \,.$$

The proof we used for the linear kernel can then be immediately adapted to the polynomial kernel. We will see later that such a mapping Φ exists for *any* kernel k. Of course, the reasoning above can be adapted to any α.

Gaussian kernel. A widely used kernel is the so-called *Gaussian* kernel, introduced by Boser et al. [1992]. It is defined by

$$k_{\mathrm{G}}(x, y) := \exp\left(\frac{-\|x - y\|^2}{2\nu^2}\right) \,,$$

for $x, y \in \mathcal{X} = \mathbb{R}^d$ and $\nu > 0$. See Fig. 1-6 a plot of $k_\ell(x, y)$ when $d = 1$ and $\nu = 1.0$. Let Z be a standard Gaussian random variable $\mathcal{N}(0, \mathrm{I}_d)$ and $\lambda_1, \ldots, \lambda_m \in \mathbb{R}$. Then

$$\sum_{i,j=1}^{m} \lambda_i \lambda_j k_{\mathrm{G}}(x_i, x_j) = \sum_{i,j=1}^{m} \lambda_i \lambda_j \, \mathbb{E}\left[\mathrm{e}^{\mathrm{i}\nu^{-1}(x_i - x_j)^\top Z}\right] = \mathbb{E}\left[\left|\sum_{i=1}^{m} \lambda_i \, \mathrm{e}^{\mathrm{i}\nu^{-1} x_i^\top Z}\right|^2\right] \geq 0\,,$$

hence k_{G} is a positive-definite kernel.

Laplace kernel. The *Laplace* kernel is very similar to the Gaussian kernel, defined by $k_{\mathrm{L}}(x, y) := \exp\left(-\left\|x - y\right\| / \nu\right)$, for $x, y \in \mathcal{X}$ and $\nu \geq 0$. It is sometimes called the exponential kernel [Genton, 2001]. Together, they belong to a larger class of kernels such that $k(x, y) = \kappa(x - y)$ for some function κ. These kernels are called *translation-invariant* since they depend only on the difference between the input vectors. They are considered in more depth in Section 4.3.1, where we give a precise definition and additional properties. We refer to Berg et al. [1984] for a systematic study of such kernels. In this example and the previous one, it is clear that $k(x, y)$ is large whenever x and y are near, since $k(x, y)$ is a decreasing function of $\|x - y\|$. See Fig. 1-6 a plot of $k_{\mathrm{L}}(x, y)$ when $d = 1$ and $\nu = 1.0$.

Graph kernels. Given two finite graphs $G_1 = (E_1, V_1)$ and $G_2 = (E_2, V_2)$, one can define the direct product of G_1 and G_2 as the graph $G_\times = (V_\times, E_\times)$ that has vertex set $V_1 \times V_2$, and edges given by the rule: "(v_1, v_1') is connected to (v_2, v_2') if and only if v_1 is connected to v_2 and v_1' is connected to v_2'." Denote by $A_\times$ the adjacency matrix of $G_\times$, and pick a sequence $(a_p)_{p \geq 0}$ of positive weights. Then the *direct product* kernel [Gärtner et al., 2003] is defined as

$$k_\times(G_1, G_2) := \sum_{i,j=1}^{|V_\times|} \left(\sum_{p=0}^{+\infty} a_p A_\times^p\right)_{i,j}$$

if the limit exists — see also Vishwanathan et al. [2010]. Note that it is already more difficult than in the previous examples to see why $k_\times$ is a similarity measure on the set of finite graphs.

1.4.3 The kernel-trick

From Definition 1.1, clearly any scalar product on $\mathcal{X}$ is a kernel. In this sense, kernels can be seen as a generalization of scalar product. [6] But in fact this analogy between kernels and dot products runs deeper, as was hinted in the previous examples. Let us recall that a Hilbert space is a real or complex inner product space that is also a complete metric space with respect to the distance induced by the inner product. The fundamental property of kernels is the following:

6. Let us stress that some key properties of dot products are not satisfied by kernels, in particular linearity.

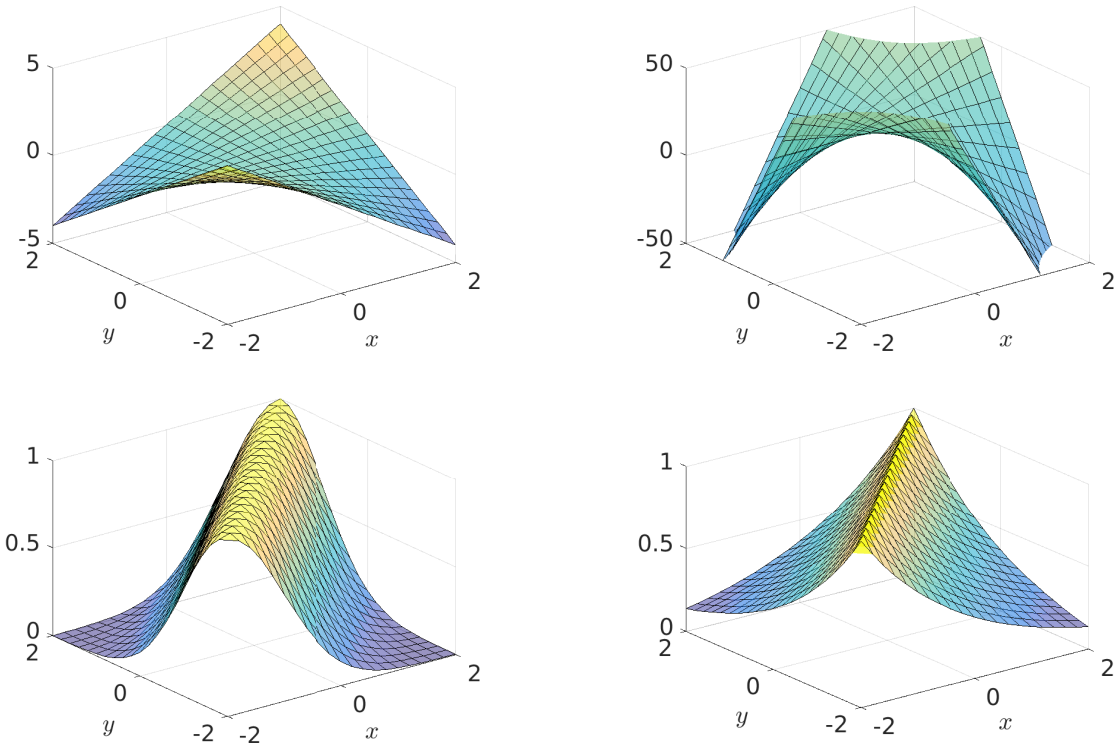

Figure 1-6 – In this figure, we plot the values of $k(x, y)$ for different positive semi-definite kernels and $x, y \in \mathbb{R}$. *Upper left:* linear kernel; *Upper right:* polynomial kernel with $\alpha = 2$ and $c = 1$; *Bottom left:* Gaussian kernel with $\nu = 1.0$; *Bottom right:* Laplace kernel with $\nu = 1.0$.

Theorem 1.1 (Moore—Aronszajn)**.** *The mapping k is a positive semi-definite kernel on $\mathcal{X}$ if, and only if, there exists a Hilbert space $\mathcal{H}$ and a mapping $\Phi : \mathcal{X} \to \mathcal{H}$ such that*

$$\forall x, y \in \mathcal{X}, \qquad k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} \,. \tag{1.12}$$

Note that the reverse sense of Theorem 1.1 is clearly true; it follows directly by setting (1.12) as a definition for k and using the properties of the dot product. The non-trivial part in Theorem 1.1 is the direct implication. It was first obtained by Mercer [1909] for $\mathcal{X} = [a, b] \subset \mathbb{R}$ and continuous k, and later proved by Kolmogorov [1941] for countable $\mathcal{X}$. The general result is stated by Aronszajn [1950], which refers to Aronszajn [1943] and attributes the result to Moore [1916, 1935, 1939]. Moreover the Hilbert space $\mathcal{H}$ is essentially unique.

Before any further comment, let us give a short proof of Theorem 1.1 in the special case where $\mathcal{X}$ is a finite set.

Proof. We only prove the direct implication. Let us assume that $\mathcal{X} = \{x_1, \dots, x_m\}$. Then any kernel k on $\mathcal{X}$ is entirely defined by the matrix $\mathrm{K} = \left(k\left(x_i, x_j\right)\right)_{1 \leq i,j \leq m}$, which is positive semi-definite. In particular, it can be diagonalized on an orthonormal basis of eigenvectors $(e_1, \dots, e_m) \in \mathbb{R}^{m \times m}$, with eigenvalues

$$0 \leq \zeta_1 \leq \cdots \leq \zeta_m \,.$$

Define $\Phi(x_i) := \left(\sqrt{\zeta_1}\left(e_1\right)_i, \dots, \sqrt{\zeta_m}\left(e_m\right)_i\right)^\top \in \mathbb{R}^m$. Then, for any $1 \leq i, j \leq m$,

$$\langle \Phi(x_i), \Phi(x_j) \rangle = \sum_{\ell=1}^{m} \zeta_\ell \left(e_\ell\right)_i \left(e_\ell\right)_j = \left(\sum_{\ell=1}^{m} \zeta_\ell e_\ell e_\ell^\top \right)_{i,j} = k\left(x_i, x_j\right) \,.$$

□

This link between positive definite kernels and Hilbert spaces has a simple, yet crucial consequence. Often referred to as the *kernel-trick*, Equation (1.12) allows to run any algorithm that depends only on the scalar products $\langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}}$ without actually doing any calculations in $\mathcal{H}$, replacing them by evaluations $k\left(x_i, x_j\right)$ of the kernel map.

This fact was first uncovered by the statistical community in the 60s for specific kernels, *cf.* Aizerman et al. [1964] for instance. The full generality of the kernel trick was made clear in the work of Schölkopf et al. [1997, 1998]. This principle has since been applied in countless settings, see Schölkopf and Smola [2002] for a full account. In particular, the possibility to extend classical methods to data living in general sets such as graphs or texts is a huge benefit from the use of kernels.

We now give an elementary example to illustrate this principle. Fix $\mathcal{X}$ and k, and let $\mathcal{H}$ be the associated Hilbert space and Φ the feature map. Take $x, y \in \mathcal{X}$, and suppose that we want to compute the distance between $\Phi(x)$ and $\Phi(y)$ in $\mathcal{H}$. We write

$$\|\Phi(x) - \Phi(y)\|_{\mathcal{H}}^2 = \langle \Phi(x) - \Phi(y), \Phi(x) - \Phi(y) \rangle_{\mathcal{H}}$$

$$= \langle \Phi(x), \Phi(x) \rangle_{\mathcal{H}} - 2\langle \Phi(x), \Phi(y) \rangle_{\mathcal{H}} + \langle \Phi(y), \Phi(y) \rangle_{\mathcal{H}}$$
$$\|\Phi(x) - \Phi(y)\|_{\mathcal{H}}^2 = k(x,x) - 2k(x,y) + k(y,y) \ ,$$

thus

$$\mathrm{d}(\Phi(x), \Phi(y)) = \sqrt{k(x,x) - 2k(x,y) + k(y,y)} \, .$$

As promised, for points in $\mathcal{X}$, pairwise distance computations in $\mathcal{H}$ require only the knowledge of the Gram matrix of k. We refer to Fig. 1-7 for an illustration.

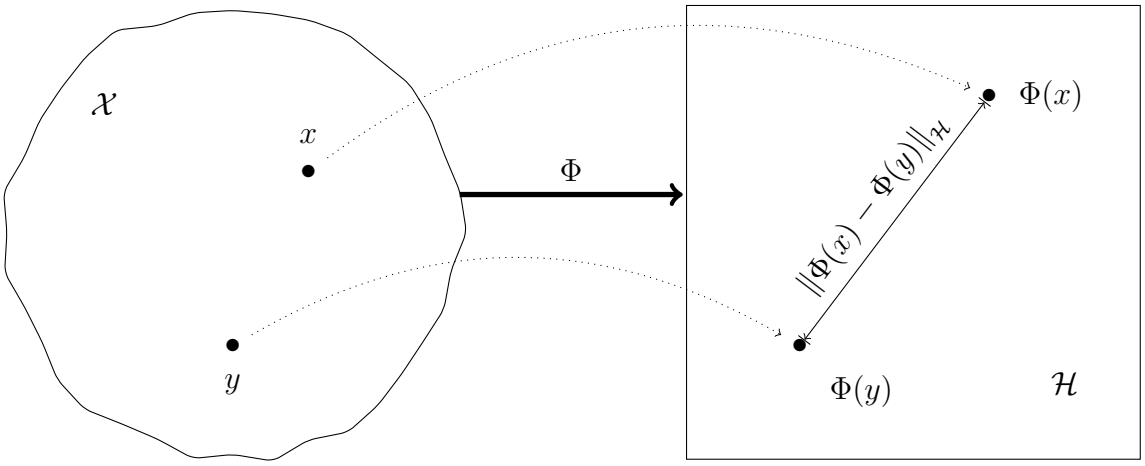

Figure 1-7 – The points x and y are mapped from $\mathcal{X}$ to the Hilbert space $\mathcal{H}$ associated to k *via* the feature map Φ. The distance between $\Phi(x)$ and $\Phi(y)$ in $\mathcal{H}$ can be computed without explicit computations in $\mathcal{H}$.

A remarkable fact is that the mapping Φ in Theorem 1.1 is explicit. Let us define a *Reproducing Kernel Hilbert Space*, latter abbreviated RKHS.

Definition 1.2 (RKHS)**.** Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a Hilbert space of functions[7] $f : \mathcal{X} \to \mathbb{R}$. Then $\mathcal{H}$ is called a RKHS if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:

(i) k has the *reproducing* property

$$\forall f \in \mathcal{H}, \quad \forall x \in \mathcal{X}, \qquad \langle f, k(x, \cdot) \rangle = f(x) \, .$$

In particular, $\langle k(x,\cdot), k(y,\cdot) \rangle = k(x,y)$.

(ii) $\mathcal{H}$ is the completion of $\mathrm{Span}\left\{ k(x,\cdot) \mid x \in \mathcal{X} \right\}$.

We then say that k is a reproducing kernel. Then the following holds:

Theorem 1.2. *A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive semi-definite kernel if, and only if, it is a reproducing kernel.*

7. That is a Hilbert space whose elements are functions $f : \mathcal{X} \to \mathbb{R}$ and such that all the evaluation functionals are continuous.

In other words, we can write $\Phi(x) = k(x, \cdot)$ in Theorem 1.1. We call Φ the *feature map*.

This concludes our short introduction to kernel methods. Positive semi-definite kernels come with other fascinating features, as the representer theorem [Kimeldorf and Wahba, 1971], a result essentially useful in the resolution of optimization problems. As we will see in Section 2.4, the optimization problem that we are to solve does not require using this result, and we chose to exclude the representer theorem from this introduction.

We now turn to the center of this manuscript, kernel change-point detection.

Chapter 2

Kernel change-point detection

Abstract

In this chapter we introduce the kernel change-point algorithm proposed by Arlot et al. [2012], which aims at locating an unknown number of change-points in the distribution of a sequence of data taking values in an arbitrary set. Before presenting the practical implementation of the algorithm, we introduce some concepts and notations related to the RKHS setting, in particular the kernel mean embedding. We then define the theoretical framework under which our analysis is conducted, and recall briefly existing results concerning KCP.

2.1 Introduction

In many situations, some properties of a time series change over time, such as the mean, the variance or higher-order moments. Change-point detection is the long standing question of finding both the number and the localization of such changes. This is an important front-end task in many applications. For instance, detecting changes occurring in comparative genomic hybridization array data (CGH arrays) is crucial to the early diagnosis of cancer [Lai et al., 2005]. In finance, some intensively examined time series like the volatility process exhibit local homogeneity and it is useful to be able to segment these time series both for modeling and forecasting [Lavielle and Teyssiere, 2006; Spokoiny, 2009]. Change-point detection can also be used to detect changes in the activity of a cell [Ritov et al., 2002], in the structure of random Markov fields [Liu et al., 2017], or a sequence of images [Kim et al., 2009; Abou-Elailah et al., 2015]. Generally speaking, it is of interest to the practitioner to segment a time series in order to calibrate its model on homogeneous sets of data-points.

Addressing the change-point problem in practice requires to face several important challenges. First, the number of changes can not be assumed to be known in advance — in particular, it can not be assumed to be equal to 0 or 1 —, hence a practical change-point procedure must be able to infer the number of changes from the data. Second, changes do not always occur in the mean or the variance of the data, as assumed by most change-point procedures. We need to be able to detect changes in other features of the distribution. Third, parametric assumptions — which are often

made for building or for analyzing change-point procedures — are often unrealistic, so that we need a fully non-parametric approach. Fourth, data points in the time series we want to segment can be high-dimensional and/or structured. If the dimensionality is larger than the number of observations, a non-asymptotic analysis is mandatory for theoretical results to be meaningful. When data are structured — for instance, histograms, graphs or strings —, taking their structure into account seems necessary for detecting efficiently the change-points.

We focus only on the *off-line* problem, that is, when all observations are given at once, as opposed to the situation where data come as a continuous stream. We refer to Tartakovsky et al. [2014] for an extensive review of sequential methods, which are adapted to the latter situation. Numerous off-line change-point procedures have been proposed since the seminal works of Page [1955], Fisher [1958] and Bellman [1961], which are mostly parametric in essence. We refer to Brodsky and Darkhovsky [2013, Chapter 2] for a review of non-parametric off-line change-point detection methods. Among recent works in this direction, we can mention the Wild Binary Segmentation (WBS, [Fryzlewicz, 2014]) and the non-parametric multiple change-point detection procedure (NMCD, [Zou et al., 2014]). Some authors also consider the case of high-dimensional data when only a few coordinates of the mean change at each change-point [Wang and Samworth, 2016, and references therein], or the problem of detecting gradual changes [Vogt and Dette, 2015]; we do not address these slightly different problems.

To the best of our knowledge, no off-line change-point procedure addressed simultaneously the four challenges mentioned above, until KCP was proposed by Arlot et al. [2012]. In short, KCP mixes the penalized least-squares approach to change-point detection [Comte and Rozenholc, 2004; Lebarbier, 2005] with positive semi-definite kernels [Aronszajn, 1950]. It is not the only procedure that uses positive semi-definite kernels to detect changes in a times series. Apart from Harchaoui and Cappé [2007], who introduced KCP for a fixed number of change-points, and Arlot et al. [2012] who extended KCP to an unknown number of change-points, we are aware of several closely related work. Maximum Mean Discrepancy [MMD, Gretton et al., 2007] has been used for building two-sample tests; a block average version of the MMD, named the M-statistic, has lead to an on-line change-point detection procedure [Li et al., 2015]. A kernel-based statistic, named kernel Fisher discriminant ratio, has been used by Harchaoui et al. [2009] for homogeneity testing and for detecting a single change-point. Sharipov et al. [2016] build an analogue of the CUSUM statistic for Hilbert-valued random variables in order to detect a single change in the mean, and could be applied in our setting to the images of the observations in the feature space. Kernel change detection [Desobry et al., 2005] is an on-line procedure that uses a kernel to build a dissimilarity measure between the near past and future of a data-point.

We first introduce in Section 2.2 and 2.3 some notations and concepts that are necessary for the rest of the manuscript. In Section 2.4, we present the computational aspect of KCP. In brief, the KCP segmentation can be computed efficiently thanks to a dynamic programming algorithm [Harchaoui and Cappé, 2007; Arlot et al., 2012].

Section 2.5 is devoted to presenting the hypothesis that will be used in Chapter 3 for the theoretical study of KCP. Our framework is common to Arlot et al. [2012], that proved an oracle inequality for KCP. We recall this result in Section 2.6.

2.2 Kernel change-point detection

We first describe the change-point problem with our notations (Section 2.2.1) and the kernel change-point procedure (Section 2.2.2).

2.2.1 Change-point problem

Set $2 \leq n < +\infty$ and consider $X_1, \ldots, X_n$ independent $\mathcal{X}$-valued random variables, where $\mathcal{X}$ is an arbitrary (measurable) space. The goal of change-point detection is to detect abrupt changes in the distribution of the X_is. For any $D \in \{1, \ldots, n\}$ and any integers $0 = \tau_0 < \tau_1 < \cdots < \tau_D = n$, we define the *segmentation* $\tau := [\tau_0, \ldots, \tau_D]$ of $\{1, \ldots, n\}$ as the collection of segments $\lambda_\ell = \{\tau_{\ell-1} + 1, \ldots, \tau_\ell\}$, $\ell \in \{1, \ldots, D\}$. We call *change-points* the right-end of the segments, that is the τ_ℓ, $\ell \in \{1, \ldots, D\}$. Let us denote by $\mathcal{T}_n^D$ the set of segmentations with D segments [1] and $\mathcal{T}_n := \bigcup_{D=1}^n \mathcal{T}_n^D$ the set of all segmentations of $\{1, \ldots, n\}$. For any $\tau \in \mathcal{T}_n$, we write D_τ for the number of segments of τ. Fig. 2-1 provides a visual example.

Figure 2-1 – A typical graphical representation of a segmentation. The black solid disks stand for the ordered elements of $\{1, \ldots, n\}$, and the vertical lines denote the changes. Here, as one can read, $n = 10$, $D_\tau = 3$, $\tau_0 = 0$, $\tau_1 = 3$, $\tau_2 = 7$ and $\tau_3 = 10$ — thus $\tau = [0, 3, 7, 10]$.

An important example to have in mind is the following.

Example 2.1 (Asymptotic setting). Let $K \geq 1$, $0 = b_0 < b_1 < \cdots < b_K < b_{K+1} = 1$ and $P_1, \ldots, P_{K+1}$ some probability distributions on $\mathcal{X}$ be fixed. Then, for any n and $i \in \{1, \ldots, n\}$, we set $t_i := i/n$ and the distribution of X_i is $P_{j(i)}$ where $j(i)$ is such that $t_i \in [b_{j(i)}, b_{j(i)+1})$. In other words, we have a fixed segmentation of $[0, 1]$, given by the b_j, a fixed distribution over each segment, given by the P_j, and we observe independent realizations from the distributions at discrete times $t_1, \ldots, t_n$. The corresponding true change-points in $\{0, \ldots, n\}$ are the $\lfloor nb_j \rfloor$, $j = 1, \ldots, K$. For n large enough, there are $K + 1$ segments. Fig. 2-2 shows an example. Let us emphasize that in this setting, n going to infinity does not mean that new observations are observed over time. Recall that we consider the change-point problem *a posteriori*: a larger n means that we have been able to observe the phenomenon of interest with a finer time discretization. This is similar to the setting presented in Chapter 1. Also note that this asymptotic setting is restrictive in the sense that segments size asymptotically are of order n; we

1. In the context of model selection, we can see D as the *dimension* of the model, hence the notation.

do not make this assumption in our analysis, which also covers asymptotic settings where some segments have a smaller size.

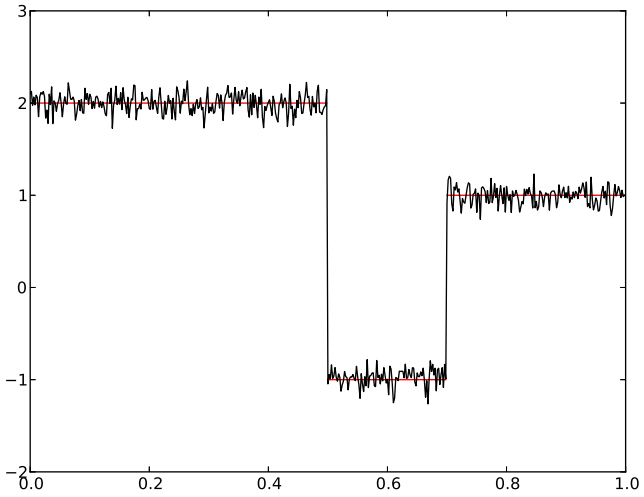

Figure 2-2 – Illustration of the asymptotic setting (Example 2.1) in the case of changes in the mean of the X_i. Here, $\mathcal{X} = \mathbb{R}$, $X_i = f(t_i) + \varepsilon_i$ with $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. and centered, and $f : [0, 1] \to \mathbb{R}$ is a (fixed) piecewise constant function (shown in red). The goal is to recover the number of abrupt changes of f (here, 2) and their locations ($b_1 = 0.5$ and $b_2 = 0.7$). Note that other kinds of changes in the distribution of the X_i can be considered, see Section 4.2.

2.2.2 Kernel change-point procedure

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semi-definite kernel, that is, a measurable function such that the matrix $(k(x_i, x_j))_{1 \leq i,j \leq m}$ is positive semi-definite for any $m \geq 1$ and $x_1, \ldots, x_m \in \mathcal{X}$ [Schölkopf and Smola, 2002]. Let us recall some classical examples of positive semi-definite kernels, some of them already encountered in the Introduction:

— the *linear kernel*: $k_\ell(x, y) = \langle x, y \rangle_{\mathbb{R}^p}$ for $x, y \in \mathcal{X} = \mathbb{R}^p$.

— the *polynomial kernel* of order $\alpha \geq 1$: $k_{\mathrm{P}}(x, y) = \left(\langle x, y \rangle_{\mathbb{R}^p} + 1\right)^\alpha$ for $x, y \in \mathcal{X} = \mathbb{R}^p$.

— the *Gaussian kernel* with bandwidth $\nu > 0$: $k_{\mathrm{G}}(x, y) = \exp[-|x - y|^2 / (2\nu^2)]$ for $x, y \in \mathcal{X} = \mathbb{R}^p$.

— the *Laplace kernel* with bandwidth $\nu > 0$: $k_{\mathrm{L}}(x, y) = \exp[-|x - y| / (2\nu)]$ for $x, y \in \mathcal{X} = \mathbb{R}^p$.

— the χ^2-kernel: $k_{\chi^2}(x, y) = \exp\left(-\frac{1}{p\nu} \sum_{i=1}^p \frac{(x_i - y_i)^2}{x_i + y_i}\right)$ for $x, y \in \mathcal{X} = \Delta_p$ the p-dimensional simplex, and the bandwidth ν is a positive constant.

As done by Harchaoui and Cappé [2007] and Arlot et al. [2012], for a given segmentation $\tau \in \mathcal{T}_n^D$, we assess the adequacy of τ with the *kernel least-squares criterion*

$$\widehat{\mathcal{R}}_n(\tau) := \frac{1}{n}\sum_{i=1}^{n} k\left(X_i, X_i\right) - \frac{1}{n}\sum_{\ell=1}^{D}\left[\frac{1}{\tau_\ell - \tau_{\ell-1}}\sum_{i=\tau_{\ell-1}+1}^{\tau_\ell}\sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} k\left(X_i, X_j\right)\right] . \quad (2.1)$$

Elementary algebra shows that, when $\mathcal{X} = \mathbb{R}^p$ and $k = k_\ell$, $\widehat{\mathcal{R}}_n$ is the usual least-squares criterion. Indeed, if we denote by $\widehat{R}_n(\tau)$ the least-squares criterion in $\mathbb{R}^p$ and set $k\left(x, y\right) = \langle x, y\rangle$,

$$\begin{aligned}
n\widehat{R}_n(\tau) &= \sum_{\ell=1}^{D}\sum_{i=\tau_{\ell-1}+1}^{\tau_\ell}\left\|X_i - \frac{1}{\tau_\ell - \tau_{\ell-1}}\sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} X_j\right\|^2 \\
&= \sum_{\ell=1}^{D}\sum_{i=\tau_{\ell-1}+1}^{\tau_\ell}\left(\|X_i\|^2 - \frac{2}{\tau_\ell - \tau_{\ell-1}}\sum_{j=\tau_{\ell-1}+1}^{\tau_\ell}\langle X_i, X_j\rangle \right. \\
&\qquad\qquad \left. + \frac{1}{\left(\tau_\ell - \tau_{\ell-1}\right)^2}\sum_{j,j'=\tau_{\ell-1}+1}^{\tau_\ell}\langle X_j, X_{j'}\rangle\right) \\
&= \sum_{i=1}^{n}\|X_i\|^2 - \sum_{\ell=1}^{d}\frac{1}{\tau_\ell - \tau_{\ell-1}}\sum_{i,j=\tau_{\ell-1}+1}^{\tau_\ell}\langle X_i, X_j\rangle \\
n\widehat{R}_n(\tau) &= n\,\widehat{\mathcal{R}}_n(\tau)\,.
\end{aligned}$$

Minimizing this criterion over the set of all segmentations always outputs the segmentation with n segments reduced to a point, which can be seen as over-fitting. To counteract this, a classical idea [Lavielle, 2005, for instance] is to minimize a penalized criterion $\mathrm{crit}\left(\tau\right) := \widehat{\mathcal{R}}_n(\tau) + \mathrm{pen}\left(\tau\right)$, where $\mathrm{pen} : \mathcal{T}_n \to \mathbb{R}_+$ is called the penalty. Formally, the kernel change-point procedure of Arlot et al. [2012] selects the segmentation

$$\widehat{\tau} \in \operatorname*{arg\,min}_{\tau\in\mathcal{T}_n}\left\{\mathrm{crit}\left(\tau\right)\right\} \qquad \text{where} \qquad \mathrm{crit}\left(\tau\right) = \widehat{\mathcal{R}}_n(\tau) + \mathrm{pen}\left(\tau\right) . \quad (2.2)$$

In this manuscript, we will use two different classes of penalty functions which were already encountered in Section 1.3.2. Let us recall their definition. The first is a classical choice in model selection, similar to AIC, BIC and C_p criteria. It is proportional to the number of segments and is often called a *linear* penalty. Namely, we consider

$$\mathrm{pen}\left(\tau\right) = \mathrm{pen}_\ell\left(\tau\right) := \frac{CM^2 D_\tau}{n}\,, \quad (2.3)$$

where C is a positive constant and M is specified in Assumption 2.1 later on. This definition coincides with Eq. (1.8) with $\beta_n = CM^2$. Note that, as mentioned in Chapter 1, different penalty shapes can be considered. For instance, as suggested

by Lebarbier [2005], it is also possible to use as a penalty shape

$$\mathrm{pen}\,(\tau) = \mathrm{pen}_{\mathrm{L}}\,(\tau) := \frac{D_\tau}{n}\left(c_1 \log\frac{n}{D_\tau} + c_2\right), \tag{2.4}$$

with c_1 and c_2 positive constants. This definition coincides with Eq. (1.9) up to a variance term. In a very similar fashion, Arlot et al. [2012] advocates for the use of

$$\mathrm{pen}\,(\tau) = \mathrm{pen}_{\mathrm{ACH}}(\tau) := \frac{1}{n}\left(c_1 \log\binom{n-1}{D_\tau - 1} + c_2 D_\tau\right).$$

We will see in Section 2.6 that the oracle inequality obtained in Arlot et al. [2012] is valid for pen_ℓ when the penalty constant C is of order $\log(n)$, as well as $\mathrm{pen}_{\mathrm{L}}$ and $\mathrm{pen}_{\mathrm{ACH}}$.

2.3 The reproducing kernel Hilbert space

Let $\mathcal{H}$ be the RKHS associated to k, together with the canonical feature map $\Phi : \mathcal{X} \to \mathcal{H}$

$$\begin{array}{rccc} \Phi & : & \mathcal{X} & \to & \mathcal{H} \\ & & x & \mapsto & \Phi(x) := k\,(\cdot, x)\ , \end{array}$$

as exposed in Section 1.4. In this section, we explain how to define a "mean element" of $\Phi(X_i)$ belonging to $\mathcal{H}$, and we rewrite the empirical risk.

2.3.1 Kernel mean embedding

Let us write $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ (resp. $\|\cdot\|_{\mathcal{H}}$) for the inner product (resp. the norm) of $\mathcal{H}$. For any $i \in \{1, \ldots, n\}$, define $Y_i := \Phi(X_i) \in \mathcal{H}$. As we have seen, in the case where $k = k_\ell$, then $Y_i = \langle \cdot, X_i \rangle$ and the empirical risk $\widehat{\mathcal{R}}_n$ reduces to the least-squares criterion

$$\widehat{R}_n(\tau) = \frac{1}{n}\sum_{\ell=1}^{D_\tau}\sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \left\|X_i - \overline{X}_\ell\right\|^2,$$

where $\overline{X}_\ell$ is the empirical mean of the X_i over the segment $\{\tau_{\ell-1}+1, \ldots, \tau_\ell\}$. It is well-known that penalized least-squares procedures detect changes in the mean of the observations X_i, see Yao [1988]. Hence the kernelized version of this least-squares procedure, KCP, should detect changes in the "mean" of the $Y_i = \Phi(X_i)$, which are a non-linear transformation of the X_i.

More precisely, assume that $\mathcal{H}$ is separable.[2] Suppose that

$$\forall i \in \{1, \ldots, n\}, \qquad \mathbb{E}\left[\sqrt{k\,(X_i, X_i)}\right] < +\infty\,. \tag{2.5}$$

2. A topological space is called *separable* if it contains a countable dense subset.

Then, for any $i \in \{1, \dots, n\}$, we define $\mu_i^\star$ as the *Bochner integral* of Y_i,

$$\mu_i^\star := \mathbb{E}\left[k\left(X_i, \cdot\right)\right] \in \mathcal{H}\,.$$

We refer to Diestel and Uhl [1977, Chapter 2] and Ledoux and Talagrand [2013] for the proper definition of the Bochner integral. Let us say that the mean element $\mu_i^\star$ can be seen as an embedding of P_{X_i}, the distribution of X_i, in $\mathcal{H}$ — for this reason, the mapping $P \mapsto \mathbb{E}_{X\sim P}\left[k\left(X, \cdot\right)\right]$ is often referred to as the *kernel embedding*, or kernel mean embedding. The kernel embedding can be seen as a generalization of the notion of characteristic function [Muandet et al., 2017].

Note that the condition (2.5) is satisfied in our setting (when either Assumption 2.1 or Assumption 2.2 holds true, see Section 2.5), and $\mathcal{H}$ is separable in most standard cases [Dieuleveut and Bach, 2016]. Hence the mean elements $\mu_i^\star$ will be well-defined in our setting. The Bochner integral commutes with continuous linear operators, hence the following property holds, which will be of use:

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

We now define the "true segmentation" $\tau^\star \in \mathcal{T}_n$ by

$$\begin{gathered}\mu_1^\star = \dots = \mu_{\tau_1^\star}^\star, \quad \mu_{\tau_1^\star+1}^\star = \dots = \mu_{\tau_2^\star}^\star, \quad \dots \quad \mu_{\tau_{D^\star-1}^\star+1}^\star = \dots = \mu_n^\star \\ \text{and} \quad \forall i \in \{1, \dots, D^\star - 1\}, \quad \mu_{\tau_i^\star}^\star \neq \mu_{\tau_{i+1}^\star}^\star\end{gathered} \tag{2.6}$$

with $1 \leq \tau_1^\star < \dots < \tau_{D^\star-1}^\star \leq n$. We call the $\tau_i^\star$s the *true* change-points. It should be clear that it is always possible to define $\tau^\star$, and that the previous display is not an assumption we make.

A kernel is said to be characteristic if the mapping $P \mapsto \mathbb{E}_{X\sim P}\left[\Phi(X)\right]$ is injective, for P belonging to the set of Borel probability measures on $\mathcal{X}$ [Fukumizu et al., 2004, 2008]. In simpler terms, when k is a characteristic kernel, X_i and X_{i+1} have the same distribution if and only if $\mu_i^\star = \mu_{i+1}^\star$, and $\tau^\star$ indeed corresponds to the set of changes in the distribution of the X_i. For instance, all integrally [3] positive definite kernels are characteristic, including the Gaussian kernel, see Sriperumbudur et al. [2010]. Therefore, in the setting of Example 2.1, for n large enough, $D^\star = K + 1$ and $\tau_\ell^\star = \lfloor n b_\ell \rfloor$ for $\ell = 1, \dots, K$.

For a general kernel, some changes of P_{X_i}, the distribution of X_i, might not appear in $\tau^\star$. For instance, with the linear kernel, $\tau^\star$ only corresponds to changes of the mean of the X_i. In most cases, a characteristic kernel is known and we can choose to use KCP with a characteristic kernel; then, as we prove in Chapter 3, KCP eventually detects any change in the distribution of the observations. But one can also choose a non-characteristic kernel on purpose, hence focusing only on some changes in the distribution of the X_i. For instance, the polynomial kernel of order d is not characteristic and leads to the detection of changes in the first d moments of the distribution; with the linear kernel, KCP detects changes in the mean of the X_i.

3. A measurable, symmetric and bounded function k is said to be *integrally* positive definite if $\iint_{\mathcal{X}} k\left(x, y\right) \mathrm{d}\,\mu(x)\,\mathrm{d}\,\mu(y) > 0$ for any finite signed Borel measure μ on $\mathcal{X}$.

From now on, we focus on the problem of detecting the changes of $\tau^\star$ only, whether the kernel is characteristic or not.

2.3.2 Rewriting the empirical risk

It is convenient to see the images of the observations by the feature map as an element of $\mathcal{H}^n$. To this extent, we define $Y := (Y_1, \dots, Y_n)$, as well as $\mu^\star := (\mu_1^\star, \dots, \mu_n^\star) \in \mathcal{H}^n$ and $\varepsilon := Y - \mu^\star \in \mathcal{H}^n$. We identify the elements of $\mathcal{H}^n$ with the set of applications $\{1, \dots, n\} \to \mathcal{H}$, naturally embedded with the inner product and norm [4] given by

$$\forall x, y \in \mathcal{H}^n, \qquad \langle x, y \rangle := \sum_{i=1}^{n} \langle x_i, y_i \rangle_{\mathcal{H}} \qquad \text{and} \qquad \|x\|^2 := \sum_{j=1}^{n} \|x_j\|_{\mathcal{H}}^2 \,.$$

We now rewrite the empirical risk as a function of τ and Y. For any segmentation $\tau \in \mathcal{T}_n$, define F_τ the set of applications $\{1, \dots, n\} \to \mathcal{H}$ that are constant over the segments of τ. We see F_τ as a subspace of $\mathcal{H}^n$ as a vector space. Take $f \in \mathcal{H}^n$, we define $\Pi_\tau f$ the orthogonal projection of f onto F_τ with respect to $\|\cdot\|$:

$$\Pi_\tau f \in \operatorname*{arg\,min}_{g \in F_\tau} \|f - g\| \,.$$

It comes without much surprise that $\Pi_\tau f$ can be computed as in the real case and, in addition, is also equal to the piece-wise constant function whose values are the empirical mean of f on each segment of τ. Namely, for any $f \in \mathcal{H}^n$ and any $\ell \in \{1, \dots, D_\tau\}$,

$$\forall i \in \{\tau_{\ell-1} + 1, \dots, \tau_\ell\}, \qquad (\Pi_\tau f)_i = \frac{1}{|\tau_\ell - \tau_{\ell-1}|} \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} f_j \,. \tag{2.7}$$

We give the proof of Eq. (2.7) as found in Arlot et al. [2012], which is another illustration of the kernel trick.

Proof. Define $\lambda_\ell := \{\tau_{\ell-1} + 1, \tau_\ell\}$ the ℓ-th segment of τ. For any $g \in F_\tau$, denote by g_{λ_ℓ} the value of g on λ_ℓ and

$$\widetilde{g}_{\lambda_\ell} := \frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{i \in \lambda_\ell} g_i \,.$$

Then we can decompose $\|f - g\|$ as

$$\|f - g\|^2 = \sum_{\ell=1}^{D_\tau} \sum_{i \in \lambda_\ell} \|f_{\lambda_\ell} - \widetilde{g}_{\lambda_\ell}\|_{\mathcal{H}}^2 + \|g_i - \widetilde{g}_{\lambda_\ell}\|_{\mathcal{H}}^2 + 2\langle f_{\lambda_\ell} - \widetilde{g}_{\lambda_\ell}, g_i - \widetilde{g}_{\lambda_\ell} \rangle_{\mathcal{H}}$$

4. Note that we slightly abuse the notations $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$, previously defined for elements of $\mathbb{R}^n$. This should create no confusion.

$$= \sum_{\ell=1}^{D_\tau} (\tau_\ell - \tau_{\ell-1}) \left\| f_{\lambda_\ell} - \widetilde{g}_{\lambda_\ell} \right\|_{\mathcal{H}}^2 + \sum_{\ell=1}^{D_\tau} \sum_{i \in \lambda_\ell} \left\| g_i - \widetilde{g}_{\lambda_\ell} \right\|_{\mathcal{H}}^2 ,$$

since $\sum_{i \in \lambda_\ell} (g_i - \widetilde{g}_{\lambda_\ell}) = 0$. Therefore, $\|f - g\|^2$ is minimal over $f \in F_\tau$ in $\widetilde{g}$, and Eq. (2.7) is proved. □

Thanks to Eq. (2.7), we are now able to write the empirical risk as

$$\widehat{\mathcal{R}}_n(\tau) = \frac{1}{n} \|Y - \widehat{\mu}_\tau\|^2 = \frac{1}{n} \sum_{\ell=1}^{D_\tau} \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \|Y_i - (\widehat{\mu}_\tau)_i\|_{\mathcal{H}}^2 , \tag{2.8}$$

where $\widehat{\mu}_\tau = \Pi_\tau Y$, following [Harchaoui and Cappé, 2007; Arlot et al., 2012].

2.4 Algorithmic aspects of KCP

In this section we discuss the implementation of the minimization problem (2.2).

As we explained before, one of the consequences of the kernel trick is that the KCP algorithm needs only the Gram matrix K to run. Hence the first step in KCP is the computation of K, which typically needs $\mathrm{O}(n^2)$ computations and $\mathrm{O}(n^2)$ storage space. Assuming that we dispose of a function `computeKernel` which can compute $k(x, y)$ for $x, y \in \mathcal{X}$, the calculation of K is straightforward and is given by Algorithm 2.1. Note that some non-negligible constant factors can be hidden in the $\mathrm{O}(\cdot)$ notation, in particular regarding the computational cost. For instance, consider $k = k_{\mathrm{G}}$ the Gaussian kernel. Then each evaluation of $k(x, y)$ requires to compute $\|x - y\|^2$, that has a cost proportional to the dimension of the data. Suppose that $\mathcal{X} = \mathbb{R}^p$, then the true computational cost of Algorithm 2.1 is $\mathrm{O}(pn^2)$.

Algorithm 2.1 Computation of the Gram matrix

```
procedure ComputeGramMatrix(x)
    n ← length of x
    K ← zeros(n, n)
    for i = 1 : n do
        for j = 1 : n do
            K(i, j) ← computeKernel(x(i), x(j))
        end for
    end for
    return K
end procedure
```

KCP without penalty term

We now explain how problem (2.2) can be solved thanks to *dynamic programming* [Bellman, 1961]. To begin with, we solve (2.2) for a pre-defined number of

segments D, without adding a penalty term, as in Harchaoui and Cappé [2007]. For any $1 \leq a < b \leq n$, we define $c\left(a:b\right)$ the cost of segment $\left\{a+1,\ldots,b\right\}$, that is

$$c\left(a:b\right) := \sum_{i=a+1}^{b} \left\| Y_i - \left(\bar{Y}_{a:b}\right)_i \right\|_{\mathcal{H}}^2 .$$

We also define $\text{cost}\left(D,t\right)$ the optimal cost of segmenting our signal up to time t in D segments, *i.e.*,

$$\text{cost}\left(D,t\right) := \inf_{\tau \in \mathcal{T}_t^D} \left\{ \frac{1}{n} \sum_{\ell=1}^{D} \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \left\| Y_i - \left(\widehat{\mu}_\tau\right)_i \right\|_{\mathcal{H}}^2 \right\} = \inf_{\tau \in \mathcal{T}_t^D} \left\{ \frac{1}{n} \sum_{\ell=1}^{D} c\left(\tau_{\ell-1}:\tau_\ell\right) \right\} .$$

Notice that $\text{cost}\left(D,n\right) = n\,\widehat{\mathcal{R}}(\widehat{\tau}(D))$ is the quantity we want to obtain. Now the key observation is that $\text{cost}\left(D,t\right)$ is additive:

$$\begin{aligned} \text{cost}\left(D,t\right) &= \inf_{\tau \in \mathcal{T}_t^D} \left\{ \sum_{\ell=1}^{D-1} c\left(\tau_{\ell-1}:\tau_\ell\right) + c\left(\tau_{D-1}:t\right) \right\} \\ &= \inf_{s \leq t} \left\{ \text{cost}\left(D-1,s\right) + c\left(s:t\right) \right\} . \end{aligned}$$

Assuming that we dispose of a matrix `S` of size $n \times n$ that contains $c\left(s:t\right)$ for every values of s,t with $0 \leq s < t \leq n$, this gives rise to Algorithm 2.2, a dynamic programming scheme for computing the optimal cost. Of course, a simple backtracking procedure allows to recover the corresponding segmentation. The computational cost of running `CostMatrix` is

$$\sum_{d=2}^{D} \sum_{t=d}^{n} (t+1) = \mathrm{O}\left(n^2 D\right) .$$

As for the memory space, it is dominated by the cost matrix and the index matrix, of size $\mathrm{O}\left(n^2 D\right)$. The `BackTracking` routine, on the other hand, runs in $\mathrm{O}\left(D\right)$.

It is important to understand that computing n^2 entries of the matrix `S` is in fact a byproduct of the computation of the Gram matrix. Indeed, for any $1 \leq s < t \leq n$, $c\left(s:t\right)$ can be expressed as a function of the cumulative sum of K. More precisely, we already noticed that

$$c\left(s:t\right) = \sum_{i=s+1}^{t} k\left(X_i, X_i\right) - \frac{1}{t-s} \sum_{i=s+1}^{t} \sum_{j=s+1}^{t} k\left(X_i, X_j\right) .$$

Algorithm 2.2 KCP dynamic programming step with given segment costs

procedure CostMatrix(S,D)
  $n \leftarrow$ length of $S(:, 1)$
  $C \leftarrow \text{zeros}(D, n)$ ▷ cost matrix
  $I \leftarrow \text{zeros}(D, n)$ ▷ index matrix
  for $t = 1 : n$ **do**
    $C(1, t) \leftarrow S(0, t)$
  end for
  for $d = 2 : D$ **do**
    for $t = d : n$ **do**
      $[I(d, t), C(d, t)] \leftarrow \min_{s \leq t} \{C(d - 1, s) + S(s, t)\}$
    end for
  end for
  output I, C
end procedure
procedure BackTracking(C,I,D)
  $\widehat{\tau} \leftarrow \text{zeros}(D)$ ▷ estimated segmentation
  $p \leftarrow n$ ▷ current position
  for $d = 1 : D$ **do**
    $p \leftarrow I(D - d + 1, p)$
  end for
  return $\widehat{\tau}$
end procedure

Define Γ the cumulative sum and T the cumulative trace of the matrix K, *i.e.*,

$$\forall 1 \leq s < t \leq n, \qquad \Gamma_{s,t} := \sum_{i=1}^{s} \sum_{j=1}^{t} k\left(X_i, X_j\right) \qquad \text{and} \qquad T_s := \sum_{i=1}^{s} k\left(X_i, X_i\right) .$$

Notice that Γ and T can be computed at the same time than K, in $\mathrm{O}\left(n^2\right)$ operations. Then

$$c\left(s:t\right) = T_t - T_s - \frac{1}{t-s}\left(\Gamma_{t,t} - 2\Gamma_{s,t} + \Gamma_{s,s}\right) .$$

Since k is a symmetric function, Γ is a symmetric matrix, and the previous computation requires only 5 reads in the matrices Γ and T and simple arithmetic operations. Let us define the following auxiliary procedure.

Algorithm 2.3 Computation of the segment costs

procedure SEGMENTCOSTS(x,K)
  $n \leftarrow$ length of x
  $\Gamma \leftarrow$ CumSum(K) ▷ cumulative sum of the Gram matrix
  $T \leftarrow$ CumSum(diag (K)) ▷ cumulative sum of the trace matrix
  $S \leftarrow \mathrm{zeros}(n, n)$
  for $t = 1 : n$ **do**
    $S(0, t) \leftarrow T(t) - \Gamma(t, t)/t$
  end for
  for $t = 2 : n$ **do**
    for $s = 1 : n - 1$ **do**
      $S(s, t) \leftarrow T(t) - T(s) - \frac{1}{t-s}\left(\Gamma(t, t) - 2\Gamma(s, t) + \Gamma(s, s)\right)$
    end for
  end for
  return S
end procedure

We are now ready to write down the kernel change-point algorithm without penalty term, Algorithm 2.4.

Algorithm 2.4 KCP algorithm without a penalty term

procedure KCPNOPENALTY(x,D)
  $n \leftarrow$ length of x
  K $\leftarrow$ ComputeGramMatrix(x)
  $S \leftarrow$ SegmentCosts(x, K)
  $[I, C] \leftarrow$ CostMatrix(S, D)
  $\widehat{\tau} \leftarrow$ BackTracking(C, I, D)
  return $\widehat{\tau}$
end procedure

Adding the penalty

Let us go back to the original problem (2.2), with a penalty function pen depending only on the number of segments. We assume that a function `ComputePenalty` is provided that takes care of the computation of $\text{pen}(\tau) = \text{pen}(D_\tau)$. We can decompose problem (2.2) as follows: First, we compute the optimal least-squares criterion for any $1 \leq D \leq D_{\max}$, that is

$$\forall 1 \leq D \leq D_{\max}, \qquad \widehat{\tau}(D) \in \underset{\tau \in \mathcal{T}_n^D}{\arg\min} \left\| Y - \widehat{\mu}_\tau \right\|^2 , \tag{2.9}$$

which is solved with the same scheme used by Algorithm 2.4. Second, we choose

$$\widehat{D} \in \underset{1 \leq D \leq D_{\max}}{\arg\min} \left\{ \left\| Y - \widehat{\mu}_{\widehat{\tau}(D)} \right\|^2 + \text{pen}\left(\widehat{\tau}(D)\right) \right\}, \tag{2.10}$$

and third we set $\widehat{\tau} = \widehat{\tau}(\widehat{D})$. From the algorithmic point of view, the cost matrix and index matrix outputted by the `CostMatrix` function contain all the relevant information to solve (2.9). Hence we just have to add the penalty term to the cost matrix after it is computed, and then select the number of segments for $\widehat{\tau}$ by minimizing the new criteria as in (2.10). This modification of Algorithm 2.4 is called Algorithm 2.5 and has the same computational complexity as Algorithm 2.4 for $D_{\max}$ segments, that is $\text{O}\left(D_{\max} n^2\right)$.

Algorithm 2.5 KCP algorithm with a penalty term

procedure KCPWITHPENALTY(x,$D_{\max}$)
  $n \leftarrow$ length of x
  K $\leftarrow$ ComputeGramMatrix(x)
  $S \leftarrow$ SegmentCosts(x, K)
  $[I, C] \leftarrow$ CostMatrix($S, D_{\max}$)
  for $d = 1 : D_{\max}$ **do**
    $C(d, n) \leftarrow C(d, n) +$ ComputePenalty(d) ▷ modifying the last column of C
  end for
  $[\widehat{D}, \sim] \leftarrow \min C(:, n)$
  $\widehat{\tau} \leftarrow$ BackTracking($C, I, \widehat{D}$)
  return $\widehat{\tau}$
end procedure

In definitive, the total computational cost for solving Problem 2.2 is thus

$$\text{O}\left(\left(c_k + D_{\max}\right) n^2\right) ,$$

where c_k is the cost of an evaluation of $k(x, y)$. We want to emphasize that this is prohibitive for sample size larger than 10^5. A path worth exploring to reduce this computational burden is to use a low-rank approximation of the Gram matrix instead of K as suggested by Celisse et al. [2016]. If we let r denote the rank of the approximation, the complexity of the dynamic programming part of the algorithm

then drops to $\mathrm{O}\left(r^2 D_{\max}\right)$. Note that, however, there is no theoretical guarantees for the approximation of the correct solution obtained with this method to the best of our knowledge.

2.5 Assumptions

We now precise the framework under which we are going to study the kernel change-point detection procedure in the next chapter.

A key ingredient of our analysis is the concentration of ε. Intuitively, the performance of KCP is better when ε concentrates strongly around its mean, since without noise we are just given the task to segment a piecewise-constant signal. It is thus natural to make assumptions on ε in order to obtain concentration results. We actually formulate assumptions on the kernel k, which translate automatically onto ε.

As in Arlot et al. [2012], the main hypothesis used in our analysis is the following.

Assumption 2.1. A positive constant M exists such that

$$\forall i \in \{1, \dots, n\}, \qquad k\left(X_i, X_i\right) \leq M^2 < +\infty \qquad \text{a.s.}$$

A simple, yet useful remark is the following: If Assumption 2.1 holds true,

$$\forall i \in \{1, \dots, n\}, \qquad \|Y_i\|_{\mathcal{H}} = \sqrt{k\left(X_i, X_i\right)} \leq M \qquad \text{a.s.}$$

and $\|\varepsilon_i\|_{\mathcal{H}} \leq 2M$ almost surely. Indeed, for any i,

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

Hence $\|\mu_i^\star\|_{\mathcal{H}}^2 \leq \mathbb{E}\left[k\left(X_i, X_i\right)\right] \leq M^2$, and the triangle inequality yields

$$\|\varepsilon_i\|_{\mathcal{H}} \leq \|Y_i\|_{\mathcal{H}} + \|\mu_i^\star\|_{\mathcal{H}} \leq 2M\,. \tag{2.11}$$

This is the reason why we sometimes refer to Assumption 2.1 as a *bounded noise* rather than bounded kernel.

Assumption 2.1 is satisfied for a large class of commonly used kernels, such as the Gaussian, Laplace and χ^2 kernels; $M = 1$ in these three examples.

Note that Assumption 2.1 is weaker than assuming k to be bounded — that is, $k(x, x) \leq M$ for any $x \in \mathcal{X}$, which is equivalent to $k\left(x, x'\right) \leq M$ for any $x, x' \in \mathcal{X}$ since k is positive semi-definite. For instance, if $\mathcal{X} = \mathbb{R}^p$ and the data X_i are bounded almost surely, Assumption 2.1 holds true for the linear kernel and all polynomial kernels, which are not bounded on $\mathbb{R}^p$.

In the setting of Example 2.1, Assumption 2.1 holds true when

$$\forall j \in \{1, \dots, K\}, \qquad k\left(x, x\right) \leq M^2 \qquad \text{for} \quad P_j\text{-a.e.} \quad x \in \mathcal{X}\,.$$

It is sometimes possible to weaken Assumption 2.1 into a finite variance assumption.

Assumption 2.2. A positive constant $V < +\infty$ exists such that

$$\max_{1 \leq i \leq n} \mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{H}}^2\right] \leq V.$$

Since $v_i := \mathbb{E}\left[\|\varepsilon_i\|_{\mathcal{H}}^2\right] = \mathbb{E}\left[k\left(X_i, X_i\right)\right] - \|\mu_i^\star\|_{\mathcal{H}}^2$, Assumption 2.2 holds true when

$$\forall i \in \big\{1, \ldots, n\big\}, \qquad \mathbb{E}\left[k\left(X_i, X_i\right)\right] \leq V\,.$$

As a consequence, Assumption 2.1 implies Assumption 2.2 with $V = M^2$. Note that Assumption 2.2 is satisfied for the polynomial kernel of order d provided that the observations satisfy a moment assumption, namely

$$\forall i \in \big\{1, \ldots, n\big\}, \qquad \mathbb{E}\big[\|X_i\|^{2d}\big] < +\infty\,.$$

In the setting of Example 2.1, Assumption 2.2 holds true with

$$V = \max_{1 \leq \ell \leq K+1} \mathbb{E}_{X \sim P_\ell}\left[k\left(X, X\right)\right]\,,$$

provided this maximum is finite.

2.6 An oracle inequality for KCP

In this section, we recall briefly the oracle inequality obtained by Arlot et al. [2012]. This is not exactly a result on change-point estimation, but a guarantee on estimation of the "mean" of the time series in the RKHS associated with the kernel chosen.

Let us define the quadratic risk of any $\mu \in \mathcal{H}^n$ as an estimator of $\mu^\star$ by

$$\mathcal{R}\left(\mu\right) := \frac{1}{n}\left\|\mu - \mu^\star\right\|^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\mu_i - \mu_i^\star\right\|_{\mathcal{H}}^2\,. \tag{2.12}$$

Then the following holds.

Theorem 2.1 (Arlot et al. [2012], Theorem 2)**.** *Let C be a non-negative constant. Assume that Assumption 2.1 holds true and that* pen $: \mathcal{T}_n \to \mathbb{R}$ *is some penalty function satisfying*

$$\forall \tau \in \mathcal{T}_n, \qquad \mathrm{pen}\left(\tau\right) \geq \frac{CM^2}{n}\left[D_\tau + \log\binom{n-1}{D_\tau - 1}\right]\,. \tag{2.13}$$

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

$$\widehat{\tau} \in \operatorname*{arg\,min}_{\tau \in \mathcal{T}_n}\big\{\widehat{\mathcal{R}}_n(\tau) + \mathrm{pen}\left(\tau\right)\big\}\,,$$

we have

$$\mathcal{R}\left(\widehat{\mu}_{\widehat{\tau}}\right) \leq 2 \inf_{\tau \in \mathcal{T}_n} \left\{ \mathcal{R}\left(\widehat{\mu}_{\tau}\right) + \operatorname{pen}\left(\tau\right) \right\} + \frac{83 y M^2}{n} \,. \tag{2.14}$$

Informally speaking, Eq. (2.14) means that $\widehat{\mu}_{\widehat{\tau}}$ estimates well the mean $\mu^\star \in \mathcal{H}^n$ of the time series $Y_1, \ldots, Y_n$. More precisely, Theorem 2.1 states that, with high probability, the quadratic risk of the least-squares estimator built on $\widehat{\tau}$ is of the same order than the quadratic risk of the estimator associated to any segmentation, up to a penalty term and a remainder term.

Clearly, Theorem 2.1 applies to $\widehat{\tau}$ given by (2.2) when $\operatorname{pen} = \operatorname{pen}_{\mathrm{ACH}}$.

If both c_1, c_2 are greater than $2L_1M^2$, Theorem 2.1 also applies to $\widehat{\tau}$ given by (2.2) when $\operatorname{pen} = \operatorname{pen}_{\mathrm{L}}$. Indeed, for any $D \in \left\{1, \ldots, n\right\}$,

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

where the last inequality is standard [see Prop. 2.5 in Massart, 2007, for a stronger result].

A linear penalty pen_ℓ also satisfies Eq. (2.13). Indeed, from the last display we deduce that

$$\forall n \geq 1, \qquad \max_{1 \leq D \leq n} \frac{1}{D} \binom{n-1}{D-1} \leq \log(n) + 1 \,.$$

As a consequence, for $C \gg \left(\log(n)\right)^{-1}$, $\operatorname{pen}_\ell\left(\tau\right) \geq \frac{CM^2}{n} \left(D_\tau + \log\left(\binom{n-1}{D_\tau - 1}\right)\right)$ and Theorem 2.1 holds for a linear penalty.

Chapter 3

Consistency of kernel change-point detection

Abstract

In this chapter, we show that, with high probability, KCP with a linear penalty function recovers the correct number of change-points, provided that the penalty constant is well-chosen. In addition, we prove that KCP estimates the change-points at the optimal rate. As a consequence, when using a characteristic kernel, KCP detects all kinds of change in the distribution (not only changes in the mean or the variance), and it is able to do so for complex structured data (not necessarily in $\mathbb{R}^p$). A key point of the proof is a concentration inequality of a quadratic form in a Hilbert space, a refinement of a result obtained in the first version of Arlot et al. [2012]. We also show an analogous statement for a different penalty function. Both these results are proved under a boundedness assumption; we prove slightly weaker results under a finite variance assumption. We also discuss in detail the various notions of distances between segmentations, and show that they are all equivalent when one of them is small enough compared to the size of the smallest segment. This chapter is based upon the article Garreau and Arlot [2016], under submission to the *Electronic Journal of Statistics*.

3.1 Introduction

At this stage, some key theoretical questions remain open: does KCP estimate correctly the number of change-points and their locations with a large probability? If yes, can we prove a consistency result similar to those introduced in Section 1.3.2?

This chapter answers these questions, showing that KCP has good theoretical properties for change-point estimation with independent data, under a boundedness assumption. (Theorem 3.1, stated for a linear penalty, and Theorem 3.2, for $\mathrm{pen}_{\mathrm{L}}$). These results are non-asymptotic, hence meaningful for high-dimensional or complex data. In the asymptotic setting — with a fixed true segmentation and more and more data points observed within each segment —, Theorem 3.1 implies that KCP

estimates consistently all changes in the "kernel mean" of the distribution of data, at speed $\log(n)/n$ with respect to the sample size n. Since we make no assumptions on the minimal size of the true segments, this matches minimax lower bounds [Brunel, 2014]. We also provide a partial result under a weaker finite variance assumption (Theorem 3.3 in Section 3.2.3) and explain in Section 3.3 how our proofs could be extended to other settings, including the dependent case. These findings are illustrated by numerical simulations in Section 4.2.

An important case is when KCP is used with a characteristic kernel [Fukumizu et al., 2004, 2008], such as the Gaussian or the Laplace kernel. Then, any change in the distribution of data induces a change in the "kernel mean". So, Theorem 3.1 implies that KCP then estimates consistently and at the minimax rate *all changes* in the distribution of the data, without any parametric assumption and without prior knowledge about the number of changes.

Our results also are interesting regarding to the theoretical understanding of least-squares change-point procedures. Indeed, when KCP is used with the linear kernel, it reduces to previously known penalized least-squares change-point procedures [Yao, 1988; Comte and Rozenholc, 2004; Lebarbier, 2005, for instance]. There are basically two kinds of results on such procedures in the change-point literature: (i) asymptotic statements on change-point estimation [Yao, 1988; Yao and Au, 1989; Bai and Perron, 1998; Lavielle and Moulines, 2000] and (ii) non-asymptotic oracle inequalities [Comte and Rozenholc, 2004; Lebarbier, 2005; Arlot et al., 2012], which are based upon concentration inequalities and model selection theory [Birgé and Massart, 2001] but do not directly provide guarantees on the estimated change-point locations. Our results and their proofs show how to conciliate the two approaches when we are interested in change-point locations, which is already new for the case of the linear kernel, and also holds for a general kernel.

Section 3.2 is dedicated to the exposition of these results. They are discussed in Section 3.3, and the proofs are collected in Section 3.4. Section 3.5 contains the proofs of technical lemmas needed in Section 3.4.

3.2 Theoretical guarantees for KCP

We state our main results in this section, which is divided as follows. In Section 3.2.1, we state Theorem 3.1, which provides simple conditions under which KCP recovers the correct number of segments and localizes the true change-points with high probability, under the bounded kernel Assumption 2.1. This is our main result. Theorem 3.1 concerns KCP with a linear penalty: we state in the same section a result for a different penalty, Theorem 3.2. Section 3.2.2 contains a review of the classical losses between segmentations which can be considered in addition to the one used in Theorem 3.1 and 3.2. Corollary 3.1 formulates a result on $\widehat{\tau}$ in terms of the Frobenius loss. Finally, Section 3.2.3 states a partial result on KCP — requiring the number of change-points $D^\star$ to be known — under the weaker Assumption 2.2.

3.2.1 Consistency under bounded kernel assumption

We first need to define some quantities. The size of the smallest jump of $\mu^\star$ in $\mathcal{H}$ is defined by

$$\underline{\Delta} := \min_{i \,/\, \mu_i^\star \neq \mu_{i+1}^\star} \left\|\mu_i^\star - \mu_{i+1}^\star\right\|_{\mathcal{H}} . \tag{3.1}$$

Intuitively, the higher $\underline{\Delta}$ is, the easier it is to detect the smallest jump with our procedure. The quantity $\left\|\mu_i^\star - \mu_{i+1}^\star\right\|_{\mathcal{H}}$ coincides with the MMD between the distributions of X_i and X_{i+1}. In the scalar setting (with the linear kernel), the ratio $\underline{\Delta}/\sigma$ (where σ^2 is the variance of the noise) is called the *signal-to-noise ratio* [Basseville and Nikiforov, 1993] and is often used as a measure of the magnitude of a change in the signal. In Example 2.1,

$$\underline{\Delta} = \min_{1 \leq j \leq K} \left\|\mu_{P_j}^\star - \mu_{P_{j+1}}^\star\right\|_{\mathcal{H}} ,$$

where $\mu_{P_j}^\star$ denotes the (Bochner) expectation of $\Phi(X)$ when $X \sim P_j$.

For any $\tau \in \mathcal{T}_n$, we denote the (normalized) sizes of its smallest and of its largest segment by

$$\underline{\Lambda}_\tau := \frac{1}{n} \min_{1 \leq \ell \leq D_\tau} |\tau_\ell - \tau_{\ell-1}| \qquad \text{and} \qquad \overline{\Lambda}_\tau := \frac{1}{n} \max_{1 \leq \ell \leq D_\tau} |\tau_\ell - \tau_{\ell-1}| . \tag{3.2}$$

It should be clear that the smaller $\underline{\Lambda}_{\tau^\star}$ is, the harder it is to detect the segment that achieves the minimum in (3.2). For instance, in the particular case of Example 2.1,

$$\underline{\Lambda}_{\tau^\star} \xrightarrow[n \to +\infty]{} \min_{0 \leq j \leq K} |b_{j+1} - b_j| \qquad \text{and} \qquad \overline{\Lambda}_{\tau^\star} \xrightarrow[n \to +\infty]{} \max_{0 \leq j \leq K} |b_{j+1} - b_j| .$$

Finally, for any τ^1 and $\tau^2 \in \mathcal{T}_n$, we define

$$\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right) := \max_{1 \leq i \leq D_{\tau^1} - 1} \left\{ \min_{1 \leq j \leq D_{\tau^2} - 1} \left|\tau_i^1 - \tau_j^2\right| \right\},$$

which is a loss function (a measure of dissimilarity) between the segmentations τ^1 and τ^2. Note that $\mathrm{d}_\infty^{(1)}$ is not a distance; other possible losses between segmentations and their relationship with $\mathrm{d}_\infty^{(1)}$ are discussed in Section 3.2.2.

We are now able to state our main result.

Theorem 3.1. *Suppose that Assumption 2.1 holds true. For any $y > 0$, an event Ω of probability at least $1 - \mathrm{e}^{-y}$ exists on which the following holds true. For any $C > 0$, let $\widehat{\tau}$ be defined by*

$$\widehat{\tau} \in \operatorname*{arg\,min}_{\tau \in \mathcal{T}_n} \left\{ \mathrm{crit}\,(\tau) \right\} \qquad \textit{where} \qquad \mathrm{crit}\,(\tau) = \widehat{\mathcal{R}}_n(\tau) + \mathrm{pen}\,(\tau) , \tag{2.2}$$

with $\mathrm{pen} = \mathrm{pen}_\ell$ *defined by*

$$\mathrm{pen}_\ell\,(\tau) := \frac{C M^2 D_\tau}{n} . \tag{2.3}$$

Set

$$C_{\min} := \frac{74}{3}(D^\star + 1)(y + \log(n) + 1) \quad \textit{and} \quad C_{\max} := \frac{\underline{\Delta}^2}{M^2}\frac{\underline{\Lambda}_{\tau^\star}}{6D^\star}n\,.$$

Then, if

$$C_{\min} < C < C_{\max}\,, \tag{3.3}$$

on Ω*, we have*

$$\widehat{D} = D^\star \qquad \textit{and} \qquad \frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^\star, \widehat{\tau}\right) \leq v_1(y) := \frac{148D^\star M^2}{\underline{\Delta}^2} \cdot \frac{y + \log(n) + 1}{n}\,.$$

We delay the proof of Theorem 3.1 to Section 3.4.4. Some remarks follow.

Theorem 3.1 is a non-asymptotic result: it is valid for any $n \geq 1$ and there is nothing hidden in $\mathrm{o}\,(1)$ remainder terms. The latter point is crucial for complex data — for instance, $\mathcal{X} = \mathbb{R}^p$ with $p > n$ — since in this case, assuming $\mathcal{X}$ fixed while $n \to +\infty$ is not realistic.

Nevertheless, it is useful to write down what Theorem 3.1 becomes in the asymptotic setting of Example 2.1. As previously noticed, $D^\star$, $\underline{\Lambda}_{\tau^\star}$, $\underline{\Delta}^2$ and M^2 then converge to positive constants as $n \to +\infty$. Therefore, $C_{\min}$ is of order $\log(n)$, $C_{\max}$ is of order n and we always have $C_{\min} < C_{\max}$ for n large enough. The upper bound on C matches classical asymptotic conditions for variable selection [Shao, 1997]. The necessity of taking C of order at least $\log(n)$ is shown by Birgé and Massart [2007] in a variable selection setting, which includes change-point detection as a particular example; Birgé and Massart [2007]; Abramovich et al. [2006] provide several arguments for the optimality of taking a constant C of order $\log(n)$. When C satisfies Eq. (3.3), the result of Theorem 3.1 implies that $\mathbb{P}\left(\widehat{D} = D^\star\right) \to 1$. For the linear kernel in $\mathbb{R}^d$, this is a well-known result when the distribution of the X_i changes only through its mean. The first result dates back to Yao [1988, Section 2] for a Gaussian noise, later extended by Liu et al. [1997] and Bai and Perron [1998, Section 3.1] under mixingale hypothesis on the error, and Lavielle and Moulines [2000] under very mild assumptions satisfied for a large family of zero-mean processes [for the precise statement of the hypothesis, see Lavielle and Moulines, 2000, Section 2.1]. Theorem 3.1 also shows that the normalized estimated change-points of $\widehat{\tau}$ converge towards the normalized true change-points at speed at least $\log(n)/n$.

Up to a logarithmic factor, this speed matches the minimax lower bound n^{-1} which has been obtained previously for various change-point procedures [Korostelev, 1988; Boysen et al., 2009; Korostelev and Tsybakov, 2012, for instance] including least-squares [Lavielle and Moulines, 2000], assuming that $\underline{\Lambda}_{\tau^\star} \geq \kappa > 0$. When $D^\star \geq 3$ and the assumption on $\underline{\Lambda}_{\tau^\star}$ is removed —that is, segments of length much smaller than n are allowed, which is compatible with Theorem 3.1 since it is non-asymptotic—, Brunel [2014, Theorem 6] shows a minimax lower bound of order $\log(n)/n$. Therefore, in this setting, KCP achieves the minimax rate. We do not know whether KCP remains minimax optimal (without the log factor) under the assumption $\underline{\Lambda}_{\tau^\star} \geq \kappa > 0$.

Note finally that KCP also performs well for finite samples, according to the simulation experiments of Arlot et al. [2012].

Theorem 3.1 emphasizes the key role of $\underline{\Delta}^2/M^2$, which can be seen as a generalization of the signal-to-noise ratio, for the change-point detection performance of KCP. The larger is this ratio, the easier it is to have Eq. (3.3) satisfied and the smaller is $v_1(y)$. This suggests to choose k (theoretically at least) by maximizing $\underline{\Delta}^2/M^2$, as we discuss in Section 3.3, and later in Section 4.3. Note that $\underline{\Delta}^2/M^2$ is invariant by a rescaling of k, hence the result of Theorem 3.1 is unchanged when k is rescaled.

The hypothesis in Eq. (3.3) is actually three-fold. First, we use that $C > C_{\min}$ to get $\widehat{D} \leq D^\star$. We have to assume C large enough since a too small penalty leads to selecting (with KCP or any other penalized least-squares procedure) the segmentation with n segments, that is $\widehat{D} = n$. Second, $C < C_{\max}$ is used to get $\widehat{D} \geq D^\star$. Such an assumption is required since taking a penalty function too large in Eq. (2.2) would result in selecting the segmentation with only one segment, that is, $\widehat{D} = 1$. Third, $C_{\max}$ has to be greater than $C_{\min}$ for providing a non-empty interval of possible values for C. This inequality is also used in the proof of the upper bound on $\mathrm{d}_\infty^{(1)}(\tau^\star, \widehat{\tau})$ when we already know that $\widehat{D} = D^\star$. In Example 2.1, the $C_{\min} < C_{\max}$ hypothesis translates into $\underline{\Lambda}_{\tau^\star} \succ \log(n)/n$. That is, the size of the smallest segment has to be of order $\log(n)/n$. This is known to be a necessary condition to obtain the minimax rate in multiple change-point detection [Brunel, 2014, section 2].

Theorem 3.1 helps choosing C, which is a key parameter of KCP, as in any penalized model selection procedure. However, in practice, we do not recommend to directly use (3.3) for choosing C for two reasons: $C_{\min}, C_{\max}$ depend on unknown quantities $D^\star, \underline{\Lambda}_{\tau^\star}, \underline{\Delta}$, and the exact values of the constants in $C_{\min}, C_{\max}$ might be pessimistic compared to what we can observe from simulation experiments. We rather suggest to use a data-driven method for choosing C, see Section 4.1.

Finally, note that if we know $D^\star$, we can replace $\widehat{\tau}$ by

$$\widehat{\tau}(D^\star) \in \operatorname*{arg\,min}_{\tau \in \mathcal{T}_n^{D^\star}} \big\{ \widehat{\mathcal{R}}_n(\tau) \big\} .$$

Then, assuming that $\underline{\Lambda}_{\tau^\star} > v_1(y)$ — which is weaker than assuming $C_{\min} < C_{\max}$ —, the proof of Theorem 3.1 shows that, on Ω, we have

$$\frac{1}{n} \mathrm{d}_\infty^{(1)}(\tau^\star, \widehat{\tau}(D^\star)) \leq v_1(y) .$$

We now present an analogue of Theorem 3.1 for another penalty function, $\mathrm{pen}_{\mathrm{L}}$.

Theorem 3.2. *Suppose that Assumption 2.1 holds true. For any $y > 0$, an event Ω of probability at least $1 - \mathrm{e}^{-y}$ exists on which the following holds true. For any $c_1, c_2 > 0$, let $\widehat{\tau}$ be defined by*

$$\widehat{\tau} \in \operatorname*{arg\,min}_{\tau \in \mathcal{T}_n} \big\{ \mathrm{crit}(\tau) \big\} \qquad \textit{where} \qquad \mathrm{crit}(\tau) = \widehat{\mathcal{R}}_n(\tau) + \mathrm{pen}(\tau) , \tag{2.2}$$

with $\mathrm{pen} = \mathrm{pen}_L$ *defined by*

$$\mathrm{pen}_L(\tau) := \frac{D_\tau}{n}\left(c_1 \log\frac{n}{D_\tau} + c_2\right), \tag{2.4}$$

Define

$$c_{1,\min} := \frac{74M^2}{3}(D^\star + 1) \quad \textit{and} \quad c_{1,\max} := \frac{n\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2}{6D^\star(y + \log(n) + 4)} - \frac{74M^2}{3}(D^\star + 2)\,.$$

Set

$$c_2 := (y+4)c_1 + \frac{74M^2}{3}(D^\star + 1)(y+4)\,.$$

Then, if

$$c_{1,\min} < c_1 < c_{1,\max}\,, \tag{3.4}$$

on Ω*, we have*

$$\widehat{D} = D^\star \qquad \textit{and} \qquad \frac{1}{n}\,\mathrm{d}_\infty^{(1)}(\tau^\star, \widehat{\tau}) \le \frac{148D^\star M^2}{\underline{\Delta}^2} \cdot \frac{y + \log(n) + 4}{n} =: v_1'(y)\,.$$

The proof of Theorem 3.2 is postponed to Section 3.4.5.

The remarks made after Theorem 3.1 remain valid up to minor changes. In particular, the bounds given for the penalty constants in Theorem 3.2 do not depend on n in the same fashion: the minimal theoretical penalty constant $c_{1,\min}$ does not depend on n, as well as the associated c_2. This should not come as a surprise, given that we "incorporated" the $\log(n)$ factor into the penalty. In some sense, $\mathrm{pen}_{\mathrm{L}}$ is more natural in the context of change-point detection, as it captures the complexity of the model. However, we want to emphasize that calibrating *both* c_1 and c_2 is not an easy task (see Lebarbier [2002, Chapter 3] for a detailed discussion of the calibration of c_1 and c_2 in the real case). Since, again, the constants given by Theorem 3.2 may be very pessimistic, we recommend using pen_ℓ for all practical purposes.

Also note that the hypothesis (3.4) is threefold, as it was the case for Assumption (3.3) in Theorem 3.1. Indeed, for $c_{1,\min} < c_{1,\max}$ to hold, in particular one must have $c_{1,\min} < c_{1,\max}$, which translates into

$$\frac{\underline{\Delta}^2}{M^2} > \frac{148D^\star(2D^\star + 3)(y + \log(n) + 4)}{n\underline{\Lambda}_{\tau^\star}}\,,$$

after some algebra. Since $\log(n)/n \to 0$, this condition is satisfied for n large enough.

Finally, as in Theorem 3.1, we recover a speed of convergence of order $\log(n)/n$, which matches the minimax rate, since v_1' is v_1 up to numerical constants.

Notice that all our results are given in terms of $\mathrm{d}_\infty^{(1)}$, which may seem as an arbitrary choice. We discuss in more depth the notion of distances between segmentations in the next section.

3.2.2 Loss functions between segmentations

Theorem 3.1 shows that $\widehat{\tau}$ is close to $\tau^\star$ in terms of $\mathrm{d}_\infty^{(1)}$. Several other loss functions (measures of dissimilarity) can be defined between segmentations [Hubert and Arabie, 1985]. Our goal in this section is to consider a few of them, which are often used or natural for the change-point problem, and to study their relationships. The main result of this section is Lemma 3.1, that states rigorously that all these notions of distance between segmentations are equal, provided they are close enough.

Let us first consider losses related to the Hausdorff distance. For any τ^1 and $\tau^2 \in \mathcal{T}_n$, we define

$$\begin{aligned}
\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right) &:= \max_{1 \leq i \leq D_{\tau^1}-1} \left\{ \min_{1 \leq j \leq D_{\tau^2}-1} \left|\tau_i^1 - \tau_j^2\right| \right\} \\
\mathrm{d}_\infty^{(2)}\left(\tau^1, \tau^2\right) &:= \max_{1 \leq i \leq D_{\tau^1}-1} \left\{ \min_{0 \leq j \leq D_{\tau^2}} \left|\tau_i^1 - \tau_j^2\right| \right\} \\
\mathrm{d}_{\mathrm{H}}^{(i)}(\tau^1, \tau^2) &:= \max\left\{\mathrm{d}_\infty^{(i)}(\tau^1, \tau^2), \mathrm{d}_\infty^{(i)}(\tau^2, \tau^1)\right\} \qquad \text{for } i \in \left\{1, 2\right\}.
\end{aligned}$$

The difference between $\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right)$ and $\mathrm{d}_\infty^{(2)}\left(\tau^1, \tau^2\right)$ is that $\mathrm{d}_\infty^{(2)}$ takes into account the length of τ^1's first and last segments. As to $\mathrm{d}_{\mathrm{H}}^{(i)}$, it is just a symmetrized version of $\mathrm{d}_\infty^{(i)}$ for $i = 1, 2$. Whenever $D_{\tau^1} = D_{\tau^2}$, we define

$$\mathrm{d}_\infty^{(3)}\left(\tau^1, \tau^2\right) := \max_{1 \leq i \leq D_{\tau^1}-1} \left|\tau_i^1 - \tau_i^2\right| .$$

Note that $\mathrm{d}_\infty^{(3)}$ is symmetric thus there is no need to define $\mathrm{d}_{\mathrm{H}}^{(3)}$. One could also define $\mathrm{d}_{\mathrm{H}}^{(1)}$ as the *Hausdorff distance* between the point sets $\{\tau_1^1, \ldots, \tau_{D_{\tau^1}-1}^1\}$ and $\{\tau_1^2, \ldots, \tau_{D_{\tau^2}-1}^2\}$ with respect to the distance $\delta(x, y) = |x - y|$ on $\mathbb{R}$. These definitions are illustrated by Fig. 3-1.

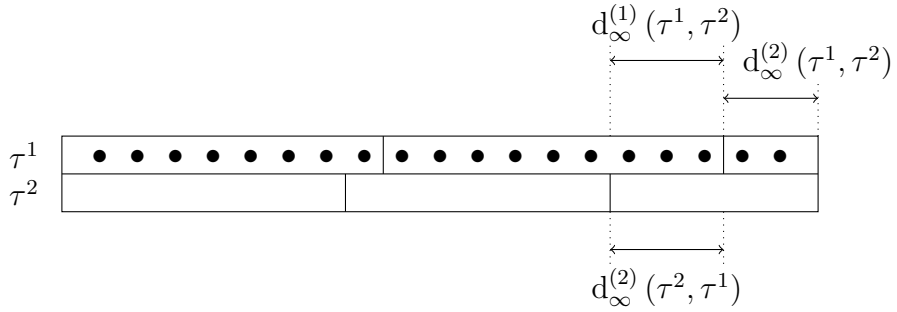

Figure 3-1 – Illustration of the definition of $\mathrm{d}_\infty^{(i)}$, with $n = 19$, $\tau^1 = \big[0, 8, 17, 19\big]$ and $\tau^2 = \big[0, 7, 14, 19\big]$. In this example, $D_{\tau^1} = D_{\tau^2} = 3$. We can compute $\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right) = \mathrm{d}_\infty^{(1)}\left(\tau^2, \tau^1\right) = \mathrm{d}_\infty^{(2)}\left(\tau^2, \tau^1\right) = \mathrm{d}_\infty^{(3)}\left(\tau^1, \tau^2\right) = 3$ and $\mathrm{d}_\infty^{(2)}\left(\tau^1, \tau^2\right) = 2$.

Interestingly, all these loss functions coincide whenever $n^{-1}\,\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right)$ is small enough. The following lemma makes this claim rigorous.

Lemma 3.1. *We have the following two properties.*

(i) For any $\tau^1, \tau^2 \in \mathcal{T}_n$ such that

$$\frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right) < \frac{1}{2}\min\left\{\underline{\Lambda}_{\tau^1}, \underline{\Lambda}_{\tau^2}\right\},$$

we have $D_{\tau^1} = D_{\tau^2}$ and

$$\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right) = \mathrm{d}_\infty^{(2)}\left(\tau^1, \tau^2\right) = \mathrm{d}_\infty^{(3)}\left(\tau^1, \tau^2\right) = \mathrm{d}_{\mathrm{H}}^{(1)}\left(\tau^1, \tau^2\right) = \mathrm{d}_{\mathrm{H}}^{(2)}\left(\tau^1, \tau^2\right).$$

(ii) For any $\tau^1, \tau^2 \in \mathcal{T}_n$ such that

$$D_{\tau^1} = D_{\tau^2} \qquad \text{and} \qquad \frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right) < \frac{\underline{\Lambda}_{\tau^1}}{2},$$

we have

$$\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right) = \mathrm{d}_\infty^{(1)}\left(\tau^2, \tau^1\right) = \mathrm{d}_{\mathrm{H}}^{(1)}\left(\tau^1, \tau^2\right).$$

Lemma 3.1 is proved in Section 3.5.1. As a direct application of Lemma 3.1 we see that the statement of Theorem 3.1 holds true with $\mathrm{d}_\infty^{(1)}$ replaced by *any* of the loss functions that we defined above, at least for n large enough.

Another loss between segmentations is the *Frobenius* loss [Lajugie et al., 2014], which is defined as follows. For any $\tau^1, \tau^2 \in \mathcal{T}_n$,

$$\mathrm{d}_{\mathrm{F}}\left(\tau^1, \tau^2\right) := \left\|\Pi_{\tau^1} - \Pi_{\tau^2}\right\|_{\mathrm{F}},$$

where Π_τ is the orthogonal projection onto F_τ, as defined in Section 2.3.2, and $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix:

$$\forall A \in \mathbb{R}^{N\times M}, \qquad \|A\|_{\mathrm{F}}^2 := \sum_{i=1}^{N}\sum_{j=1}^{M} A_{ij}^2.$$

A closed-form formula for d_{F} can be derived from the matrix representation of Π_τ that is given by (2.7): for any $i, j \in \{1, \ldots, n\}$,

$$(\Pi_\tau)_{i,j} = \begin{cases} \frac{1}{|\lambda|} & \text{if } i \text{ and } j \text{ belong to the same segment } \lambda \text{ of } \tau \\ 0 & \text{otherwise.} \end{cases}$$

An interesting feature of the Frobenius loss is that it is smaller than 1 only when τ^1 and τ^2 have the same number of segments, whereas Hausdorff distances can be small with very different numbers of segments. Indeed, we prove in Section 3.5.2 that

$$|D_{\tau^1} - D_{\tau^2}| \le \mathrm{d}_{\mathrm{F}}\left(\tau^1, \tau^2\right)^2 \le D_{\tau^1} + D_{\tau^2}. \tag{3.5}$$

The next proposition shows that there is an equivalence (up to constants) between the Hausdorff and Frobenius losses between segmentations, provided that they are close enough.

Proposition 3.1. *Suppose that* $D_{\tau^1} = D_{\tau^2}$ *and* $\frac{1}{n}\,\mathrm{d}_\infty^{(1)}(\tau^1,\tau^2) < \underline{\Lambda}_{\tau^1}/2$*, then*

$$\left(\mathrm{d}_{\mathrm{F}}\left(\tau^1,\tau^2\right)\right)^2 \leq \frac{12 D_{\tau^1}}{\underline{\Lambda}_{\tau^1}} \cdot \frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^1,\tau^2\right) .$$

If in addition $\frac{1}{n}\,\mathrm{d}_\infty^{(1)}(\tau^1,\tau^2) < \underline{\Lambda}_{\tau^1}/3$*, then*

$$\frac{2}{3\overline{\Lambda}_{\tau^1}} \frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^1,\tau^2\right) \leq \left(\mathrm{d}_{\mathrm{F}}\left(\tau^1,\tau^2\right)\right)^2 .$$

Prop. 3.1 was first stated and proved by [Lajugie et al., 2014, Theorem B.2]. We prove it in Section 3.5.2 for completeness.

As a corollary of Theorem 3.1 and Prop. 3.1, we get the following guarantee on the Frobenius loss between $\tau^\star$ and the segmentation $\widehat{\tau}$ estimated by KCP.

Corollary 3.1. *Under the assumptions of Theorem 3.1, on the event* Ω *defined by Theorem 3.1, for any* $\widehat{\tau}$ *satisfying* (2.2) *with* pen *defined by* (2.3)*, we have:*

$$\mathrm{d}_{\mathrm{F}}\left(\tau^\star,\widehat{\tau}\right) \leq \frac{43 D^\star}{\sqrt{\underline{\Lambda}_{\tau^\star}}} \cdot \frac{M}{\underline{\Delta}} \sqrt{\frac{y+\log(n)+1}{n}} .$$

Note that Corollary 3.1 gives a better result (at least for large n) than the obvious bound

$$\mathrm{d}_{\mathrm{F}}\left(\tau^\star,\widehat{\tau}\right) \leq D^\star + \widehat{D} - 2 .$$

Proof. On the event Ω, we have $\frac{1}{n}\,\mathrm{d}_\infty^{(1)}(\tau^\star,\widehat{\tau}) < \underline{\Lambda}_{\tau^\star}/(D^\star+1)$ and $D^\star = \widehat{D}$. Therefore, according to Prop. 3.1,

$$\left(\mathrm{d}_{\mathrm{F}}\left(\tau^\star,\widehat{\tau}\right)\right)^2 \leq \frac{12 D^\star}{\underline{\Lambda}_{\tau^\star}} \cdot \frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^\star,\widehat{\tau}\right) \leq \frac{1776\left(D^\star\right)^2\left(y+\log(n)+1\right)}{n\underline{\Lambda}_{\tau^\star}} \cdot \frac{M^2}{\underline{\Delta}^2} .$$

□

Up to this point, we assessed the quality of the segmentation τ by considering the proximity of τ with $\tau^\star$. Another natural idea is to measure the distance between $\mu^\star$ and $\mu_\tau^\star$ in $\mathcal{H}^n$. It is closely related to the oracle inequality proved by Arlot et al. [2012], which implies an upper bound on $\|\mu^\star - \widehat{\mu}_{\widehat{\tau}}\|^2$. We can also observe that there is a simple relationship between $\|\mu^\star - \mu_\tau^\star\|^2$ and the Frobenius distance between τ and $\tau^\star$. Indeed,

$$\|\mu^\star - \mu_\tau^\star\|^2 = \|(\Pi_{\tau^\star} - \Pi_\tau)\mu^\star\|^2 \leq \|\Pi_{\tau^\star} - \Pi_\tau\|_2^2\,\|\mu^\star\|^2 \leq \left(\mathrm{d}_{\mathrm{F}}\left(\tau^\star,\widehat{\tau}\right)\right)^2\|\mu^\star\|^2 . \quad (3.6)$$

Eq. (3.17) in the proof of Theorem 3.1 shows that on Ω, under the assumptions of Theorem 3.1,

$$\|\mu^\star - \mu_{\widehat{\tau}}^\star\|^2 \leq 74\big(y+\log(n)+1\big) D^\star M^2$$

which is slightly better (but similar) to what Corollary 3.1, (3.6) and the bound $\|\mu^\star\|^2 \leq M^2 n$ imply together.

3.2.3 Extension to the finite variance case

Theorem 3.1 and 3.2 are valid under a boundedness assumption (Assumption 2.1). What happens under the weaker Assumption 2.2? As a first step, we provide a result for

$$\widehat{\tau}(D^\star, \delta_n) \in \underset{\tau \in \mathcal{T}_n^{D^\star} \,/\, \underline{\Lambda}_\tau \geq n\delta_n}{\arg\min} \left\{ \widehat{\mathcal{R}}_n(\tau) \right\} \tag{3.7}$$

for some $\delta_n > 0$. In other words, we restrict our search to segmentations τ of the correct size — hence $D^\star$ must be known *a priori* — and having no segment with less than $n\delta_n$ observations. We discuss how to relax this restriction right after the statement of Theorem 3.3. Note that, since we know $D^\star$, there is no need for a penalty function in the new problem given by Eq. (3.7). Also notice that the dynamic programming algorithm of Harchaoui and Cappé [2007] can be used for computing $\widehat{\tau}(D^\star, \delta_n)$ efficiently.

Similarly to $\underline{\Delta}$, we define $\overline{\Delta} := \max_i \left\| \mu_i^\star - \mu_{i+1}^\star \right\|_{\mathcal{H}}$.

Theorem 3.3. *Suppose that Assumption 2.2 holds true. For any $\delta_n, y > 0$, define*:

$$v_2(y, \delta_n) := 24 (D^\star)^2 \frac{\overline{\Delta}\sqrt{V}}{\underline{\Delta}^2} \frac{y}{\sqrt{n}} + 8 D^\star \frac{V}{\underline{\Delta}^2} \frac{y^2}{n\delta_n} \,.$$

For any $y > 0$, an event Ω_2 exists such that

$$\mathbb{P}\left(\Omega_2\right) \geq 1 - \frac{1}{y^2}$$

and, on Ω_2, we have the following: *for any $\delta_n \in (0, \underline{\Lambda}_{\tau^\star}]$ and any $\widehat{\tau}(D^\star, \delta_n)$ satisfying Eq.* (3.7)*, if $v_2(y, \delta_n) \leq \underline{\Lambda}_{\tau^\star}$,*

$$\frac{1}{n} \mathrm{d}_\infty^{(1)} \left(\tau^\star, \widehat{\tau}(D^\star, \delta_n) \right) \leq v_2(y, \delta_n) \,. \tag{3.8}$$

We postpone the proof of Theorem 3.3 to Section 3.4.6. Let us make a few remarks.

As for Theorem 3.1, our result is non-asymptotic. However, it is interesting to write it down in the setting of Example 2.1. If n goes to infinity, then the assumption $\underline{\Lambda}_{\tau^\star} \geq \delta_n$ is satisfied whenever $\delta_n \to 0$. If we furthermore require that $n\delta_n \to \infty$, then Eq. (3.8) implies that

$$\frac{1}{n} \mathrm{d}_\infty^{(1)} \left(\tau^\star, \widehat{\tau}(D^\star, \delta_n) \right) \xrightarrow[n \to +\infty]{\mathbb{P}} 0 \,,$$

by taking a well-chosen y of order $\sqrt{n} + \sqrt{n\delta_n}$. In the particular case of the linear kernel, this result is known under various hypothesis [Lavielle and Moulines, 2000, for instance]; it is new for a general kernel.

More precisely, if we take $\delta_n = n^{-1/2}$, Theorem 3.3 implies that

$$\frac{1}{n} \mathrm{d}_\infty^{(1)} \left(\tau^\star, \widehat{\tau}(D^\star, n^{-1/2}) \right)$$

goes to zero at least as fast as $\ell_n / \sqrt{n}$, where $(\ell_n)_{n \geq 1}$ is any sequence tending to

infinity, for instance $\ell_n = \log(n)$. This speed seems suboptimal compared to previous results [Lavielle and Moulines, 2000, for instance] — which do not consider the case of a general kernel —, but we have not been able to prove tight enough deviation bounds for getting the localization rate $\log(n)/n$ under Assumption 2.2.

How does Theorem 3.3 compares to Theorem 3.1 and 3.2? First, as noticed in Remark 3.4 in Section 3.4.4, the result of Theorem 3.1 also holds true for $\widehat{\tau}(D^\star, \delta_n)$ as long as $\underline{\Lambda}_{\tau^\star} \geq \delta_n$. Second, $v_1(y)$ is usually smaller than $v_2(y, \delta_n)$ — its order of magnitude is smaller when $n \to +\infty$ —, and the lower bound on the probability of Ω is better than the one for Ω_2. There is no surprise here: the stronger Assumption 2.1 helps us proving a stronger result for $\widehat{\tau}(D^\star, \delta_n)$. Nevertheless, these only are upper bounds, so we do not know whether the performance of $\widehat{\tau}(D^\star, \delta_n)$ actually changes much depending on the noise assumption. For instance, as already noticed, we do not believe that the localization speed $\log(n)/n$ requires a boundedness assumption; in particular cases at least, it has been obtained for unbounded data [Lavielle and Moulines, 2000; Boysen et al., 2009].

The dependency in k of the speed of convergence of $\widehat{\tau}(D^\star, \delta_n)$ is slightly less clear than in Theorem 3.1. The signal-to-noise ratio appears through $\underline{\Delta}^2/V$, as expected, but the size $\overline{\Delta}$ of the largest true jump also appears in v_2. At the very least, it is clear that $\underline{\Delta}^2/V$ should not be too small.

As noted by Lavielle and Moulines [2000], it may be possible to get rid of the minimal segment length δ_n, either by imposing stronger conditions on ε — which are not met in our setting — or by constraining the values of $\widehat{\mu}$ to lie in a compact subset $\Theta \subset \mathcal{H}^{D^\star+1}$.

3.3 Discussion

Before proving our main results, let us discuss some of their consequences regarding the KCP procedure.

Fully non-parametric consistent change-point detection. We have proved that for any kernel satisfying some reasonably mild hypotheses, the KCP procedure outputs a segmentation close by the true segmentation with high probability.

An important particular example is the "asymptotic setting" of Example 2.1, where we have a fixed true segmentation $\tau^\star$ and fixed distributions $P_1, \ldots, P_{K+1}$ from which more and more points are sampled. How fast can KCP recover $\tau^\star$, without any prior information on the number of segments $D^\star$ or on the distributions $P_1, \ldots, P_{K+1}$?

Let us take a bounded characteristic kernel — for instance the Gaussian or the Laplace kernel if $\mathcal{X} = \mathbb{R}^d$ —, so that Assumption 2.1 holds true. Then, Theorem 3.1 shows that KCP detects consistently all changes in the distribution of the X_i, and localizes them at speed $\log(n)/n$. This speed also depends on the kernel k and the size of the differences between the P_j, through the ratio $\underline{\Delta}^2/M^2$. Obtaining such a fully non-parametric result for multiple change-points with a general set $\mathcal{X}$ — we only need to know a bounded characteristic kernel on $\mathcal{X}$ — has never been obtained before. To the best of our knowledge, non-parametric consistency results for the detection

of arbitrary changes in the distribution of the data have only been obtained for real-valued data [Zou et al., 2014] or for the case of a single change-point [Carlstein, 1988; Brodsky and Darkhovsky, 2013].

Choice of k. An important question remains: how to choose the kernel k? In Theorem 3.1, k only appears through the "signal-to-noise ratio" $\underline{\Delta}^2/M^2$, leading to better theoretical guarantees when this signal-to-noise ratio is larger: a larger value for $C_{\max}$ and a smaller bound v_1 on $\mathrm{d}_\infty^{(1)}(\tau^\star, \widehat{\tau})$. Therefore, a simple strategy for choosing the kernel is to pick k that maximizes $\underline{\Delta}^2/M^2$, at least among a family of kernels, for instance Gaussian kernels. This first idea requires to know the distributions of the X_i, or at least to have prior information on them. As we have noticed, when the change-points locations are known, $\underline{\Delta}^2$ corresponds to the minimum of the MMDs between the distributions of the X_i over contiguous segments. In this particular setting, it may be feasible to estimate and to maximize $\underline{\Delta}^2$ with respect to the kernel k, as done by Gretton et al. [2012b]. This question is addressed in more details in Section 4.3.1. An interesting future development would be to build an estimator of $\underline{\Delta}^2$ without knowing the change-point locations and to maximize this estimator with respect to the kernel k. We refer to Arlot et al. [2012, section 7.2] for a complementary discussion about the choice of k for KCP.

Choice of C. Another important parameter of the KCP procedure is the constant C that appears in the linear penalty function. As mentioned below Theorem 3.1, our theoretical guarantees provide some guidelines for choosing C, but these are not sufficient to choose precisely C in practice. We recommend to follow the advice of [Arlot et al., 2012, section 6.2] on this point, which is to choose C from data with the "dimension jump" heuristic [Baudry et al., 2012]. The exact procedure is explained in details in Section 4.1.

Modularity of the proofs and possible extensions. Finally, we would like to emphasize what we believe to be an important contribution of this thesis. The structure of the proofs of Theorems 3.1 and 3.3 — which follow the same strategy — is modular, so that one can easily adapt it to different sets of assumptions. This is also the case for the proof of Theorem 3.2, to some extent: the behavior of the penalty function near $D^\star$ has to be controlled more precisely.

Our proof strategy is not fully new, since it is similar to the one of almost all previous papers analyzing the consistency of least-squares change-point detection procedures. In particular, we adapted some ideas of the proofs of Lavielle and Moulines [2000] to the Hilbert space setting. Nevertheless, these papers formulate their main results in asymptotic terms, which can be seen as a limitation — especially when n is small or $\mathcal{X}$ is of large dimension. Another approach is the one of Lebarbier [2005]; Comte and Rozenholc [2004]; Arlot et al. [2012] where non-asymptotic oracle inequalities — using concentration inequalities and following the model selection results of Birgé and Massart [2001] — are provided as theoretical guarantees on some

penalized least-squares change-point procedures. Up to now, these two approaches seemed difficult to combine. The proofs of Theorems 3.1 and 3.3 show how they can be reconciled, which allows us to mix their strengths.

Indeed, assumptions on the distributions of the X_i — Assumptions 2.1 and 2.2 — are only used for proving bounds on two quantities depending on ε — a linear term L_τ and a quadratic term Q_τ —, uniformly over $\tau \in \mathcal{T}_n$. Under Assumption 2.1, this is done thanks to concentration inequalities (Lemmas 3.9 and 3.8) which have been proved first by Arlot et al. [2012] in order to get an oracle inequality. Under Assumption 2.2, this is done by generalizing the method of Lavielle and Moulines [2000] to Hilbert-space valued data, through two deterministic bounds (Lemmas 3.6 and 3.7) and a deviation inequality for

$$B_n := \max_{1 \leq k \leq n} \left\| \sum_{j=1}^{k} \varepsilon_j \right\|_{\mathcal{H}}$$

(Lemma 3.11). The rest of the proofs does not use any information about the distribution of $X_1, \ldots, X_n$.

As a consequence, if one can generalize these bounds to another setting, a straightforward consequence is that a result similar to Theorem 3.1, 3.2, or 3.3 holds true for the KCP procedure in this new setting. In particular, this could be used for dealing with the case of dependent data $X_1, \ldots, X_n$. We could also consider an intermediate assumption between Assumption 2.2 and Assumption 2.1, of the form:

$$\max_{1 \leq i \leq n} \mathbb{E}\big[k(X_i, X_i)^\alpha\big] \leq B_\alpha < +\infty \,,$$

for some $\alpha \in (1, +\infty)$.

Without further ado, we now turn to the proofs of Theorems 3.1, 3.2 and 3.3.

3.4 Proofs

Let us start by describing our general strategy for proving Theorems 3.1 and 3.3. Our goal is to build a large probability event on which any $\widehat{\tau} \in \arg\min_{\tau \in \mathcal{T}_n} \mathrm{crit}\,(\tau)$ belongs to some subset $\mathcal{E}$ of $\mathcal{T}_n$. For proving this, we use the key fact that $\mathrm{crit}\,(\tau^\star) \geq \mathrm{crit}\,(\widehat{\tau})$, together with a lower bound on $\mathrm{crit}\,(\tau)$ holding simultaneously for all $\tau \in \mathcal{T}_n$ — hence for $\tau = \widehat{\tau}$.

In order to get such a lower bound on the empirical penalized criterion, we start by decomposing it in Section 3.4.1 into terms that are simpler to control individually: two random terms — a linear function of ε and a quadratic function of ε —, and two deterministic terms — the approximation error and the penalty. Then, we control these terms thanks to deterministic bounds (Section 3.4.2) and deviation/concentration inequalities (Section 3.4.3). Finally, we prove Theorem 3.1 in Section 3.4.4 and Theorem 3.3 in Section 3.4.6.

The proof of Theorem 3.2 shares the same ideas, but is slightly different due to the more complicated form of the penalty. It can be found in Section 3.4.5.

3.4.1 Decomposition of the empirical risk

The first step in the proofs of Theorems 3.1, 3.2 and 3.3 is to decompose the empirical risk (2.8).

Lemma 3.2. *Let $\tau \in \mathcal{T}_n$ be a segmentation. Define $\mu_\tau^\star = \Pi_\tau \mu^\star$. Then we can write*

$$n\,\widehat{\mathcal{R}}_n(\tau) = \|Y - \widehat{\mu}_\tau\|^2 = \|\mu^\star - \mu_\tau^\star\|^2 + 2\langle \mu^\star - \mu_\tau^\star, \varepsilon\rangle - \|\Pi_\tau \varepsilon\|^2 + \|\varepsilon\|^2 \,. \tag{3.9}$$

Proof. First, recall that $\widehat{\mu}_\tau = \Pi_\tau Y$ and that $Y = \mu^\star + \varepsilon$, hence

$$\begin{aligned}
\|Y - \widehat{\mu}_\tau\|^2 &= \|Y - \Pi_\tau Y\|^2 \\
&= \|\mu^\star + \varepsilon - \Pi_\tau(\mu^\star + \varepsilon)\|^2 \\
&= \|\mu^\star - \Pi_\tau \mu^\star\|^2 + \|\varepsilon - \Pi_\tau \varepsilon\|^2 + 2\langle \mu^\star - \Pi_\tau \mu^\star, \varepsilon - \Pi_\tau \varepsilon\rangle.
\end{aligned}$$

Since Π_τ is an orthogonal projection,

$$\begin{aligned}
\|Y - \widehat{\mu}_\tau\|^2 &= \|\mu^\star - \mu_\tau^\star\|^2 + \|\varepsilon\|^2 - 2\langle \varepsilon, \Pi_\tau \varepsilon\rangle + \|\Pi_\tau \varepsilon\|^2 + 2\langle (\mathrm{I} - \Pi_\tau)\mu^\star, \varepsilon\rangle \\
&= \|\mu^\star - \mu_\tau^\star\|^2 + \|\varepsilon\|^2 - \|\Pi_\tau \varepsilon\|^2 + 2\langle (\mathrm{I} - \Pi_\tau)\mu^\star, \varepsilon\rangle \,.
\end{aligned}$$

□

Since each term of Eq. (3.9) behaves differently and is controlled via different techniques depending on the result to be proven, we name each of these terms:

$$L_\tau := \langle \mu^\star - \mu_\tau^\star, \varepsilon\rangle \,, \qquad Q_\tau := \|\Pi_\tau \varepsilon\|^2 \qquad \text{and} \qquad A_\tau := \|\mu^\star - \mu_\tau^\star\|^2 \,. \tag{3.10}$$

It should be clear that L stands for "linear", Q stands for "quadratic" and A stands for "approximation error". We also define

$$\psi_\tau := 2L_\tau - Q_\tau + A_\tau \,. \tag{3.11}$$

Therefore a reformulation of Lemma 3.2 is

$$n\,\widehat{\mathcal{R}}_n(\tau) = \psi_\tau + \|\varepsilon\|^2 \,.$$

Notice that $L_{\tau^\star} = A_{\tau^\star} = 0$ and $Q_{\tau^\star} \geq 0$, hence $\psi_{\tau^\star} \leq 0$. Also note that ψ, L and Q are random quantities depending on ε.

3.4.2 Deterministic bounds

In this section, we provide some deterministic bounds that are used in the proofs of Theorems 3.1, 3.2 and 3.3.

Approximation error A_τ

We begin by the following result, which is the reason for the $\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2$ term in the minimal penalty constants in Theorem 3.1 and 3.2.

Lemma 3.3. *Let $\tau \in \mathcal{T}_n$ be a segmentation such that $D := D_\tau < D^\star$. Then*

$$\frac{1}{n} A_\tau = \frac{1}{n} \left\| \mu^\star - \mu^\star_\tau \right\|^2 \geq \frac{1}{2} \underline{\Lambda}_{\tau^\star} \underline{\Delta}^2 \,. \tag{3.12}$$

The proof of Lemma 3.3 can be found in Section 3.5.3.

Remark 3.1. The inequality in Lemma 3.3 is tight. Indeed, consider the simple case $D_\tau = 1$ and $D^\star = 2$. Assume that $n = 2m$ is an even number, and let $\tau^\star_1 = m$. It follows from definitions (3.1) and (3.2) that, in this case,

$$\underline{\Delta} = \left\| \mu^\star_1 - \mu^\star_n \right\|_{\mathcal{H}} \qquad \text{and} \qquad \underline{\Lambda}_{\tau^\star} = \frac{1}{2} \,.$$

According to Eq. (2.7), $(\mu^\star_\tau)_i = \frac{1}{2}\left(\mu^\star_1 + \mu^\star_n\right)$, which yields

$$\frac{1}{n} A_\tau = \frac{1}{4} \left\| \mu^\star_1 - \mu^\star_n \right\|^2_{\mathcal{H}} = \frac{1}{2} \underline{\Lambda}_{\tau^\star} \underline{\Delta}^2 \,.$$

Thus, in this particular class of examples, equality holds in (3.12).

We next state an analogous result, valid for any $\tau \in \mathcal{T}_n$, which plays a key role in the proofs of Theorems 3.1, 3.2 and 3.3.

Lemma 3.4. *For any $\tau \in \mathcal{T}_n$,*

$$\frac{1}{n} A_\tau \geq \frac{1}{2} \min\left\{ \underline{\Lambda}_{\tau^\star}, \frac{1}{n} \mathrm{d}^{(1)}_\infty\left(\tau^\star, \tau\right) \right\} \underline{\Delta}^2 \,. \tag{3.13}$$

Lemma 3.4 is proved in Section 3.5.4.

Linear term L_τ and quadratic term Q_τ

The proof of Theorem 3.3 relies on some deterministic bounds on L_τ and Q_τ. We start with a preliminary lemma.

Lemma 3.5. *For any $\varepsilon_1, \ldots, \varepsilon_n \in \mathcal{H}$,*

$$\frac{1}{2} \max_{1 \leq a < b \leq n} \left\| \sum_{j=a}^{b} \varepsilon_j \right\|_{\mathcal{H}} \leq \max_{1 \leq k \leq n} \left\| \sum_{j=1}^{k} \varepsilon_j \right\|_{\mathcal{H}} =: B_n \,. \tag{3.14}$$

Proof. For every $a < b$, we have:

$$\left\| \sum_{j=a}^{b} \varepsilon_j \right\|_{\mathcal{H}} = \left\| \sum_{j=1}^{b} \varepsilon_j - \sum_{j=1}^{a-1} \varepsilon_j \right\|_{\mathcal{H}} \leq \left\| \sum_{j=1}^{b} \varepsilon_j \right\|_{\mathcal{H}} + \left\| \sum_{j=1}^{a-1} \varepsilon_j \right\|_{\mathcal{H}} \leq 2B_n \,.$$

□

The following result is a deterministic bound on Q_τ in terms of B_n.

Lemma 3.6. *Let $\tau \in \mathcal{T}_n$ be a segmentation. Then*

$$Q_\tau \leq \frac{4D_\tau B_n^2}{n\underline{\Lambda}_\tau}\,.$$

Proof. By Eq. (2.7), for any $\tau_{\ell-1}+1 \leq i \leq \tau_\ell$,

$$(\Pi_\tau \varepsilon)_i = \frac{1}{|\tau_\ell - \tau_{\ell-1}|} \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_j\,.$$

Since $Q_\tau = \|\Pi_\tau \varepsilon\|^2$,

$$\begin{aligned}
Q_\tau &= \sum_{i=1}^{n} \|(\Pi_\tau \varepsilon)_i\|_{\mathcal{H}}^2 \\
&= \sum_{\ell=1}^{D_\tau} \sum_{i=\tau_{\ell-1}+1}^{\tau_\ell} \left\| \frac{1}{|\tau_\ell - \tau_{\ell-1}|} \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_j \right\|^2 \\
&= \sum_{\ell=1}^{D_\tau} \frac{1}{|\tau_\ell - \tau_{\ell-1}|} \left\| \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_j \right\|_{\mathcal{H}}^2 \\
&\leq D_\tau \max_{1\leq \ell \leq D_\tau} \left\{ \frac{1}{|\tau_\ell - \tau_{\ell-1}|} \left\| \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_j \right\|_{\mathcal{H}}^2 \right\} \\
&\leq \frac{D_\tau}{n\underline{\Lambda}_\tau} \max_{1\leq \ell \leq D_\tau} \left\| \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \varepsilon_j \right\|_{\mathcal{H}}^2 \\
Q_\tau &\leq \frac{4D_\tau}{n\underline{\Lambda}_\tau} B_n^2\,,
\end{aligned}$$

where we used Lemma 3.5 for the last inequality. □

The following result is a deterministic bound on L_τ.

Lemma 3.7. *For any $\tau \in \mathcal{T}_n$,*

$$|L_\tau| \leq 6D^\star \max\{D^\star, D_\tau\}\, \overline{\Delta} B_n\,.$$

Lemma 3.7 is proved in Section 3.5.5.

3.4.3 Concentration

In this subsection, we present concentration results on Q_τ, L_τ, and deviation bounds for B_n — which will imply deviation bounds on Q_τ and L_τ by Lemmas 3.6 and 3.7). For any $j \in \{1, \dots, n\}$, $\tau \in \mathcal{T}_n$ and $\ell \in \{1, \dots, D_\tau\}$, we define

$$v_j := \mathbb{E}\left[\|\varepsilon_j\|_{\mathcal{H}}^2\right] \qquad v_{\tau,\ell} := \frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} v_j \qquad \text{and} \qquad v_\tau := \sum_{\ell=1}^{D} v_{\tau,\ell}\,.$$

Concentration under Assumption 2.1. We present in this section concentration results that we already identified as essential to the proofs of Theorems 3.1 and 3.2.

The first result that we provide [1] in this section concerns the quadratic term Q_τ when Assumption 2.1 is satisfied.

Lemma 3.8. *Suppose that Assumption 2.1 holds true. Then for any $x > 0$, with probability at least $1 - \mathrm{e}^{-x}$,*

$$Q_\tau - v_\tau \leq \left(x + 2\sqrt{2xD_\tau}\right) \frac{14M^2}{3}\,. \tag{3.15}$$

We want to emphasize that obtaining concentration results for $Q_\tau = \|\Pi_\tau \varepsilon\|^2$ is not an easy task at first sight, since ε_j belongs to $\mathcal{H}$, a potentially infinite-dimensional space. A way to tackle the challenge of concentrating Q_τ is to use concentration results for U-statistics of order 2 with values in a general set [Giné and Nickl, 2015, Th. 3.4.8], but in this case a term of order M^2x^2 would appear in the right-hand side of Eq. (3.15). Another would be the use of Talagrand's inequality [Boucheron et al., 2013, Cor. 12.12], leading to the same deviation term of order M^2x^2. But we need a deviation term of order M^2x for our proof machinery to operate. We refer to Arlot et al. [2012, Section 5.4.3] for a detailed discussion of such matters and to Ledoux and Talagrand [2013] for related questions regarding concentration of random variables in Banach spaces. Before turning to the proof of Lemma 3.8, we recall Bernstein's inequality and Pinelis-Sakhanenko's inequality, two results that play a crucial role in this proof.

Bernstein's inequality is a cornerstone in concentration inequality theory. It provides sub-exponential concentration bounds for a sum of independent random variables under an hypothesis on the growth of the moments of this sum. First proved by Bernstein in the 20s [Bernstein, 1924], it implies as a special cases standard tools in concentration, *e.g.*, Bennett's inequality. We give here the version that can be found in Boucheron et al. [2013], which is a bit more general than the classical form.

Proposition 3.2 (Bernstein's inequality)**.** *Let $Z_1, \dots, Z_m$ be independent real-valued*

1. Lemma 3.8 is a refinement of a result obtained by Arlot, Celisse and Harchaoui in the first version of Arlot et al. [2012].

random variables. Assume that there exist some positive constants v and c such that

$$\forall q \geq 2, \qquad \sum_{i=1}^{m} \mathbb{E}\left[|Z_i|^q\right] \leq \frac{q!}{2} v c^{q-2} .$$

Then, for every $x > 0$,

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

Pinelis-Sakhanenko's inequality [Pinelis and Sakhanenko, 1986, Cor. 1] is close in spirit to Bernstein's inequality, with the noteworthy difference that it concerns a sum of Hilbert-valued random variables.

Proposition 3.3 (Pinelis-Sakhanenko's inequality)**.** *Let $Z_1, \ldots, Z_m$ be independent random variables with values in some Hilbert space $\mathfrak{H}$. Assume that the Z_i are centered and that there exists some positive constants v and c such that*

$$\forall q \geq 2, \qquad \sum_{i=1}^{m} \mathbb{E}\left[\|Z_i\|_{\mathfrak{H}}^q\right] \leq \frac{q!}{2} v c^{q-2} .$$

Then, for every $x > 0$,

$$\mathbb{P}\left(\left\| \sum_{i=1}^{m} Z_i \right\|_{\mathfrak{H}} > x \right) \leq 2 \exp\left(\frac{-x^2}{2(cx + v)} \right) .$$

We are now ready to prove Lemma 3.8.

Proof of Lemma 3.8. Let us define $T_\lambda := \frac{1}{|\lambda|} \left\| \sum_{j \in \lambda} \varepsilon_j \right\|_{\mathcal{H}}^2$. Remark that $(T_\lambda)_\lambda$ is a sequence of independent real-valued random variables. Since $Q_\tau = \sum_{\lambda \in \tau} T_\lambda$, we can obtain a concentration inequality for Q_τ *via* Bernstein's inequality as long as T_λ satisfies some moment conditions. We will use the Pinelis-Sakhanenko's deviation inequality to prove such bounds.

By the independence property of the ε_js, for any $\lambda \in \tau$, we have $\mathbb{E}\left[T_\lambda\right] = v_\lambda$. Then, for any $q \geq 2$,

$$\mathbb{E}\left[T_\lambda^q\right] = \frac{1}{|\lambda|^q} \mathbb{E}\left[\left\| \sum_{j \in \lambda} \varepsilon_j \right\|_{\mathcal{H}}^{2q} \right] = \frac{1}{|\lambda|^q} \int_0^{2|\lambda|M} 2q x^{2q-1} \, \mathbb{P}\left(\left\| \sum_{j \in \lambda} \varepsilon_j \right\|_{\mathcal{H}} \geq x \right) \mathrm{d}\, x ,$$

because for any j, $\|\varepsilon_j\|_{\mathcal{H}} \leq 2M$ almost surely — *cf.* Eq. 2.11. The boundedness of $\|\varepsilon_j\|_{\mathcal{H}}$ also implies that, for any $p \geq 2$ and $\lambda \in \tau$,

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

$$\leq \frac{p!}{2} \times |\lambda| \, M^2 \times \left(\frac{2M}{3}\right)^{p-2} ,$$

that is, the assumptions of Pinelis-Sakhanenko's deviation inequality hold true with $c = 2M/3$ and $v = |\lambda| \, M^2$. Therefore, for any $x \in [0, 2 \, |\lambda| \, M]$,

$$\begin{aligned} \mathbb{P}\left(\left\| \sum_{j \in \lambda} \varepsilon_j \right\|_{\mathcal{H}} \geq x \right) &\leq 2 \exp\left(\frac{-x^2}{2\left(|\lambda| \, M^2 + \frac{2Mx}{3}\right)} \right) \\ &\leq 2 \exp\left(\frac{-3x^2}{14 \, |\lambda| \, M^2} \right) . \end{aligned}$$

We now make use of the change of variables $u = \sqrt{3/(7 \, |\lambda|)} x / M$, and write

$$\begin{aligned} \mathbb{E}\left[T_\lambda^q\right] &\leq \frac{4q}{|\lambda|^q} \int_0^{2|\lambda|M} x^{2q-1} \exp\left(\frac{-3x^2}{14 \, |\lambda| \, M^2} \right) \mathrm{d}\, x \\ &\leq 4q \left(\frac{7M^2}{3}\right)^q \int_0^{+\infty} u^{2q-1} \, \mathrm{e}^{-u^2/2} \, \mathrm{d}\, u \\ &\leq 2(q!) \left(\frac{14M^2}{3}\right)^q , \end{aligned}$$

where we used

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

Summing over $\lambda \in \tau$, it comes

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

Thus the condition of Bernstein's inequality holds with

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

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

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

The linear term L_τ can be controlled directly *via* Bernstein's inequality. This is

achieved in the next lemma, which is a specialization of Arlot et al. [2012, Prop. 3] under Assumption 2.1.

Lemma 3.9. *Suppose that Assumption 2.1 holds true. Then for any $x > 0$, with probability at least $1 - 2\,\mathrm{e}^{-x}$, for any $\theta > 0$,*

$$|L_\tau| \leq \theta A_\tau + \left(\frac{4}{3} + \frac{1}{2\theta}\right) M^2 x\,.$$

Proof. Let us define $S_\tau := \langle \mu^\star - \mu^\star_\tau, \varepsilon \rangle$. We note that

$$S_\tau = \sum_{i=1}^{n} Z_i \qquad \text{with} \qquad Z_i := \langle (\mu^\star - \mu^\star_\tau)_i\,, \varepsilon_i \rangle_{\mathcal{H}}\,.$$

The Z_is are independent centered real-valued random variables. Let us prove that they satisfy the hypothesis of Bernstein's inequality.

Set $i \in \{1, \dots, n\}$, the Cauchy-Schwarz inequality yields

$$|Z_i| \leq \|(\mu^\star - \mu^\star_\tau)_i\|_{\mathcal{H}} \cdot \|\varepsilon_i\|_{\mathcal{H}}\,.$$

We have already proved that $\|\varepsilon_i\|_{\mathcal{H}} \leq 2M$. According to Eq. (2.7),

$$(\mu^\star - \mu^\star_\tau)_i = \mu^\star_i - \frac{1}{\tau_\ell - \tau_{\ell-1}} \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \mu^\star_j$$

for some $1 \leq \ell \leq D_\tau$, and thus we can write

$$\begin{aligned}\|(\mu^\star - \mu^\star_\tau)_i\|_{\mathcal{H}} &= \frac{1}{\tau_\ell - \tau_{\ell-1}} \left\| \sum_{j=\tau_{\ell-1}+1}^{\tau_\ell} \left(\mu^\star_i - \mu^\star_j\right) \right\|_{\mathcal{H}} \\ &\leq \sup_{\tau_{\ell-1} < j \leq \tau_\ell} \left\|\mu^\star_i - \mu^\star_j\right\|_{\mathcal{H}}\,.\end{aligned}$$

Triangle inequality together with Eq. (2.11) yields $\|(\mu^\star - \mu^\star_\tau)_i\|_{\mathcal{H}} \leq 2M$, and we have proved that $|Z_i| \leq 4M^2$.

Furthermore, again due to Cauchy-Schwarz inequality,

$$\mathbb{E}\left[|Z_i|^2\right] \leq \mathbb{E}\left[\|(\mu^\star - \mu^\star_\tau)_i\|^2_{\mathcal{H}} \cdot \|\varepsilon_i\|^2_{\mathcal{H}}\right]\,.$$

Recall that under Assumption 2.1, $\mathbb{E}\left[\|\varepsilon_i\|^2_{\mathcal{H}}\right] \leq M^2$. Thus

$$\sum_{i=1}^{n} \mathbb{E}\left[|Z_i|^2\right] \leq \|\mu^\star - \mu^\star_\tau\|^2 \cdot \max_i \mathbb{E}\left[\|\varepsilon_i\|^2_{\mathcal{H}}\right] \leq \|\mu^\star - \mu^\star_\tau\|^2 \cdot M^2\,.$$

We now show that the conditions of Bernstein's inequality are satisfied. Let $q \geq 2$,

then

$$\sum_{i=1}^{n} \mathbb{E}\left[|Z_i|^q\right] \leq \left\|\mu^\star - \mu_\tau^\star\right\|^2 M^2 \cdot \left(\frac{4M^2}{3}\right)^{q-2}$$
$$= \left\|\mu^\star - \mu_\tau^\star\right\|^2 M^2 \cdot 3^{q-2} \cdot \left(\frac{4M^2}{3}\right)^{q-2} .$$

Since $3^{q-2} \leq q!/2$ for any $q \geq 2$,

$$\sum_{i=1}^{n} \mathbb{E}\left[|Z_i|^q\right] \leq \frac{q!}{2} \left\|\mu^\star - \mu_\tau^\star\right\|^2 M^2 \left(\frac{4M^2}{3}\right)^{q-2} ,$$

and the conditions of Bernstein's inequality are satisfied with $v = \left\|\mu^\star - \mu_\tau^\star\right\|^2 M^2$ and $c = 4M^2/3$. As a consequence, for any $x > 0$, with probability higher than $1 - 2\,\mathrm{e}^{-x}$,

$$\left|\sum_{i=1}^{n} Z_i\right| \leq \sqrt{2\left\|\mu^\star - \mu_\tau^\star\right\|^2 M^2 x} + \frac{4M^2}{3} x .$$

We conclude the proof by applying the inequality $2ab \leq \theta a^2 + \theta^{-1} b^2$ to $a = \sqrt{A_\tau}$ and $b = \frac{\sqrt{2}}{2} M\sqrt{x}$. □

We merge Lemmas 3.8 and 3.9 for convenience.

Lemma 3.10. *Suppose that Assumption 2.1 holds true. Take any $\lambda > 1$ and $\tau \in \mathcal{T}_n$ be a segmentation. Then, there exists an event $\Omega_{\tau,\lambda}^{(0)}$ of probability greater than $1 - 3\,\mathrm{e}^{-\lambda D_\tau}$ on which*:

$$\psi_\tau \geq \frac{1}{3} A_\tau - \frac{74}{3} \lambda D_\tau M^2 .$$

Proof. According to Lemma 3.9 with $\theta = 1/3$ and $x = \lambda D_\tau$, there exists an event $\Omega_{\tau,\lambda}^{(1)}$ on which

$$|L_\tau| \leq \frac{1}{3} A_\tau + \frac{17}{6} \lambda D_\tau M^2 ,$$

with $\mathbb{P}\left(\Omega_{\tau,\lambda}^{(1)}\right) \geq 1 - 2\,\mathrm{e}^{-\lambda D_\tau}$. Lemma 3.8 with $x = \lambda D_\tau$ gives $\Omega_{\tau,\lambda}^{(2)}$ on which

$$Q_\tau - v_\tau \leq \frac{14}{3} \left(\lambda + 2\sqrt{2\lambda}\right) D_\tau M^2 ,$$

with $\mathbb{P}\left(\Omega_{\tau,\lambda}^{(2)}\right) \geq 1 - \mathrm{e}^{-\lambda D_\tau}$. Then, $\Omega_{\tau,\lambda}^{(0)} := \Omega_{\tau,\lambda}^{(1)} \cap \Omega_{\tau,\lambda}^{(2)}$ has a probability larger than $1 - 3\,\mathrm{e}^{-\lambda D_\tau}$ by the union bound. Since for any $1 \leq \ell \leq D_\tau$, $v_{\tau,\ell} \leq M^2$, we have $v_\tau = \sum_{\ell=1}^{D_\tau} v_{\tau,\ell} \leq D_\tau M^2$. Hence, by definition (3.11) of ψ_τ and using that $\lambda \geq 1$, on the event $\Omega_{\tau,\lambda}^{(0)}$, we have:

$$\psi_\tau \geq \frac{1}{3} A_\tau - \left(\frac{31}{3}\lambda + \frac{28}{3}\sqrt{2}\sqrt{\lambda} + 1\right) D_\tau M^2$$

$$\geq \frac{1}{3}A_\tau - \lambda\left(\frac{31}{3} + \frac{28}{3}\sqrt{2} + 1\right) D_\tau M^2 \,.$$

□

Remark 3.2. It is also possible to obtain an upper bound for ψ_τ: by Lemma 3.9, for every $\lambda \geq 0$, on the event $\Omega^{(2)}_{\tau,\lambda} \subset \Omega^{(0)}_{\tau,\lambda}$,

$$\psi_\tau \leq \frac{5}{3}A_\tau + \frac{17}{3}\lambda D_\tau M^2 \,.$$

However, we do not need this result thereafter.

Concentration under Assumption 2.2. Lemma 3.6 and 3.7 directly translate upper bounds on B_n into controls of L_τ and Q_τ. Under Assumption 2.2, this is achieved via the following lemma, a Kolmogorov-like inequality for the noise in the RKHS. This result is a straightforward generalization of the inequality obtained by Kolmogorov [1928] into the Hilbert setting. A more precise result (for real random variables only) can be found in [Hájek and Rényi, 1955], of which we follow the proof. The scheme of Hájek and Rényi [1955] adapts well in our setting even though we do not need the full result.

Lemma 3.11. *If Assumption 2.2 holds true, then, for any $x > 0$,*

$$\mathbb{P}\left(B_n \geq x\right) \leq \frac{1}{x^2}\sum_{j=1}^{n} v_j \,. \tag{3.16}$$

We prove Lemma 3.11 in Section 3.5.6.

Remark 3.3. We can reformulate Lemma 3.11 as follows. For any $y > 0$, there exists an event of probability at least $1 - y^{-2}$ on which $B_n < y\sqrt{\sum_{i=j}^{n} v_j} \leq y\sqrt{nV}$. Equivalently, for any $z \geq 0$, there exists an event of probability at least $1 - \mathrm{e}^{-z}$ such that $B_n < \mathrm{e}^{z/2}\sqrt{\sum_{i=j}^{n} v_j} \leq \mathrm{e}^{z/2}\sqrt{nV}$.

3.4.4 Proof of Theorem 3.1

We follow the strategy described at the beginning of Section 3.4.

Definition of Ω. Let us define $\Omega := \bigcap_{\tau\in\mathcal{T}_n} \Omega^{(0)}_{\tau,\lambda}$ with $\lambda = y + \log(n) + 1 > 1$, where we recall that $\Omega^{(0)}_{\tau,\lambda}$ is defined in Lemma 3.10. By the union bound, and since the $\Omega^{(0)}_{\tau,\lambda}$ have probability greater than $1 - 3\,\mathrm{e}^{-\lambda D_\tau}$,

$$\mathbb{P}\left(\Omega\right) \geq 1 - 3\sum_{\tau\in\mathcal{T}_n} \mathrm{e}^{-\lambda D_\tau} \,.$$

The inequality $\mathbb{P}(\Omega) \geq 1 - \mathrm{e}^{-y}$ follows since

$$\begin{aligned}\sum_{\tau \in \mathcal{T}_n} \mathrm{e}^{-\lambda D_\tau} = \sum_{d=1}^{n} \binom{n-1}{d-1} \mathrm{e}^{-\lambda d} &= \mathrm{e}^{-\lambda} \left(1 + \mathrm{e}^{-\lambda}\right)^{n-1} \\ &\leq \mathrm{e}^{-\lambda} \exp\left((n-1)\,\mathrm{e}^{-\lambda}\right) \\ &= \frac{\mathrm{e}^{-y}}{n\,\mathrm{e}} \exp\left(\frac{n-1}{n}\,\mathrm{e}^{-1-y}\right) \\ &\leq \mathrm{e}^{-y}\,\frac{\exp\left(\mathrm{e}^{-1}\right)}{n\,\mathrm{e}} \leq 0.27\,\mathrm{e}^{-y}\,,\end{aligned}$$

where the last inequality uses that $n \geq 2$. From now on, we work exclusively on Ω.

Key argument. We now make the simple (but crucial) observation that $\mathrm{crit}(\tau^\star) \geq \mathrm{crit}(\widehat{\tau})$, hence

$$n\,\mathrm{pen}(\widehat{\tau}) + \psi_{\widehat{\tau}} \leq n\,\mathrm{pen}(\tau^\star) + \psi_{\tau^\star} \leq n\,\mathrm{pen}(\tau^\star) = CD^\star M^2\,.$$

Since we work on Ω, by definition of $\Omega^{(0)}_{\tau,\lambda}$ in Lemma 3.10, for any $\tau \in \mathcal{T}_n$, we have:

$$\psi_\tau \geq \frac{1}{3} A_\tau - \frac{74}{3}\lambda D_\tau M^2\,.$$

Therefore, we get:

$$CD^\star M^2 \geq \frac{1}{3} A_{\widehat{\tau}} + \left(C - \frac{74}{3}\lambda\right) \widehat{D} M^2\,. \tag{3.17}$$

Proof that $\widehat{D} \leq D^\star$. Since $C > 74\lambda/3$ (by the lower bound in assumption (3.3)), $M^2 > 0$ and $A_{\widehat{\tau}} \geq 0$, Eq. (3.17) implies that

$$\widehat{D} \leq \frac{C}{C - \frac{74}{3}\lambda} D^\star\,.$$

The lower bound in assumption (3.3) ensures that

$$\frac{C}{C - \frac{74}{3}\lambda} < \frac{D^\star + 1}{D^\star}\,,$$

hence $\widehat{D} \leq D^\star$ on Ω.

Proof that $\widehat{D} \geq D^\star$. Since $C > 74\lambda/3$ (by the lower bound in assumption (3.3)), Eq. (3.17) implies that $A_{\widehat{\tau}} \leq 3CD^\star M^2$. A direct consequence of (3.3) is that $A_{\widehat{\tau}} < \frac{1}{2} n \underline{\Lambda}_{\tau^\star} \underline{\Delta}^2$, hence $\widehat{D} \geq D^\star$ by Lemma 3.3.

Loss between $\widehat{\tau}$ and $\tau^\star$. We have proved that $\widehat{D} = D^\star$ on Ω, therefore, Eq. (3.17) can be rewritten

$$A_{\widehat{\tau}} \leq 74\lambda D^\star M^2 \,.$$

By Lemma 3.4 and the definition of λ, we get

$$\min\left\{\underline{\Lambda}_{\tau^\star}, \frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^\star, \widehat{\tau}\right)\right\} \leq \frac{148 D^\star M^2}{\underline{\Delta}^2} \cdot \frac{y + \log(n) + 1}{n} = v_1(y)\,. \tag{3.18}$$

Remark that Assumption (3.3) implies that

$$\frac{\underline{\Delta}^2}{M^2}\frac{\underline{\Lambda}_{\tau^\star}}{6D^\star} n > \frac{74}{3}(D^\star + 1)(y + \log(n) + 1)$$

hence

$$\underline{\Lambda}_{\tau^\star} > (D^\star + 1)\frac{148 D^\star M^2}{\underline{\Delta}^2} \cdot \frac{y + \log(n) + 1}{n} > v_1(y)\,.$$

Therefore, Eq. (3.18) can be simplified into

$$\frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^\star, \widehat{\tau}\right) \leq v_1(y)\,. \quad \square$$

Remark 3.4. The proof of Theorem 3.1 generalizes to $\widehat{\tau}$ defined by

$$\widehat{\tau} \in \underset{\tau \in \mathcal{T}_n \,/\, \underline{\Lambda}_\tau \geq \delta_n}{\arg\min} \left\{\mathrm{crit}\left(\tau\right)\right\}$$

instead of (2.2), for any $\delta_n \geq 0$ such that $\underline{\Lambda}_{\tau^\star} \geq \delta_n$. Indeed, this assumption allows to write $\mathrm{crit}\left(\tau^\star\right) \geq \mathrm{crit}\left(\widehat{\tau}\right)$ in the key argument, and the rest of the proof can stay unchanged (with the same event Ω). More generally, any constraint can be added in the minimization problem defining $\widehat{\tau}$, provided that $\tau^\star$ satisfies this constraint.

3.4.5 Proof of Theorem 3.2

The structure of the proof is very similar to the proof of Theorem 3.1, the main difference being the need for a more precise control of the penalty function near $D^\star$. We refer to Section 3.5.7 for all the technical results required for this control.

Definition of Ω. Set $y > 0$. For any τ, we set $\lambda_\tau := y + \log\frac{n}{D_\tau} + 4$ and $\Omega_{\tau,\lambda_\tau}$ as in Lemma 3.10. Define $\Omega := \bigcap_{\tau \in \mathcal{T}_n} \Omega_{\tau,\lambda_\tau}$. By the union bound, and since the $\Omega_{\tau,\lambda_\tau}$ have probability greater than $1 - 3\,\mathrm{e}^{-\lambda_\tau D_\tau}$,

$$\mathbb{P}\left(\Omega\right) = \mathbb{P}\left(\bigcap_{\tau \in \mathcal{T}_n} \Omega_{\tau,\lambda_\tau}\right) \geq 1 - 3\sum_{\tau \in \mathcal{T}_n} \mathrm{e}^{-\lambda_\tau D_\tau}\,.$$

Recall that, for any $1 \leq d \leq n$, $\binom{n-1}{d-1} \leq \binom{n}{d} \leq (n\,\mathrm{e}\,/d)^d$, hence

$$\begin{aligned}
\sum_{\tau \in \mathcal{T}_n} \mathrm{e}^{-\lambda_\tau D_\tau} &= \sum_{d=1}^{n} \binom{n-1}{d-1} \mathrm{e}^{-d(y+\log(n/d)+4)} \\
&\leq \sum_{d=1}^{n} \exp\left(d + d\log(n) - d\log(d) - 4d - dy - d\log(n) + d\log(d)\right) \\
&= \sum_{d=1}^{n} \exp\left(d(-3-y)\right) = \sum_{d=1}^{n} \left(\frac{\mathrm{e}^{-y}}{\mathrm{e}^3}\right)^d \\
&= \mathrm{e}^{-3-y} . \frac{1-(\mathrm{e}^{-3-y})^n}{1-\mathrm{e}^{-3-y}} \leq \mathrm{e}^{-y}/3.
\end{aligned}$$

The last inequality holds because $3\,\mathrm{e}^{-3} \leq 1$ and $\mathrm{e}^{-3-y} < 1$, and we have proved that $\mathbb{P}(\Omega) \geq 1 - \mathrm{e}^{-y}$. We now work solely on the event Ω.

Proof that $\widehat{D} \leq D^\star$ on Ω. Let τ be a segmentation such that $D_\tau > D^\star$. Let us show that $\mathrm{crit}(\tau) > \mathrm{crit}(\tau^\star)$. In this case, we do not have a better lower bound than 0 for the approximation error A_τ. We still can use the fact that

$$\psi_\tau \geq \frac{-74}{3} \lambda_\tau D_\tau M^2 ,$$

and we know that $\psi_{\tau^\star} \leq 0$. Thus

$$\begin{aligned}
n\left(\mathrm{crit}(\tau) - \mathrm{crit}(\tau^\star)\right) &= \psi_\tau + n\,\mathrm{pen}_{\mathrm{L}}(\tau) - \psi_{\tau^\star} - n\,\mathrm{pen}_{\mathrm{L}}(\tau^\star) \\
&\geq \frac{-74}{3}\left(y + \log\frac{n}{D_\tau} + 4\right) D_\tau M^2 + D_\tau\left(c_1 \log\frac{n}{D_\tau} + c_2\right) \\
&\qquad - D^\star\left(c_1 \log\frac{n}{D^\star} + c_2\right) \\
&\geq D_\tau\left(\left(c_1 - \frac{74}{3}M^2\right)\log\frac{n}{D_\tau} + \left(c_2 - \frac{74M^2}{3}(y+4)\right)\right) \\
&\qquad - D^\star\left(c_1 \log\frac{n}{D^\star} + c_2\right) .
\end{aligned}$$

Let us set $a := \left(c_1 - \frac{74}{3}M^2\right)$ and $b := c_2 - \frac{74M^2}{3}(y+4)$. Given that $c_1 > c_{1,\min}$ and $c_2 = (c_1 + 74M^2(D^\star+1)/3)(y+4)$, we have $a > 0$ and $b > 0$. Moreover, $b - a = 74M^2(D^\star(y+4)+1)/3 + c_1(3+y)$ is a positive quantity. Therefore, according to Lemma 3.13 together with $D_\tau > D^\star$,

$$\begin{aligned}
n\left(\mathrm{crit}(\tau) - \mathrm{crit}(\tau^\star)\right) \geq (D^\star+1)\left[\left(c_1 - \tfrac{74}{3}M^2\right)\log\tfrac{n}{D^\star+1} + \left(c_2 - \tfrac{74M^2}{3}(y+4)\right)\right] \\
- D^\star\left(c_1 \log\frac{n}{D^\star} + c_2\right) .
\end{aligned}$$

After rewriting the right-hand side of the last display, we obtain

$$\begin{aligned} n\left(\operatorname{crit}(\tau)-\operatorname{crit}(\tau^\star)\right) \geq c_1\big(\log(n)+D^\star\log D^\star-(D^\star+1)\log\left(D^\star+1\right)\big) \\ -\tfrac{74M^2}{3}\left(D^\star+1\right)\log\tfrac{n}{D^\star+1}+\Big(c_2-\tfrac{74M^2}{3}\left(D^\star+1\right)(y+4)\Big)\,. \end{aligned}$$

We now use the relationship between c_2 and c_1, together with Lemma 3.14 applied to $x = D^\star$ to write

$$\begin{aligned} n\left(\operatorname{crit}(\tau)-\operatorname{crit}(\tau^\star)\right) \geq c_1\big(2+y+\log(n)-\log D^\star\big) \\ -\frac{74M^2}{3}\left(D^\star+1\right)\left(\log(n)-\log\left(D^\star+1\right)\right)\,. \end{aligned}$$

This last display is positive if

$$c_1 > \frac{\frac{74M^2}{3}\left(D^\star+1\right)\left(\log(n)-\log\left(D^\star+1\right)\right)}{2+y+\log(n)-\log\left(D^\star\right)}\,.$$

This is indeed the case according to Lemma 3.15 applied to $x = D^\star$ together with $c_1 > c_{1,\min}$, and we can conclude.

Proof that $\widehat{D} \geq D^\star$ on Ω. Let $\tau \in \mathcal{T}_n$ be such that $D_\tau < D^\star$. We are going to show that $\operatorname{crit}(\tau) > \operatorname{crit}(\tau^\star)$. By construction of Ω, it holds that

$$\psi_\tau \geq \frac{1}{3}A_\tau - \frac{74\lambda_\tau D_\tau}{3}M^2\,.$$

Because of Lemma 3.3, this leads to

$$\psi_\tau \geq \frac{n\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2}{6} - \frac{74\lambda_\tau D_\tau}{3}M^2\,.$$

According to Lemma 3.13, and since $D_\tau < D^\star$,

$$\psi_\tau > \frac{n\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2}{6} - \frac{74D^\star}{3}\left(y+\log\frac{n}{D^\star}+4\right)M^2\,.$$

On the other side, $\psi_{\tau^\star} \leq 0$, and

$$n\operatorname{pen}_{\mathrm{L}}(\tau^\star) - n\operatorname{pen}_{\mathrm{L}}(\tau) < D^\star\left(c_1\log\frac{n}{D^\star}+c_2\right)\,.$$

Thus

$$\begin{aligned} n\left(\operatorname{crit}(\tau)-\operatorname{crit}(\tau^\star)\right) &= \psi_\tau + n\operatorname{pen}_{\mathrm{L}}(\tau) - \psi_{\tau^\star} - n\operatorname{pen}_{\mathrm{L}}(\tau^\star) \\ &> \frac{n\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2}{6} - \frac{74D^\star}{3}\left(y+\log\frac{n}{D^\star}+4\right)M^2 \\ &\quad - D^\star\left(c_1\log\frac{n}{D^\star}+c_2\right)\,. \end{aligned}$$

Let us use the relationship between c_1 and c_2. The last display is positive if

$$\begin{aligned}\frac{n\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2}{6D^\star} &> c_1\left(y+\log(n)-\log\left(D^\star\right)+4\right)+\frac{74}{3}M^2\left(y+\log\frac{n}{D^\star}+4\right)\\ &\quad+\frac{74}{3}M^2\left(D^\star+1\right)(y+4)\\ &= c_1\left(y+\log\frac{n}{D^\star}+4\right)+\frac{74}{3}M^2\left((y+4)\left(D^\star+2\right)+\log\frac{n}{D^\star}\right),\end{aligned}$$

which is true by Lemma 3.16 applied to $x = D^\star$ together with the definition of $c_{1,\max}$.

At this point, we have proved the first part of Theorem 3.2: on the event Ω, $\widehat{D} = D^\star$. We now address the problem of finding an upper bound for $\frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^\star,\widehat{\tau}\right)$.

Loss between $\widehat{\tau}$ and $\tau^\star$. Since $\widehat{D} = D^\star$, we have $\mathrm{pen}_{\mathrm{L}}\left(\widehat{\tau}\right) = \mathrm{pen}_{\mathrm{L}}\left(\tau^\star\right)$. Therefore, $\psi_{\widehat{\tau}} \le \psi_{\tau^\star}$. Since $\psi_{\tau^\star} \le 0$ and $\psi_{\widehat{\tau}} \ge \frac{1}{3}A_\tau - \frac{74}{3}\lambda_{\widehat{\tau}}\widehat{D}M^2$, we have

$$\frac{1}{3}A_{\widehat{\tau}} \le \frac{74}{3}\lambda_{\widehat{\tau}}\widehat{D}M^2 \le \frac{74D^\star M^2}{3}\left(y+\log(n)+4\right).$$

By Lemma 3.4,

$$\min\left\{\underline{\Lambda}_{\tau^\star}, \frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^\star,\widehat{\tau}\right)\right\} \le \frac{148D^\star M^2}{\underline{\Delta}^2}\cdot\frac{y+\log(n)+4}{n} = v_1'(y)\,. \tag{3.19}$$

Rewriting $c_{1,\min} < c_{1,\max}$ as a condition on $\underline{\Lambda}_{\tau^\star}$, it follows that

$$\underline{\Lambda}_{\tau^\star} > \frac{148M^2D^\star\left(2D^\star+3\right)(y+\log(n)+4)}{n\underline{\Delta}^2} = v_1'\cdot\left(2D^\star+3\right).$$

From $D^\star \ge 1$,we deduce that $\underline{\Lambda}_{\tau^\star} > v_1'$. Eq. (3.19) now yields

$$\frac{1}{n}\,\mathrm{d}_\infty^{(1)}\left(\tau^\star,\widehat{\tau}\right) < v_1'(y)\,,$$

which concludes the proof of Theorem 3.2. □

3.4.6 Proof of Theorem 3.3

We follow the strategy described at the beginning of Section 3.4. Throughout the proof, we write $\widehat{\tau}_2$ as a shortcut for $\widehat{\tau}(D^\star,\delta_n)$.

Key argument. By definition (3.7) of $\widehat{\tau}_2 = \widehat{\tau}(D^\star,\delta_n)$, since we assume $\underline{\Lambda}_{\tau^\star} \ge \delta_n$,

$$\widehat{\mathcal{R}}_n(\tau^\star) \ge \widehat{\mathcal{R}}_n(\widehat{\tau}_2)$$

hence

$$0 \ge \psi_{\tau^\star} \ge \psi_{\widehat{\tau}_2} = A_{\widehat{\tau}_2} + 2L_{\widehat{\tau}_2} - Q_{\widehat{\tau}_2}\,.$$

By Lemma 3.6, Lemma 3.7 and the facts that $D_{\widehat{\tau}_2} = D^\star$ and $\underline{\Lambda}_{\widehat{\tau}_2} \geq \delta_n$, we get

$$0 \geq \psi_{\widehat{\tau}_2} \geq A_{\widehat{\tau}_2} - 12\,(D^\star)^2\,\overline{\Delta} B_n - \frac{4D^\star B_n^2}{n\delta_n}\,.$$

Hence, using Lemma 3.4,

$$\min\left\{\underline{\Lambda}_{\tau^\star}, \frac{1}{n}\,\mathrm{d}_\infty^{(1)}\,(\tau^\star, \widehat{\tau}_2)\right\} \leq \frac{24\,(D^\star)^2\,\overline{\Delta}}{\underline{\Delta}^2}\frac{B_n}{n} + \frac{8D^\star}{\underline{\Delta}^2}\frac{B_n^2}{n^2\delta_n}\,. \tag{3.20}$$

Definition of Ω_2. We define

$$\Omega_2 := \{B_n \leq y\sqrt{nV}\}\,.$$

By Lemma 3.11, under Assumption 2.2, $\mathbb{P}\,(\Omega_2) \geq 1 - y^{-2}$.

Conclusion. By definition of Ω_2, Eq. (3.20) implies that on Ω_2:

$$\min\left\{\underline{\Lambda}_{\tau^\star}, \frac{1}{n}\,\mathrm{d}_\infty^{(1)}\,(\tau^\star, \widehat{\tau}_2)\right\} \leq 24(D^\star)^2\frac{\overline{\Delta}\sqrt{V}}{\underline{\Delta}^2}\frac{y}{\sqrt{n}} + 8D^\star\frac{V}{\underline{\Delta}^2}\frac{y^2}{n\delta_n} = v_2(y, \delta_n)\,.$$

Since we assume $v_2(y, \delta_n) < \underline{\Lambda}_{\tau^\star}$, the result follows. □

3.5 Additional proofs

In this section are collected a large part of the technical details of the proofs that precede. Some additional notation used solely in this section are introduced below.

We denote by $\lambda_1^\star, \ldots, \lambda_{D^\star}^\star$ the segments of $\tau^\star$, that is,

$$\lambda_i^\star = \left\{\tau_{i-1}^\star + 1, \ldots, \tau_i^\star\right\}.$$

For any segment λ of $\tau \in \mathcal{T}_n$, we denote by $\mu_\lambda^\star$ the value of $\mu_\tau^\star$ on λ, which does not depend on τ and is given by (2.7):

$$\mu_\lambda^\star = \frac{1}{|\lambda|}\sum_{j\in\lambda}\mu_j^\star\,. \tag{3.21}$$

We will sometimes write $\sum_{\lambda\in\tau}$ instead of the more cumbersome $\sum_{\ell=1}^{D_\tau}$, when the dependency in τ. is not apparent in the summation. More generally, we will abuse the notation τ to denote the set of segments associated to the segmentation τ.

3.5.1 Proof of Lemma 3.1

Proof of (i). We set $D^i := D_{\tau^i}$ for $i \in \{1,2\}$. Let us show first that $\mathrm{d}_\infty^{(2)}(\tau^1,\tau^2) = \mathrm{d}_\infty^{(1)}(\tau^1,\tau^2)$. Take any $i \in \{1,\dots,D^1-1\}$, by the definition of $\underline{\Lambda}_{\tau^1}$,

$$\left|\tau_i^1 - \tau_{D^2}^2\right| = \left|\tau_i^1 - n\right| \geq n\underline{\Lambda}_{\tau^1} > n\underline{\Lambda}_{\tau^1}/2 \geq n\min\left\{\underline{\Lambda}_{\tau^1}, \underline{\Lambda}_{\tau^2}\right\}/2\,,$$

which is greater than $\mathrm{d}_\infty^{(1)}(\tau^1,\tau^2)$ by assumption. In the same fashion we can prove that $|\tau_i^1 - \tau_0^2| > \mathrm{d}_\infty^{(1)}(\tau^1,\tau^2)$. Hence, for any $i \in \{1,\dots,D^1-1\}$,

$$\min_{0\leq j\leq D^2}\left|\tau_i^1 - \tau_j^2\right| = \min_{1\leq j\leq D^2-1}\left|\tau_i^1 - \tau_j^2\right|,$$

which proves that $\mathrm{d}_\infty^{(2)}(\tau^1,\tau^2) = \mathrm{d}_\infty^{(1)}(\tau^1,\tau^2)$.

Next, we prove that $D^1 = D^2$ and $\mathrm{d}_\infty^{(3)}(\tau^1,\tau^2) = \mathrm{d}_\infty^{(1)}(\tau^1,\tau^2)$. Define the mapping $\phi : \{1,\dots,D^1-1\} \to \{1,\dots,D^2-1\}$ such that

$$\left\{\phi(i)\right\} = \operatorname*{arg\,min}_{1\leq j\leq D^2-1}\left|\tau_i^1 - \tau_j^2\right|$$

for all $i \in \{1,\dots,D^1-1\}$. This mapping is well-defined: indeed, suppose that $j,k \in \{1,\dots,D^2-1\}$ both realize the minimum for some $i \in \{1,\dots,D^1-1\}$. Since we assumed $\frac{1}{n}\mathrm{d}_\infty^{(1)}(\tau^1,\tau^2) < \min\{\underline{\Lambda}_{\tau^1},\underline{\Lambda}_{\tau^2}\}/2$,

$$\left|\tau_i^1 - \tau_j^2\right| = \left|\tau_i^1 - \tau_k^2\right| \leq \mathrm{d}_\infty^{(1)}\left(\tau^1,\tau^2\right) < n\min\left\{\underline{\Lambda}_{\tau^1},\underline{\Lambda}_{\tau^2}\right\}/2.$$

By the triangle inequality,

$$\left|\tau_j^2 - \tau_k^2\right| < n\min\left\{\underline{\Lambda}_{\tau^1},\underline{\Lambda}_{\tau^2}\right\} \leq n\underline{\Lambda}_{\tau^2},$$

hence $j = k$. Next, we show that ϕ is increasing. Take $i,j \in \{1,\dots,D^1-1\}$ such that $i < j$. Recall that $\tau_\cdot^k$ is increasing ($k = 1,2$). Then

$$\begin{aligned}
\tau_{\phi(i)}^2 - \tau_{\phi(j)}^2 &= \tau_{\phi(i)}^2 - \tau_i^1 + \tau_i^1 - \tau_j^1 + \tau_j^1 - \tau_{\phi(j)}^2 \\
&= \tau_{\phi(i)}^2 - \tau_i^1 - \left|\tau_i^1 - \tau_j^1\right| + \tau_j^1 - \tau_{\phi(j)}^2 \\
&\leq \left|\tau_{\phi(i)}^2 - \tau_i^1\right| - \left|\tau_i^1 - \tau_j^1\right| + \left|\tau_j^1 - \tau_{\phi(j)}^2\right| \\
&\leq 2\,\mathrm{d}_\infty^{(1)}\left(\tau^1,\tau^2\right) - \left|\tau_i^1 - \tau_j^1\right| \\
&< n\min\left\{\underline{\Lambda}_{\tau^1},\underline{\Lambda}_{\tau^2}\right\} - n\underline{\Lambda}_{\tau^1} \leq 0\,.
\end{aligned}$$

Hence $\phi(i) < \phi(j)$, so ϕ is increasing. As a consequence, ϕ is injective and we get $D^1 \leq D^2$. The same argument, exchanging τ^1 and τ^2, shows that $D^2 \leq D^1$. Therefore, $D^1 = D^2$ and ϕ is an increasing permutation of $\{1,\dots,D^1-1\}$, hence it is the identity. Thus $\mathrm{d}_\infty^{(3)}(\tau^1,\tau^2) = \mathrm{d}_\infty^{(1)}(\tau^1,\tau^2)$.

Finally, since $\mathrm{d}_\infty^{(3)}$ is symmetric, $\mathrm{d}_\infty^{(i)}(\tau^1,\tau^2) = \mathrm{d}_{\mathrm{H}}^{(i)}(\tau^1,\tau^2)$ for any $i \in \{1,2,3\}$.

Proof of (ii). Since $D_{\tau^1} = D_{\tau^2}$, we can set $D = D_{\tau^1} = D_{\tau^2}$. Next, define $\phi(i) := \arg\min_{1\leq j\leq D-1} \left|\tau_i^1 - \tau_j^2\right|$ and $C_\phi(i) := |\phi(i)|$ for all $i \in \{1,\dots,D-1\}$. Clearly, $C_\phi(i) \geq 1$ for any i. Let us show that we actually have $C_\phi(i) = 1$.

Take i and j distinct elements of $\{1,\dots,D-1\}$, and suppose that $\phi(i) \cap \phi(j)$ is non-empty. Let k be any element of $\phi(i) \cap \phi(j)$. By the triangle inequality and the definition of $\mathrm{d}_\infty^{(1)}$,

$$n\underline{\Lambda}_{\tau^1} \leq \left|\tau_i^1 - \tau_j^1\right| \leq \left|\tau_i^1 - \tau_k^2\right| + \left|\tau_k^2 - \tau_j^1\right| \leq 2\,\mathrm{d}_\infty^{(1)}\left(\tau^1,\tau^2\right) < n\underline{\Lambda}_{\tau^1}\,.$$

Hence, the $\phi(i)$ are disjoint and we can write $\sum_{i=1}^{D-1} C_\phi(i) = D-1$, which clearly implies that $C_\phi(i) = 1$.

From now on, we identify $\phi(i)$ with its unique element. Let us show that ϕ is increasing similarly to what we have done for proving (i). Take $i,j \in \{1,\dots,D-1\}$ such that $i<j$. We showed that

$$\tau^2_{\phi(i)} - \tau^2_{\phi(j)} \leq 2\,\mathrm{d}_\infty^{(1)}\left(\tau^1,\tau^2\right) - \left|\tau_i^1 - \tau_j^1\right|,$$

thus according to the definition of $\underline{\Lambda}_{\tau^1}$, and our assumption,

$$\tau^2_{\phi(i)} - \tau^2_{\phi(j)} < n\underline{\Lambda}_{\tau^1} - n\underline{\Lambda}_{\tau^1} \leq 0\,.$$

Hence $\phi(i) < \phi(j)$: ϕ is increasing. As a consequence,

$$\mathrm{d}_\infty^{(1)}\left(\tau^1,\tau^2\right) = \mathrm{d}_\infty^{(1)}\left(\tau^2,\tau^1\right) = \mathrm{d}_{\mathrm{H}}^{(1)}\left(\tau^1,\tau^2\right)\,.$$

□

3.5.2 The Frobenius loss

A formula for $\mathrm{d}_{\mathrm{F}}^2$

We start by proving a general formula for d_{F}, which is stated by Lajugie et al. [2014], we prove it here for completeness:

$$\forall \tau^1,\tau^2 \in \mathcal{T}_n, \qquad \mathrm{d}_{\mathrm{F}}\left(\tau^1,\tau^2\right)^2 = D_{\tau^1} + D_{\tau^2} - 2\sum_{k=1}^{D_{\tau^1}}\sum_{\ell=1}^{D_{\tau^2}} \frac{\left|\lambda_k^1 \cap \lambda_\ell^2\right|^2}{|\lambda_k^1| \times |\lambda_\ell^2|}\,. \tag{3.22}$$

Indeed, by definition, we have

$$\mathrm{d}_{\mathrm{F}}\left(\tau^1,\tau^2\right)^2 = \mathrm{Tr}\big((\Pi_{\tau^1} - \Pi_{\tau^2})^2\big) = \underbrace{\mathrm{Tr}(\Pi_{\tau^1})}_{=D_{\tau^1}} + \underbrace{\mathrm{Tr}(\Pi_{\tau^2})}_{=D_{\tau^2}} - 2\,\mathrm{Tr}(\Pi_{\tau^1}\Pi_{\tau^2})$$

$$\text{and} \quad \mathrm{Tr}(\Pi_{\tau^1}\Pi_{\tau^2}) = \sum_{i=1}^n\sum_{j=1}^n \frac{\mathbb{1}_{\lambda_1(i)=\lambda_1(j) \text{ and } \lambda_2(i)=\lambda_2(j)}}{|\lambda_1(i)|\,|\lambda_2(i)|} = \sum_{k=1}^{D_{\tau^1}}\sum_{\ell=1}^{D_{\tau^2}} \frac{\left|\lambda_k^1 \cap \lambda_\ell^2\right|^2}{|\lambda_k^1| \times |\lambda_\ell^2|}\,,$$

where we denoted by $\lambda_k(i)$ the segment of τ^k to which $i \in \{1,\dots,n\}$ belongs.

Proof of Eq. (3.5)

Eq. (3.5) is stated by Lajugie et al. [2014]. The upper bound is a straightforward consequence of Eq. (3.22). We prove the lower bound here for completeness. We remark that

$$\sum_{k=1}^{D_{\tau^1}} \sum_{\ell=1}^{D_{\tau^2}} \frac{|\lambda_k^1 \cap \lambda_\ell^2|^2}{|\lambda_k^1| \times |\lambda_\ell^2|} \leq \sum_{k=1}^{D_{\tau^1}} \sum_{\ell=1}^{D_{\tau^2}} \frac{|\lambda_k^1 \cap \lambda_\ell^2|}{|\lambda_k^1|} = D_{\tau^1} \,,$$

hence Eq. (3.22) shows that

$$\mathrm{d_F}\left(\tau^1, \tau^2\right)^2 \geq D_{\tau^2} - D_{\tau^1} \,.$$

The lower bound follows since τ^1 and τ^2 play symmetric roles. □

Proof of Prop. 3.1

Throughout the proof, we write $D = D_{\tau^1} = D_{\tau^2}$, $\varepsilon = n^{-1}\, \mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right)$ and we denote by $(\lambda_k^1)_{1 \leq k \leq D}$ and $(\lambda_k^2)_{1 \leq k \leq D}$ the segments of τ^1 and τ^2, respectively.

Preliminary remark. Since we assume that $D_{\tau^1} = D_{\tau^2}$ and $\frac{1}{n}\, \mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right) = \varepsilon < \underline{\Lambda}_{\tau^1}/2$, point (ii) in Lemma 3.1 shows that $\mathrm{d}_\infty^{(1)}\left(\tau^1, \tau^2\right) = \mathrm{d}_{\mathrm{H}}^{(1)}\left(\tau^1, \tau^2\right) = \mathrm{d}_\infty^{(3)}(\tau^1, \tau^2)$. In other words, for every $k \in \{1, \ldots, D-1\}$, we have $|\tau_k^1 - \tau_k^2| \leq n\varepsilon$, and some $k_0 \in \{1, \ldots, D-1\}$ exists such that $\left|\tau_{k_0}^1 - \tau_{k_0}^2\right| = n\varepsilon$. As a consequence, for every $k \in \{1, \ldots, D-1\}$,

$$\left|\left|\lambda_k^1\right| - \left|\lambda_k^2\right|\right| \leq 2n\varepsilon \qquad \text{and} \qquad \left|\lambda_k^1 \cap \lambda_k^2\right| \geq \left|\lambda_k^1\right| - 2n\varepsilon \,. \tag{3.23}$$

Upper bound for $\mathrm{d_F}\left(\tau^1, \tau^2\right)^2$**.** We focus on the sum appearing in the right-hand side of Eq. (3.22). Using Eq. (3.23), we get:

$$\begin{aligned}
\sum_{k=1}^{D} \sum_{\ell=1}^{D} \frac{|\lambda_k^1 \cap \lambda_\ell^2|^2}{|\lambda_k^1| \times |\lambda_\ell^2|} &\geq \sum_{k=1}^{D} \frac{|\lambda_k^1 \cap \lambda_k^2|^2}{|\lambda_k^1| \times |\lambda_k^2|} \\
&\geq \sum_{k=1}^{D} \left[\frac{\left(|\lambda_k^1| - 2n\varepsilon\right)^2}{|\lambda_k^1| \times \left(|\lambda_k^1| + 2n\varepsilon\right)}\right] = \sum_{k=1}^{D} \frac{\left(1 - \frac{2n\varepsilon}{|\lambda_k^1|}\right)^2}{1 + \frac{2n\varepsilon}{|\lambda_k^1|}} \\
&\geq \sum_{k=1}^{D} \left(1 - \frac{6n\varepsilon}{|\lambda_k^1|}\right) \geq D - \frac{6\varepsilon D}{\underline{\Lambda}_{\tau^1}} \,,
\end{aligned}$$

since for any $x \geq 0$, $\frac{(1-x)^2}{1+x} \geq 1 - 3x$. The upper bound follows, using Eq. (3.22).

Lower bound for $\mathrm{d_F}\left(\tau^1, \tau^2\right)^2$**.** As shown in the preliminary remark, there exists some $k_0 \in \{1, \ldots, D-1\}$ such that $\left|\tau_{k_0}^1 - \tau_{k_0}^2\right| = n\varepsilon$. First consider the case where

$\tau^1_{k_0} < \tau^2_{k_0}$. Then, by definition of $\mathrm{d_F}$ and Π_τ, we have:

$$\begin{aligned}
\mathrm{d_F}\left(\tau^1, \tau^2\right)^2 &:= \sum_{1 \le i,j \le n} \left(\Pi_{\tau^1} - \Pi_{\tau^2}\right)^2_{i,j} \\
&\geq \sum_{i \in \lambda^1_{k_0+1} \cap \lambda^2_{k_0}} \; \sum_{j \in \lambda^1_{k_0+1} \cap \lambda^2_{k_0+1}} \frac{1}{\left|\lambda^1_{k_0+1}\right|^2} \\
&\qquad + \sum_{i \in \lambda^1_{k_0+1} \cap \lambda^2_{k_0+1}} \; \sum_{j \in \lambda^1_{k_0+1} \cap \lambda^2_{k_0}} \frac{1}{\left|\lambda^1_{k_0+1}\right|^2} \\
&= \frac{2\left|\lambda^1_{k_0+1} \cap \lambda^2_{k_0}\right| \cdot \left|\lambda^1_{k_0+1} \cap \lambda^2_{k_0+1}\right|}{\left|\lambda^1_{k_0+1}\right|^2} .
\end{aligned}$$

Now, remark that $\left|\lambda^1_{k_0+1} \cap \lambda^2_{k_0}\right| = n\varepsilon$, by the preliminary remark and our assumption $\tau^2_{k_0} > \tau^1_{k_0}$. Using also Eq. (3.23), we get:

$$\mathrm{d_F}\left(\tau^1, \tau^2\right)^2 \geq \frac{2n\varepsilon\left(\left|\lambda^1_{k_0+1}\right| - 2n\varepsilon\right)}{\left|\lambda^1_{k_0+1}\right|^2} \geq \frac{2n\varepsilon}{3\overline{\Lambda}_{\tau^1}} ,$$

since $\left|\lambda^1_{k_0+1}\right| - 2n\varepsilon \geq \left|\lambda^1_{k_0+1}\right|/3$ and $\left|\lambda^1_{k_0+1}\right| \leq \overline{\Lambda}_{\tau^1}$. When $\tau^1_{k_0} > \tau^2_{k_0}$, we apply the same reasoning, restricting the sum over i, j in the definition of $\mathrm{d_F}$ to $i \in \lambda^1_{k_0} \cap \lambda^2_{k_0}$ and $j \in \lambda^1_{k_0} \cap \lambda^2_{k_0+1}$ (plus its symmetric). We obtain the same lower bound, which concludes the proof. □

3.5.3 Lower bounds on the approximation error

This section provides the proofs of Lemmas 3.3 and 3.4.

Preliminary lemma

We start with a lemma useful in the two proofs.

Lemma 3.12. *If a segment $\lambda \subset \{1, \dots, n\}$ intersects only two segments of $\tau^\star$, $\lambda^\star_i$ and $\lambda^\star_{i+1}$, then we have*:

$$\sum_{j \in \lambda} \left\|\mu^\star_j - \mu^\star_\lambda\right\|^2_{\mathcal{H}} = \frac{\left|\lambda \cap \lambda^\star_i\right| \cdot \left|\lambda \cap \lambda^\star_{i+1}\right|}{\left|\lambda \cap \lambda^\star_i\right| + \left|\lambda \cap \lambda^\star_{i+1}\right|} \left\|\mu^\star_{\lambda^\star_{i+1}} - \mu^\star_{\lambda^\star_i}\right\|^2_{\mathcal{H}} \tag{3.24}$$

$$\geq \left(\frac{\left|\lambda \cap \lambda^\star_i\right|}{\left|\lambda^\star_i\right|} \wedge \frac{\left|\lambda \cap \lambda^\star_{i+1}\right|}{\left|\lambda^\star_{i+1}\right|}\right) \cdot \frac{\left|\lambda^\star_i\right| \cdot \left|\lambda^\star_{i+1}\right|}{\left|\lambda^\star_i\right| + \left|\lambda^\star_{i+1}\right|} \cdot \left\|\mu^\star_{\lambda^\star_{i+1}} - \mu^\star_{\lambda^\star_i}\right\|^2_{\mathcal{H}} . \tag{3.25}$$

Proof. We first prove Eq. (3.24). Since λ only intersects $\lambda^\star_i$ and $\lambda^\star_{i+1}$, we have:

$$\sum_{j \in \lambda} \left\|\mu^\star_j - \mu^\star_\lambda\right\|^2_{\mathcal{H}} = \sum_{j \in \lambda \cap \lambda^\star_i} \left\|\mu^\star_j - \mu^\star_\lambda\right\|^2_{\mathcal{H}} + \sum_{j \in \lambda \cap \lambda^\star_{i+1}} \left\|\mu^\star_j - \mu^\star_\lambda\right\|^2_{\mathcal{H}}$$

$$= |\lambda \cap \lambda_i^\star| \cdot \left\| \mu^\star_{\lambda_i^\star} - \mu^\star_\lambda \right\|^2_{\mathcal{H}} + \left|\lambda \cap \lambda^\star_{i+1}\right| \cdot \left\| \mu^\star_{\lambda^\star_{i+1}} - \mu^\star_\lambda \right\|^2_{\mathcal{H}} . \qquad (3.26)$$

Since $\mu^\star_\lambda$ is given by Eq. (3.21), we obtain

$$\left\| \mu^\star_{\lambda_i^\star} - \mu^\star_\lambda \right\|^2_{\mathcal{H}} = \left\| \frac{1}{|\lambda|} \sum_{j \in \lambda} \left(\mu^\star_{\lambda_i^\star} - \mu^\star_j \right) \right\|^2_{\mathcal{H}} = \left\| \frac{1}{|\lambda|} \sum_{j \in \lambda \cap \lambda^\star_{i+1}} \left(\mu^\star_{\lambda_i^\star} - \mu^\star_{\lambda^\star_{i+1}} \right) \right\|^2_{\mathcal{H}}$$
$$= \frac{\left|\lambda \cap \lambda^\star_{i+1}\right|^2}{|\lambda|^2} \left\| \mu^\star_{\lambda^\star_{i+1}} - \mu^\star_{\lambda_i^\star} \right\|^2_{\mathcal{H}} .$$

The same computation on $\lambda \cap \lambda^\star_{i+1}$ yields

$$\left\| \mu^\star_{\lambda^\star_{i+1}} - \mu^\star_\lambda \right\|^2_{\mathcal{H}} = \frac{|\lambda \cap \lambda_i^\star|^2}{|\lambda|^2} \left\| \mu^\star_{\lambda^\star_{i+1}} - \mu^\star_{\lambda_i^\star} \right\|^2_{\mathcal{H}} .$$

Therefore, Eq. (3.26) and the fact that $|\lambda| = |\lambda \cap \lambda_i^\star| + \left|\lambda \cap \lambda^\star_{i+1}\right|$ yield Eq. (3.24).

Now, we remark that for any $a, b, c, d > 0$,

$$\frac{abcd}{ab+cd} = \frac{1}{\frac{ab}{\max(a,c)} + \frac{cd}{\max(a,c)}} \times \min(a,c) \times bd \geq \min(a,c) \frac{bd}{b+d} .$$

Taking $a = |\lambda \cap \lambda_i^\star| / |\lambda_i^\star|$, $b = |\lambda_i^\star|$, $c = \left|\lambda \cap \lambda^\star_{i+1}\right| / \left|\lambda^\star_{i+1}\right|$ and $d = \left|\lambda^\star_{i+1}\right|$, we get Eq. (3.25). □

Proof of Lemma 3.3

In fact, we prove a slightly stronger statement. We show that, for any $n \geq 2$, for any $D^\star \in \{2, \dots, n\}$, for any $D \in \{1, \dots, D^\star - 1\}$ and any $\tau \in \mathcal{T}_n^D$,

$$\|\mu^\star - \mu^\star_\tau\|^2 \geq \min_{1 \leq i \leq D^\star - 1} \left\{ \frac{|\lambda_i^\star| \cdot \left|\lambda^\star_{i+1}\right|}{|\lambda_i^\star| + \left|\lambda^\star_{i+1}\right|} \cdot \left\| \mu^\star_{\lambda^\star_{i+1}} - \mu^\star_{\lambda_i^\star} \right\|^2_{\mathcal{H}} \right\} . \qquad (3.27)$$

Then,

$$\|\mu^\star - \mu^\star_\tau\|^2 \geq \underline{\Gamma} \cdot \underline{\Delta}^2 \qquad \text{where} \qquad \underline{\Gamma} = \left(n \max_{1 \leq i \leq D^\star - 1} \left\{ \frac{1}{|\lambda_i^\star|} + \frac{1}{\left|\lambda^\star_{i+1}\right|} \right\} \right)^{-1} .$$

Since we always have

$$\underline{\Lambda}_{\tau^\star} \geq \underline{\Gamma} \geq \frac{1}{2} \underline{\Lambda}_{\tau^\star} ,$$

Eq. (3.12) follows.

Proof of Eq. (3.27) by induction. We show by strong induction on $D^\star$ that, for any $D^\star \geq 2$, for any $D \in \{1, \dots, D^\star - 1\}$, any $n \geq D^\star$ and any $\tau \in \mathcal{T}_n^D$, Eq. (3.27)

holds true.

First, if $D^\star = 2$, the result follows by Eq. (3.25) in Lemma 3.12 since we then have $i = 1$ and

$$\frac{|\lambda \cap \lambda_1^\star|}{|\lambda_1^\star|} = \frac{|\lambda \cap \lambda_2^\star|}{|\lambda_2^\star|} = 1\,.$$

Suppose now that the result is proved for all $D^\star \in \{2, \ldots, p\}$ and consider a change-point problem $(\tau^\star, \mu^\star)$ with $D^\star = D^\star = p+1$ and $n \geq p+1$. Let $D < p+1$ and some segmentation $\tau \in \mathcal{T}_n^D$ be fixed. Then one of these two scenarios occurs: (i) there exists $\lambda_i^\star$ with $2 \leq i \leq D^\star - 1$ that does not contain any change-point of τ, or (ii) $\lambda_2^\star$,...,$\lambda_{D^\star-1}^\star$ all contain a change-point of τ.

Case (i). Suppose that there exists an inner segment $\lambda_i^\star$ of $\tau^\star$, $2 \leq i \leq D^\star - 1$, that does not contain any change-point of τ (see Fig. 3-2). Therefore, there exists $k \in \{1, \ldots, D\}$ such that $\lambda_i^\star \subsetneqq \lambda_k$. By definition, there are $i-1$ change-points of $\tau^\star$ to the left of $\lambda_i^\star$ and $k-1$ change-points of τ to the left of $\lambda_i^\star$. Suppose that $k < i$. We define τ° as the segmentation obtained by adding $\tau_i^\star$ to τ (see Fig. 3-2). Then $\|\mu^\star - \mu_\tau^\star\|^2 \geq \|\mu^\star - \mu_{\tau^\circ}^\star\|^2$ because τ° is finer than τ. Reducing τ° to a segmentation $\widetilde{\tau}^\circ$ of $\{1, 2, \ldots, \tau_i^\star\}$ in k segments and $\tau^\star$ to a segmentation $\widetilde{\tau}^\star$ of $\{1, 2, \ldots, \tau_i^\star\}$ in i segments and defining $\widetilde{\mu}^\star = (\mu_1^\star, \ldots, \mu_{\tau_i^\star}^\star) \in \mathcal{H}^i$, we get back to a situation covered by the induction since $i \leq D^\star - 1$ and $k < i$. So,

$$\begin{aligned}
\|\widetilde{\mu}^\star - \widetilde{\mu}_{\widetilde{\tau}^\circ}^\star\|^2 &\geq \inf_{1 \leq j \leq i-1} \left\{ \frac{\left|\lambda_j^\star\right| \cdot \left|\lambda_{j+1}^\star\right|}{\left|\lambda_j^\star\right| + \left|\lambda_{j+1}^\star\right|} \cdot \left\| \widetilde{\mu}_{\lambda_{j+1}^\star}^\star - \widetilde{\mu}_{\lambda_j^\star}^\star \right\|_{\mathcal{H}}^2 \right\} \\
&\geq \inf_{1 \leq j \leq D^\star - 1} \left\{ \frac{\left|\lambda_j^\star\right| \cdot \left|\lambda_{j+1}^\star\right|}{\left|\lambda_j^\star\right| + \left|\lambda_{j+1}^\star\right|} \cdot \left\| \mu_{\lambda_{j+1}^\star}^\star - \mu_{\lambda_j^\star}^\star \right\|_{\mathcal{H}}^2 \right\}
\end{aligned}$$

and we get the result since $\|\mu^\star - \mu_{\tau^\circ}^\star\|^2 \geq \|\widetilde{\mu}^\star - \widetilde{\mu}_{\widetilde{\tau}^\circ}^\star\|^2$. A symmetric reasoning can be applied if $k \geq i$, considering change-points to the right of $\lambda_i^\star$ and using that $D - k + 1 < D^\star - i + 1$ since $D < D^\star$.

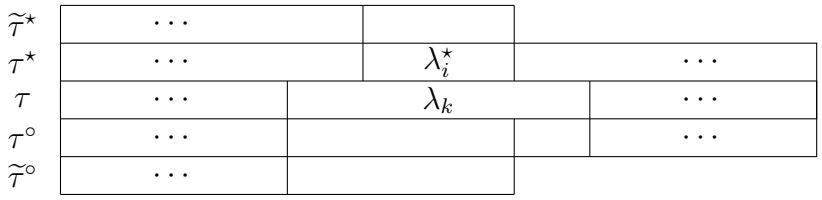

Figure 3-2 – Proof of Lemma 3.3, Case (i): $\lambda_i^\star$ is a segment of $\tau^\star$ that is included in a segment of τ. The segmentation τ° is obtained by joining $\tau_i^\star$ to the segmentation τ.

Case (ii). Suppose that each inner segment of $\tau^\star$ contains a change-point of τ. Since there are $D^\star - 2$ inner segments of $\tau^\star$ and $D - 1 \leq D^\star - 2$ change-points of τ, there is at most (hence exactly) one change-point of τ in each inner segment of $\tau^\star$. Then $D = D^\star - 1$ and we are in the situation depicted in Fig. 3-3.

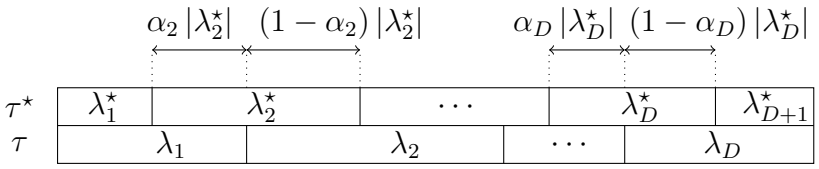

Figure 3-3 – Proof of Lemma 3.3, Case (ii): $D = D^\star - 1$ and each inner segment of $\tau^\star$ contains exactly one change-point of τ.

We can use Eq. (3.25) in Lemma 3.12 to lower bound the contribution of each $\lambda \in \tau$ to $\|\mu^\star - \mu_\tau^\star\|^2$. For $2 \leq i \leq D = D^\star - 1$, define $\alpha_i := |\lambda_i^\star \cap \lambda_{i-1}| \,/\, |\lambda_i^\star|$. Then, we have

$$\begin{aligned}
\|\mu^\star - \mu_\tau^\star\|^2 &\geq (1 \wedge \alpha_2)\frac{|\lambda_1^\star| \cdot |\lambda_2^\star|}{|\lambda_1^\star| + |\lambda_2^\star|} \cdot \left\|\mu_{\lambda_2^\star}^\star - \mu_{\lambda_1^\star}^\star\right\|_{\mathcal{H}}^2 \\
&\quad + \sum_{j=2}^{D-1} \left(\left[(1-\alpha_j) \wedge \alpha_{j+1}\right] \cdot \frac{\left|\lambda_j^\star\right| \cdot \left|\lambda_{j+1}^\star\right|}{\left|\lambda_j^\star\right| + \left|\lambda_{j+1}^\star\right|} \cdot \left\|\mu_{\lambda_{j+1}^\star}^\star - \mu_{\lambda_j^\star}^\star\right\|_{\mathcal{H}}^2 \right) \\
&\quad + \left[(1-\alpha_D) \wedge 1\right] \frac{\left|\lambda_D^\star\right| \cdot \left|\lambda_{D+1}^\star\right|}{\left|\lambda_D^\star\right| + \left|\lambda_{D+1}^\star\right|} \cdot \left\|\mu_{\lambda_{D+1}^\star}^\star - \mu_{\lambda_D^\star}^\star\right\|_{\mathcal{H}}^2 \\
&\geq \left[1 \wedge \alpha_2 + (1-\alpha_2) \wedge \alpha_3 + \cdots + (1-\alpha_{D-1}) \wedge \alpha_D + (1-\alpha_D) \wedge 1\right] \\
&\quad \times \inf_{1 \leq j \leq D^\star - 1} \left\{ \frac{\left|\lambda_j^\star\right| \cdot \left|\lambda_{j+1}^\star\right|}{\left|\lambda_j^\star\right| + \left|\lambda_{j+1}^\star\right|} \cdot \left\|\mu_{\lambda_{j+1}^\star}^\star - \mu_{\lambda_j^\star}^\star\right\|_{\mathcal{H}}^2 \right\} .
\end{aligned}$$

Since $\alpha_i \geq 0$ for any $2 \leq i \leq D^\star - 1$, it is straightforward to show that

$$\alpha_2 + (1-\alpha_2) \wedge \alpha_3 + \cdots + (1-\alpha_D) \geq 1 \,,$$

which concludes the proof. □

3.5.4 Proof of Lemma 3.4

Let us define $\delta := \min\left\{n\underline{\Lambda}_{\tau^\star}, \mathrm{d}_\infty^{(1)}(\tau^\star, \tau)\right\}$. If $\delta = 0$, then Eq. (3.13) holds true. We assume from now on that $\delta > 0$.

Because $n\underline{\Lambda}_{\tau^\star} \geq \delta$, for any $1 \leq i \leq D^\star - 1$, we can write $\left|\tau_{i+1}^\star - \tau_i^\star\right| \geq \delta$. On the other hand, because $\mathrm{d}_\infty^{(1)}(\tau^\star, \tau) \geq \delta$, there exists $i \in \{1, \ldots, D^\star - 1\}$ such that, for any $j \in \{1, \ldots, D-1\}$, $|\tau_i^\star - \tau_j| \geq \delta$. Since $\delta \leq n\underline{\Lambda}_{\tau^\star}$, this also holds true for $j = 0$ and $j = D$. Let us define, as illustrated by Fig. 3-4,

$$\lambda^\circ := \left\{\tau_i^\star - \delta + 1, \ldots, \tau_i^\star, \tau_i^\star + 1, \ldots, \tau_i^\star + \delta\right\} .$$

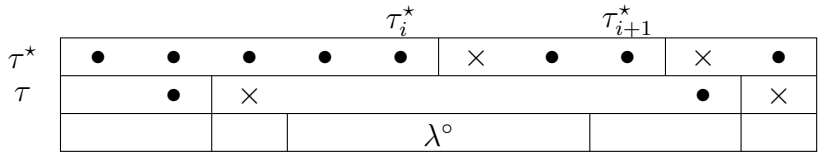

Figure 3-4 – Construction of λ° in the proof of Lemma 3.4. In this case, $\delta = 2$ since $\underline{\Lambda}_{\tau^\star} = 2/10$ (the rightmost segment of $\tau^\star$ is of size 2) and $\mathrm{d}_\infty^{(1)}(\tau^\star, \tau) = 3$ (achieved in $\tau_i^\star$).

Since λ° is included in a segment of τ,

$$\|\mu^\star - \mu_\tau^\star\|^2 \geq \sum_{j \in \lambda^\circ} \left\|\mu_j^\star - (\mu_\tau^\star)_j\right\|_{\mathcal{H}}^2 \geq \sum_{j \in \lambda^\circ} \left\|\mu_j^\star - \mu_{\lambda^\circ}^\star\right\|_{\mathcal{H}}^2 .$$

Because of the hypothesis we made, λ° only intersects $\lambda_i^\star$ and $\lambda_{i+1}^\star$ among the segments of $\tau^\star$, so Eq. (3.24) in Lemma 3.12 shows that

$$\begin{aligned}\sum_{j \in \lambda^\circ} \left\|\mu_j^\star - \mu_{\lambda^\circ}^\star\right\|_{\mathcal{H}}^2 &= \frac{\left|\lambda^\circ \cap \lambda_i^\star\right| \cdot \left|\lambda^\circ \cap \lambda_{i+1}^\star\right|}{\left|\lambda^\circ \cap \lambda_i^\star\right| + \left|\lambda^\circ \cap \lambda_{i+1}^\star\right|} \left\|\mu_{\lambda_{i+1}^\star}^\star - \mu_{\lambda_i^\star}^\star\right\|_{\mathcal{H}}^2 \\ &= \frac{\delta}{2} \left\|\mu_{\lambda_{i+1}^\star}^\star - \mu_{\lambda_i^\star}^\star\right\|_{\mathcal{H}}^2 \geq \frac{\delta}{2} \underline{\Delta}^2 ,\end{aligned}$$

hence the result. □

3.5.5 Proof of Lemma 3.7

In this proof, since τ is fixed, we denote by $\lambda_1, \ldots, \lambda_D$ the segments of τ, that is, $\lambda_i = \{\tau_{i-1} + 1, \ldots, \tau_i\}$.

First, notice that

$$L_\tau = \langle \mu^\star - \mu_\tau^\star, \varepsilon \rangle = \sum_{i=1}^{D^\star} \langle \mu_{\lambda_i^\star}^\star, \sum_{j \in \lambda_i^\star} \varepsilon_j \rangle_{\mathcal{H}} - \sum_{i=1}^{D_\tau} \langle \mu_{\lambda_i}^\star, \sum_{j \in \lambda_i} \varepsilon_j \rangle_{\mathcal{H}} . \tag{3.28}$$

Now, if $D_\tau < D^\star$ we *arbitrarily* define $\lambda_{D_\tau+1} = \cdots = \lambda_{D^\star} = \emptyset$, so that $\sum_{j \in \lambda_i} \varepsilon_j = 0$ for every $i \in \{D_\tau + 1, \ldots, D^\star\}$. Similarly, if $D^\star < D_\tau$, we define $\lambda_{D^\star+1}^\star = \cdots = \lambda_{D_\tau}^\star = \emptyset$. We also define $\mu_\emptyset^\star = \mu_n^\star$ by convention. Then, defining $D^+ := \max\{D^\star, D_\tau\}$, we can rewrite Eq. (3.28) as follows:

$$\begin{aligned}L_\tau &= \sum_{i=1}^{D^+} \langle \mu_{\lambda_i^\star}^\star, \sum_{j \in \lambda_i^\star} \varepsilon_j \rangle_{\mathcal{H}} - \sum_{i=1}^{D^+} \langle \mu_{\lambda_i}^\star, \sum_{j \in \lambda_i} \varepsilon_j \rangle_{\mathcal{H}} \\ &= \sum_{i=1}^{D^+} \langle \mu_{\lambda_i^\star}^\star - \mu_{\lambda_i}^\star, \sum_{j \in \lambda_i^\star} \varepsilon_j \rangle_{\mathcal{H}} + \sum_{i=1}^{D^+} \langle \mu_{\lambda_i}^\star, \sum_{j \in \lambda_i^\star} \varepsilon_j - \sum_{j \in \lambda_i} \varepsilon_j \rangle_{\mathcal{H}}\end{aligned}$$

$$= \sum_{i=1}^{D^+} \langle \mu^\star_{\lambda_i^\star} - \mu^\star_{\lambda_i}, \sum_{j\in\lambda_i^\star} \varepsilon_j \rangle_{\mathcal{H}} + \sum_{i=1}^{D^+} \langle \mu^\star_{\lambda_i} - \mu^\star_n, \sum_{j\in\lambda_i^\star} \varepsilon_j - \sum_{j\in\lambda_i} \varepsilon_j \rangle_{\mathcal{H}} \,,$$

since

$$\sum_{i=1}^{D^+} \Big(\sum_{j\in\lambda_i^\star} \varepsilon_j - \sum_{j\in\lambda_i} \varepsilon_j \Big) = 0 \,.$$

Then, by the triangle inequality and Cauchy-Schwarz inequality,

$$\begin{aligned}
|L_\tau| &\leq \sum_{i=1}^{D^+} \left\| \mu^\star_{\lambda_i^\star} - \mu^\star_{\lambda_i} \right\|_{\mathcal{H}} \left\| \sum_{j\in\lambda_i^\star} \varepsilon_j \right\|_{\mathcal{H}} + \sum_{i=1}^{D^+} \left\| \mu^\star_{\lambda_i} - \mu^\star_n \right\|_{\mathcal{H}} \left\| \sum_{j\in\lambda_i^\star} \varepsilon_j - \sum_{j\in\lambda_i} \varepsilon_j \right\|_{\mathcal{H}} \\
&\leq \operatorname{diam} \operatorname{conv}\big\{ \mu^\star_j \,/\, j \in \big\{1,\dots,n\big\} \big\} \\
&\qquad \times \left[\sum_{i=1}^{D^+} \left\| \sum_{j\in\lambda_i^\star} \varepsilon_j \right\|_{\mathcal{H}} + \sum_{i=1}^{D^+} \left(\left\| \sum_{j\in\lambda_i^\star} \varepsilon_j \right\|_{\mathcal{H}} + \left\| \sum_{j\in\lambda_i} \varepsilon_j \right\|_{\mathcal{H}} \right) \right] \\
&\leq 3D^+ \operatorname{diam} \operatorname{conv}\big\{ \mu^\star_j \,/\, j \in \big\{1,\dots,n\big\} \big\} \times \sup_{1\leq a<b\leq n} \left\| \sum_{j=a}^{b} \varepsilon_j \right\|_{\mathcal{H}}
\end{aligned}$$

where we used that $\mu^\star_\lambda \in \operatorname{conv}\big\{ \mu^\star_j \,/\, j \in \big\{1,\dots,n\big\} \big\}$ for any segment λ. Since the diameter of the convex hull of a finite set of points is equal to the diameter of the set, we have

$$\begin{aligned}
\operatorname{diam} \operatorname{conv}\big\{ \mu^\star_j \,/\, j \in \big\{1,\dots,n\big\} \big\} &= \operatorname{diam}\big\{ \mu^\star_j \,/\, j \in \big\{1,\dots,n\big\} \big\} \\
&\leq (D^\star - 1)\overline{\Delta} < D^\star \overline{\Delta} \,.
\end{aligned}$$

Using also Lemma 3.5, we get the result. □

3.5.6 Proof of Lemma 3.11

Let us put $\zeta := \|\varepsilon_1 + \cdots + \varepsilon_n\|^2_{\mathcal{H}}$. Since for any $j \neq k$, $\mathbb{E}\left[\langle \varepsilon_j, \varepsilon_k \rangle_{\mathcal{H}}\right] = 0$ (see Remark 3.5), by definition of v_j,

$$\mathbb{E}\left[\zeta\right] = \mathbb{E}\left[\|\varepsilon_1 + \cdots + \varepsilon_n\|^2_{\mathcal{H}}\right] = \sum_{j=1}^{n} v_j \,.$$

We recognize the right-hand side of (3.16) up to $1/x^2$. For any $r > 1$, let us denote by A_r the event

$$\forall 1 \leq s < r, \qquad \|\varepsilon_1 + \cdots + \varepsilon_s\|_{\mathcal{H}} < x \qquad \text{and} \qquad \|\varepsilon_1 + \cdots + \varepsilon_r\|_{\mathcal{H}} \geq x \,,$$

and by A_1 the event $\|\varepsilon_1\|_{\mathcal{H}} \geq x$. These events are disjoints, thus we can write

$$\mathbb{P}\left(\max_{1\leq k\leq n} \|\varepsilon_1 + \cdots + \varepsilon_k\|_{\mathcal{H}} \geq x\right) = \mathbb{P}\left(\bigcup_{r=1}^{n} A_r\right) = \sum_{r=1}^{n} \mathbb{P}\left(A_r\right) . \tag{3.29}$$

The law of total expectation and the positiveness of ζ yield

$$\mathbb{E}\left[\zeta\right] \geq \sum_{r=1}^{n} \mathbb{E}\left[\zeta \middle| A_r\right] \mathbb{P}\left(A_r\right) .$$

Finally, let $\ell \leq r < k$ be integers. Since ε_ℓ is independent from ε_k conditionally to $\sigma(\varepsilon_1, \ldots, \varepsilon_r)$, ε_ℓ is independent from ε_k conditionally to A_r. Furthermore, ε_k is independent from A_r and

$$\mathbb{E}\left[\langle \varepsilon_k, \varepsilon_\ell \rangle_{\mathcal{H}} \middle| A_r\right] = \langle \mathbb{E}\left[\varepsilon_k\right], \mathbb{E}\left[\varepsilon_\ell \middle| A_r\right] \rangle_{\mathcal{H}} = 0 .$$

Because of this relation and the positivity of the (real) conditional expectation, for any integers $r \leq k \leq j$,

$$\mathbb{E}\left[\zeta \middle| A_r\right] = \mathbb{E}\left[\|\varepsilon_1 + \cdots + \varepsilon_n\|_{\mathcal{H}}^2 \middle| A_r\right] \geq \mathbb{E}\left[\|\varepsilon_1 + \cdots + \varepsilon_r\|_{\mathcal{H}}^2 \middle| A_r\right] \geq x^2 .$$

Therefore, $\mathbb{E}\left[\zeta \middle| A_r\right] \geq x^2$, which gives $\mathbb{E}\left[\zeta\right] \geq x^2 \sum \mathbb{P}\left(A_r\right)$. This concludes the proof, thanks to Eq. (3.29). □

Remark 3.5. The independence between ε_j and ε_k for $j \neq k$ yields $\mathbb{E}\left[\langle \varepsilon_j, \varepsilon_k \rangle_{\mathcal{H}}\right] = 0$. Indeed, we dispose of a conditional expectation on $\mathcal{H}$ [Diestel and Uhl, 1977, chapter 5], which satisfies the same properties than the conditional expectation with real random variables. Hence we can write

$$\begin{aligned}\mathbb{E}\left[\langle \varepsilon_j, \varepsilon_k \rangle_{\mathcal{H}}\right] = \mathbb{E}\left[\mathbb{E}\left[\langle \varepsilon_j, \varepsilon_k \rangle_{\mathcal{H}} \middle| \varepsilon_k\right]\right] &= \mathbb{E}\left[\langle \mathbb{E}\left[\varepsilon_j \middle| \varepsilon_k\right], \varepsilon_k \rangle_{\mathcal{H}}\right] \\ &= \mathbb{E}\left[\langle \mathbb{E}\left[\varepsilon_j\right], \varepsilon_k \rangle_{\mathcal{H}}\right] = 0.\end{aligned}$$

Note that the ε_js expectation vanishes by hypothesis.

3.5.7 Technical lemmas for the proof of Theorem 3.2

In this we state and prove technical results used in the proof of Theorem 3.2 in Section 3.4.5.

Our first assert that the penalty shape $\mathrm{pen}_{\mathrm{L}}$ is increasing under some assumption on the penalty coefficients.

Lemma 3.13. *Take $n \in \mathbb{N}^\star$. Then, for any $a, b > 0$ such that $a < b$, the mapping*

$$f_{a,b} : x \mapsto x\left(a \log \frac{n}{x} + b\right)$$

takes positive values and is increasing on $[1, n]$.

Proof. Write $f'_{a,b}(x) = a\left(\log(n) - \log(x)\right) + b - a > 0$. □

The second result of this section is a simple inequality used in the proof of Theorem 3.2.

Lemma 3.14. *For any $x \geq 1$,*

$$x\log(x) - (x+1)\log(x+1) \geq -\log(x) - 2\,.$$

Proof. Let $x \geq 1$. Since $\log\left(1 + 1/x\right) \leq 1/x$, we have

$$(x+1)\log(x+1) - x\log(x) \leq \log(x) + 1 + 1/x\,,$$

and we can conclude. □

The next result collects some computations that intervene at the end of the proof of Theorem 3.2.

Lemma 3.15. *For any $1 \leq x \leq n$ and $y > 0$, it holds that*

$$\frac{\frac{74M^2}{3}\,(x+1)\left(\log(n) - \log\left(x+1\right)\right)}{2 + y + \log(n) - \log(x)} \leq \frac{74M^2}{3}\,(x+1)\;.$$

Proof. We write

$$\frac{\frac{74M^2}{3}\,(x+1)\left(\log(n) - \log\left(x+1\right)\right)}{2 + y + \log(n) - \log(x)} \leq \frac{\frac{74M^2}{3}\,(x+1)\left(\log(n) - \log(x)\right)}{2 + y + \log(n) - \log(x)}\,.$$

Since $\log(n) - \log(x) \geq 0$ and $y > 0$, we have

$$\frac{\log(n) - \log(x)}{2 + y + \log(n) - \log(x)} \leq 1\,,$$

and we can conclude. □

The following result is a computation regarding $c_{1,\max}$.

Lemma 3.16. *Under the Assumptions of Theorem 3.2,*

$$\frac{n\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2}{6D^\star} > c_1\left(y + \log\frac{n}{D^\star} + 4\right) + \frac{74}{3}M^2\left((y+4)\left(D^\star+2\right) + \log\frac{n}{D^\star}\right)\,.$$

Proof. We assumed that

$$c_1 < \frac{n\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2}{6D^\star(y + \log(n) + 4)} - \frac{74M^2}{3}\left(D^\star + 2\right)\,.$$

Simple algebra yields

$$\frac{n\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2}{6D^\star} > \left(c_1 + \frac{74M^2}{3}\left(D^\star+2\right)\right)\left(y + \log(n) + 4\right),$$

and since $D^\star + 2 > 0$,

$$\frac{n\underline{\Lambda}_{\tau^\star}\underline{\Delta}^2}{6D^\star} > c_1(y + \log(n) + 4) + \frac{74M^2}{3}\big((y+4)\,(D^\star+2) + \log(n)\big)\,.$$

We can conclude since $x \mapsto \log(x)$ is increasing and $c_1 > 0$. □

Chapter 4

Experimental results

Abstract

In this chapter, we show how the dimension jump heuristic can be a reasonable choice for the linear penalty in simulations. We also provide empirical evidence supporting the claims of Chapter 3. Finally, we demonstrate how to compute the key quantity $\underline{\Delta}$ that appears in our theoretical results, for translation-invariant kernels. Thanks to these computations, some of them novel, we are able to study precisely the link between the maximal penalty constant and $\underline{\Delta}$. We show that, as suggested by Theorem 3.1, it is proportional to $\underline{\Delta}^2$ all other things being equal. This chapter is partly based upon Garreau and Arlot [2016].

Whereas Chapter 3 was dedicated to theoretical results regarding KCP, we turn to more practical issues in this chapter. The pertinence of the dimension jump heuristic for KCP is studied in Section 4.1. Section 4.2 is devoted to the demonstration of the consistency of KCP on synthetic data. Finally, in Section 4.3.1, we focus on translation-invariant kernels and study the connection between the maximal penalty constant and $\underline{\Delta}^2$.

4.1 Choice of the penalty constant

From now on, we focus exclusively on KCP with a penalty proportional to the number of segments, that is

$$\mathrm{pen}\,(\tau) = \mathrm{pen}_\ell\,(\tau) = \frac{CD_\tau}{n} \quad \text{for some} \quad C > 0\,,$$

as defined in Eq. (2.3). A key practical question is the following: how do we choose a penalty constant C such that KCP recovers the correct number of change-points?

Let us denote by $\widehat{\tau}(C)$ the segmentation estimated by KCP for a penalty constant C, and $\widehat{D}(C)$ the number of segments of $\widehat{\tau}(C)$. We put $c_{\min} = c_{\min}(X_1, \ldots, X_n)$ and $c_{\max} = c_{\max}(X_1, \ldots, X_n)$ such that $\widehat{D}(c) = D^\star$ for any $c \in [c_{\min}, c_{\max}]$, a potentially empty interval. We call $c_{\min}$ (resp. $c_{\max}$) the minimal (resp. maximal) penalty constant. Assuming that $[c_{\min}, c_{\max}]$ is non-empty, we can reformulate the previous

question as: is it possible to choose $c \in [c_{\min}, c_{\max}]$ in a data-driven way?

Theorem 3.1 provides theoretical bounds for $c_{\min}$ and $c_{\max}$. Namely, with high probability,

$$c_{\min} \leq C_{\min} \approx D^\star \log(n) \qquad \text{and} \qquad c_{\max} \geq C_{\max} \approx \frac{\underline{\Delta}^2}{M^2} \frac{\underline{\Lambda}_{\tau^\star}}{D^\star} n \,, \tag{4.1}$$

However, as discussed in Section 3.2.1 after Theorem 3.1, $C_{\min}$ and $C_{\max}$ are not of much use in practice for choosing an adequate penalty constant c. Indeed, they both depend on the unknown quantities $D^\star$, $\underline{\Lambda}_{\tau^\star}$ and $\underline{\Delta}$. Furthermore, the numerical constants are such that $C_{\min} > C_{\max}$ more often than not for small values of n.

In Section 4.1.1, we present the dimension jump heuristic, an empirical method that aims at choosing $c \in [c_{\min}, c_{\max}]$, giving a partial answer to our opening question. We demonstrate its pertinence on synthetic data. In Section 4.1.3, we show that $c_{\min}$ and $c_{\max}$ have the same dependency on n than their theoretical counterparts $C_{\min}$ and $C_{\max}$.

4.1.1 The dimension jump heuristic

A well-understood phenomenon in penalized model selection is the existence of a *minimal penalty*, that is a penalty function $\text{pen}_{\min}$ such that if $\text{pen} = \alpha \, \text{pen}_{\min}$ with $\alpha < 1$ then the dimension of the estimated model tends to be close to the dimension of the largest models. This question has been first addressed by Birgé and Massart [2001, 2007] in the fixed-design Gaussian regression framework. In particular, it is shown that the optimal penalty is twice the minimal penalty, which gives rise to the following heuristic: (i) identify the minimal penalty $\text{pen}_{\min}$, (ii) choose as a penalty function $\text{pen} = 2 \, \text{pen}_{\min}$ [Birgé and Massart, 2007, Section 4].

Then the dimension jump heuristic for choosing C from the data is (i) compute $c_{\min}$ such that $\widehat{D}(c_{\min})$ is very large for $c < c_{\min}$ and reasonable for $c > c_{\min}$, (ii) define $\widehat{\tau} := \widehat{\tau}\,(2c_{\min})$.

Given the data of $\widehat{D}(c)$ as a function of c, it is usually not too hard to identify $c_{\min}$, that is when there is a clear jump in the values taken by $\widehat{D}(c)$. In this situation, it is customary to define

$$c_{\min} \in \operatorname*{arg\,max}_{c>0} \Big\{ \widehat{D}(c-) - \widehat{D}(c+) \Big\} \,, \tag{4.2}$$

that is the penalty constant achieving the maximal dimension jump. In case of equality, we take the largest constant, in order to select preferentially segmentations with fewer segments. Let us emphasize that a clear jump is not always present. In this case, one can for instance restrict $\big\{c > 0\big\}$ in Eq. (4.2) to $\big\{c > 0 \mid \widehat{D}(c) \leq D_{\max}\big\}$, where $D_{\max}$ is a user-defined constant. We provide examples of both situations in Fig. 4-1.

Fortunately, when running the KCP algorithm, it is not necessary to compute $\widehat{D}(c)$ for every value of $c > 0$ in order to find (4.2). The idea is to build recursively a sequence of critical constants c_i together with a sequence D_i for $i \geq 0$, such that

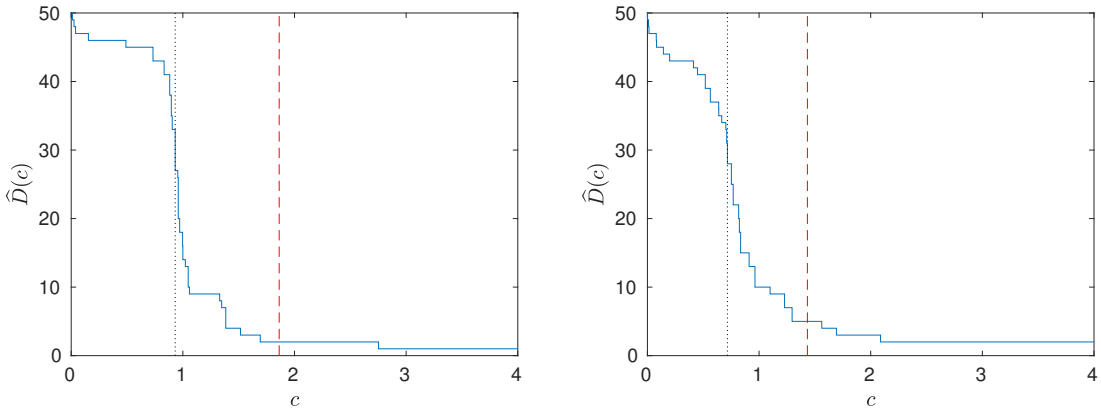

Figure 4-1 – Two plots of the estimated dimension as a function of the penalty constant c. In both experiments, we considered $n = 50$ data-points with a single jump in the mean of size 10, with unit variance Gaussian noise. *Left panel*: Gaussian kernel with bandwidth $\nu = 0.01$. *Right panel*: Gaussian kernel with bandwidth $\nu = 0.1$. The black dotted line denotes the maximal jump, located at $c = 0.93$ in the left panel (resp. $c = 0.72$ in the right panel); the red dashed line marks the penalty chosen by the dimension jump heuristic. We recover the true dimension $D^\star = 2$ in the left panel, but not in the right panel, where $\widehat{D} = 4$.

$\widehat{D}(c) = D_i$ for any $c \in [c_i, c_{i+1}]$. The first terms of these sequences are $c_0 = 0$ and $D_0 = n$, since $\widehat{D}(0) = n$. Let $i \geq 1$, and suppose that we successfully computed c_{i-1} and D_{i-1}. At the end of the dynamic programming step, we obtained a vector $R \in \mathbb{R}^n$ such that the D-th component of R is $\widehat{\tau}(D)$; it is the last column of the cost matrix `C` in the KCP algorithm, c.f. Section 2.4. Then, since $R(\cdot)$ is non-increasing and pen is increasing as a function of D_τ, for any $i \geq 1$, c_i is such that

$$R\left(D^i\right) + \frac{cD^i}{n} = R\left(D^{i-1}\right) + \frac{cD^{i-1}}{n}\,.$$

Algorithm 4.1 takes R as an argument and output sequences c_i and D_i. It can be seen as a simplified [1] version of Algorithm 3.2 in Arlot [2007] in the case of a linear penalty.

The overall complexity of Algorithm 4.1 is $\mathrm{O}\left(n^2\right)$ in the worst case, but usually a lot less steps are needed. Note also that it is possible to replace n by $D_{\max}$ in Algorithm 4.1, which decreases the complexity as well.

4.1.2 Empirical performance of the dimension jump heuristic

We now present some empirical results on synthetic data, suggesting that the dimension jump heuristic is generally a reasonable choice of penalty constant for the linear penalty. In all the tables, $x(\sigma)$ denotes a result x with standard deviation σ

1. Indeed, this procedure can be implemented for any non-decreasing penalty.

Algorithm 4.1 Computation of the critical constants

procedure ComputeJumps(R)
  $n \leftarrow$ length of R
  $D_0 \leftarrow n$
  $c_0 \leftarrow 0$
  for $i = 1 : n$ **do**
    if $D_{i-1} = 1$ **then**
      $c_i \leftarrow +\infty$
      $D_i \leftarrow 0$
      break
    else
      $[c_i, D_i] \leftarrow \min\left\{ \frac{R(D) - R(D_{i-1})}{(D_{i-1} - D)/n} \quad \text{s.t.} \quad D < D_{i-1} \right\}$
    end if
  end for
  return c, D
end procedure

on the experiments repetitions, $x^\star$ stands for a performance strictly better than 0.01, and **x** indicates the best performance on the line.

In the following experiments, we test KCP against a variety of generated data. These data are piecewise-constant in $\mathbb{R}$, with standard independent Gaussian noise. The number of jumps is fixed, say D. The jump locations are then chosen so that the segment sizes follow a multinomial distribution [2]with parameter $(100; 1/D, \dots, 1/D)$. Finally, the jump sizes are chosen uniformly at random in a pre-specified range. Experiments are repeated 10^2 times, and the randomness comes from the segments sizes, the breakpoints locations and the noise; standard deviation is computed over these repetitions.

We choose to focus on translation-invariant kernels, namely the Gaussian kernel k_{G} in Tables 4.1, 4.3, and 4.4, and the Laplace kernel k_{L} in Table 4.2. Keeping in mind the quadratic complexity of KCP, experiments over more than 10^3 data-points are quite expensive thus we limit ourself to $n \simeq 10^2$. We choose to present results only for $n = 200$, as we obtained similar results for values of n in the same range. The maximum authorized number of segments is set to $D_{\max} = 20$, which is much larger than the true number of segments in the simulations.

In our results, we choose to use $\mathrm{d}_{\mathrm{H}}^{(2)}$, whereas our theoretical results are all stated with $\mathrm{d}_{\infty}^{(1)}$ — see Chapter 3. This should not confuse the reader, recall that for segmentations that are close enough, these functions coincide, as we proved in Lemma 3.1. Note that $\mathrm{d}_{\mathrm{H}}^{(2)}$ is upper bounded by 0.5, and takes special values in the limit cases. For instance, if τ_0 is the n-segments segmentation, then $\mathrm{d}_{\mathrm{H}}^{(2)}(\tau, \tau_0) = \left\lfloor \overline{\Lambda}_\tau / 2 \right\rfloor$. Conversely, if τ^0 is the one-segment segmentation, $\mathrm{d}_{\mathrm{H}}^{(2)}(\tau, \tau_0) = 1/2$.

2. Let there be k distinct outcomes with probability p_i. Then, if N_i indicates the number of times outcome number i is observed over the n trials, we say that the vector $(N_1, \dots, N_k)$ follows a

Table 4.1 – Performances of KCP with Gaussian kernel on synthetic data with a single jump of size $\delta \in [1, 10]$, measured by $\mathrm{d}_{\mathrm{H}}^{(2)}$. The "DJH" column corresponds to a penalty constant chosen with the dimension jump heuristic.

$\nu \backslash c$	1.0	5.0	10	50	DJH
0.05	0.24 (0.02)	0.46 (0.07)	0.47 (0.02)	0.47 (0.02)	**0.02** (0.08)
0.1	0.24 (0.02)	0.04 (0.13)	0.47 (0.02)	0.47 (0.02)	**0.01*** (0.07)
0.5	0.24 (0.02)	**0.01*** (0.01)	0.02 (0.09)	0.47 (0.02)	0.03 (0.06)
1.0	0.23 (0.03)	**0.01*** (0.01)	0.01* (0.01)	0.14 (0.22)	0.05 (0.08)

Table 4.2 – Performances of KCP with Laplace kernel on synthetic data with a single jump of size $\delta \in [1, 10]$, measured by $\mathrm{d}_{\mathrm{H}}^{(2)}$. The "DJH" column corresponds to a penalty constant chosen with the dimension jump heuristic.

$\nu \backslash c$	1.0	5.0	10	50	DJH
0.05	0.24 (0.02)	0.08 (0.17)	0.47 (0.02)	0.47 (0.02)	**0.01** (0.05)
0.1	0.24 (0.02)	0.02 (0.09)	0.12 (0.21)	0.47 (0.02)	**0.01** (0.01)
0.5	0.24 (0.03)	**0.01*** (0.01)	0.01 (0.08)	0.47 (0.02)	0.04 (0.08)
1.0	0.14 (0.09)	**0.01*** (0.01)	0.02 (0.08)	0.17 (0.23)	0.03 (0.06)

Table 4.3 – Performances of KCP with Gaussian kernel on synthetic data with 5 jumps of sizes in the range $[1, 10]$, measured by $\mathrm{d}_{\mathrm{H}}^{(2)}$. The "DJH" column corresponds to a penalty constant chosen with the dimension jump heuristic.

$\nu \backslash c$	1.0	5.0	10	50	DJH
0.5	0.09 (0.01)	0.04 (0.07)	0.22 (0.09)	0.48 (0.02)	**0.02** (0.03)
1.0	0.08 (0.02)	0.02 (0.05)	0.08 (0.08)	0.47 (0.02)	**0.02** (0.03)
5.0	0.03 (0.06)	0.13 (0.09)	0.19 (0.08)	0.42 (0.10)	**0.03** (0.03)
10	0.09 (0.09)	0.23 (0.10)	0.30 (0.11)	0.45 (0.08)	**0.03** (0.03)

We can see in the results reported in Table 4.1 and 4.2 that the dimension jump heuristic provides a reasonable way to tune the penalty constant in the setting described at the beginning of this section. Indeed, the dimension jump heuristic generally manages to find a trade-off between (i) too small c, that yields an over-segmentation which translates into performances close to $\mathbb{E}\left\lfloor \overline{\Lambda}_\tau \right\rfloor \simeq 0.25$, and (ii) too large c, that yields an under-segmentation which translates into performances close to 0.5 in Table 4.1 and 4.2. Even though the dimension jump heuristic does not always finds a penalty constant in the correct region, we want to point out that the results obtained are nevertheless far better than those observed for *ad hoc* penalty constants chosen too small or too large.

We report in Table 4.3 further results that confirms the good behavior of the dimension jump heuristic in the same setting, this time for multiple change-points. We do not report results for the Laplace kernel, that are similar.

Finally, we also present some results for a change in the variance of a Gaussian sequence of observations. We consider as before $n = 200$ data-points, with a single random break-point drawn uniformly (though the minimal segment size is fixed to 5). The standard deviation of the observations before the change-point is set to 1.0, and 5.0 after, while the mean stays at zero. This data generation procedure is repeated 100 times for each choice of C and ν. The results of this experiment are reported in Table 4.4. As before, we observe that the dimension jump heuristic provides a reliable choice for C, if not the best. We do not report results for the Laplace kernel, which are similar.

Note that the sampling of the change-point locations is slightly different from before. It is more probable to observe short segments with uniform sampling rather than binomial sampling — a multinomial with parameters $(n; 1/2, 1/2)$ is a binomial with parameters $(n, 1/2)$ —, hence it is a slightly more difficult problem.

multinomial distribution with parameters $(n; p_1, \ldots, p_k)$.

Table 4.4 – Performances of KCP with Gaussian kernel on synthetic data with a single jump in the variance, measured by $\mathrm{d}_{\mathrm{H}}^{(2)}$. The "DJH" column corresponds to a penalty constant chosen with the dimension jump heuristic.

$\nu \backslash c$	1.0	5.0	10	50	DJH
0.5	0.36 (0.07)	0.22 (0.14)	0.22 (0.14)	0.22 (0.14)	**0.15** (0.13)
1.0	0.35 (0.08)	0.24 (0.15)	0.24 (0.15)	0.24 (0.15)	**0.06** (0.07)
5.0	0.35 (0.08)	**0.01** (0.03)	0.16 (0.14)	0.24 (0.14)	0.03 (0.07)
10	0.34 (0.09)	**0.01** (0.02)	0.03 (0.05)	0.24 (0.15)	0.03 (0.06)

4.1.3 Minimal and maximal penalty constant

In this section, we show how a modification of Algorithm 4.1 can give the values of $c_{\min}$ and $c_{\max}$, when they exist, and we compare these values to the theoretical bounds $C_{\min}$ and $C_{\max}$.

Computing $c_{\min}$ and $c_{\max}$. Suppose that we have access to the true dimension $D^\star$. Since Algorithm 4.1 provides all the critical constants, it is a straightforward modification to recover those corresponding to $D^\star$ when they exist. We call Algorithm 4.2 this modification. It outputs $c_{\min}$ and $c_{\max}$ such that $\widehat{D}(c) = D^\star$ for any $c \in [c_{\min}, c_{\max}]$. As a convention, if there is no penalty constant c such that $\widehat{D}(c) = D^\star$, the algorithm outputs the null value for both $c_{\min}$ and c_{max}. The computational complexity of this procedure is at most $\mathrm{O}\left(D^\star\right)$.

Experimental results. We present in Fig. 4-2 the results of the following experiment: for each n ranging from 5 to 300, we generate a piecewise-constant function with 5 segments of equal length and values on each segment alternating between 0 and 5. We then add noise sampled from a standard Gaussian distribution, and find $c_{\min}$ and $c_{\max}$ according to Algorithm 4.2 for Gaussian kernel KCP ($\nu = 1.0$). The preceding process is repeated 100 time for each value of n.

In the present setting, $D^\star$, $\underline{\Delta}$, M and $\underline{\Lambda}_{\tau^\star}$ are fixed. Therefore, Theorem 3.1 states that, with high probability, $C_{\min}$ is proportional to $\log(n)$ and $C_{\max}$ is proportional to n. We observe that the empirical evidence suggests that the dependency on n of $c_{\min}$ and $c_{\max}$ is the same than the dependency on n of $C_{\min}$ and $C_{\max}$: $c_{\min} \propto \log(n)$ and $c_{\max} \propto n$.

In the next section, we vary n as well in our experiments, but we interest ourselves in $\mathrm{d}_{\infty}^{(2)}\left(\widehat{\tau}_n, \tau^\star\right)$.

Algorithm 4.2 Min / Max penalty constant

procedure MinMaxPencsts(R,$D^\star$)
  $n \leftarrow$ length of R
  $D_0 \leftarrow n$
  $c_0 \leftarrow 0$
  for $i = 1 : n$ **do**
    if $D_{i-1} \geq D^\star$ **then**
      break
    else
      $[c_i, D_i] \leftarrow \min\left\{ \dfrac{R(D) - R(D_{i-1})}{(D_{i-1} - D)/n} \quad \text{s.t.} \quad D < D_{i-1} \right\}$
    end if
  end for
  $i_0 \leftarrow \inf_{i \geq 0} \{ D_i = D^\star \}$
  if $i_0 > 0$ **then**
    return $[c_{\min}, c_{\max}] = [c_{i_0}, c_{i_0+1}]$
  else
    return $[c_{\min}, c_{\max}] = [0, 0]$
  end if
end procedure

4.2 Consistency

A consequence of our main result, Theorem 3.1, is that for a bounded kernel, the KCP is consistent in the asymptotic setting presented in Example 2.1. In this section, we illustrate this by a simulation study.

Detecting changes in the mean with the Gaussian kernel. Let us consider the archetype change-point detection problem —finding changes in the mean of a sequence of independent random variables— and show how these changes are localized more precisely when more data are available.

We define three functions $\mu^m : [0,1] \to \mathbb{R}$, $1 \leq m \leq 3$, previously used by Arlot and Celisse [2011], which cover a variety of situations (see Fig. 4-3). For each $m \in \{1,2,3\}$ and several values of n between 10^2 and 10^3, we repeat 10^3 times the following:

— Sample n independent Gaussian random variables $g_i \sim \mathcal{N}(0,1)$;

— Set $X_i = \mu^m(i/n) + g_i$ —Fig. 4-3 shows one sample for each $m \in \{1,2,3\}$;

— Perform KCP with Gaussian kernel and linear penalty on $X_1, \ldots, X_n$; the penalty constant is chosen as indicated in Section 3.3, the bandwidth is set to 0.1, and the maximum number of change-points is set to 30;

— Compute $\mathrm{d_H}^{(2)}(\tau^\star, \widehat{\tau}_n)$.

The results are collected in Fig. 4-4, where each graph corresponds to a regression function μ^m. We represent in logarithmic scale the mean distance between the true

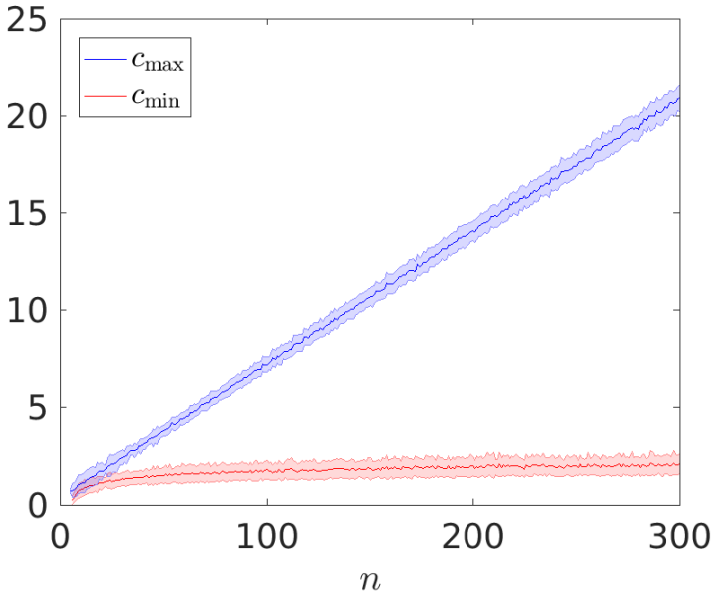

Figure 4-2 – Minimum and maximum penalty constants leading to exact recovery of the true dimension as a function of the sample size. The solid lines inside the shaded areas are the empirical mean over 100 repetitions of the experiments; the error bars correspond to the standard deviation. Null values were removed (3.5% of the experiments).

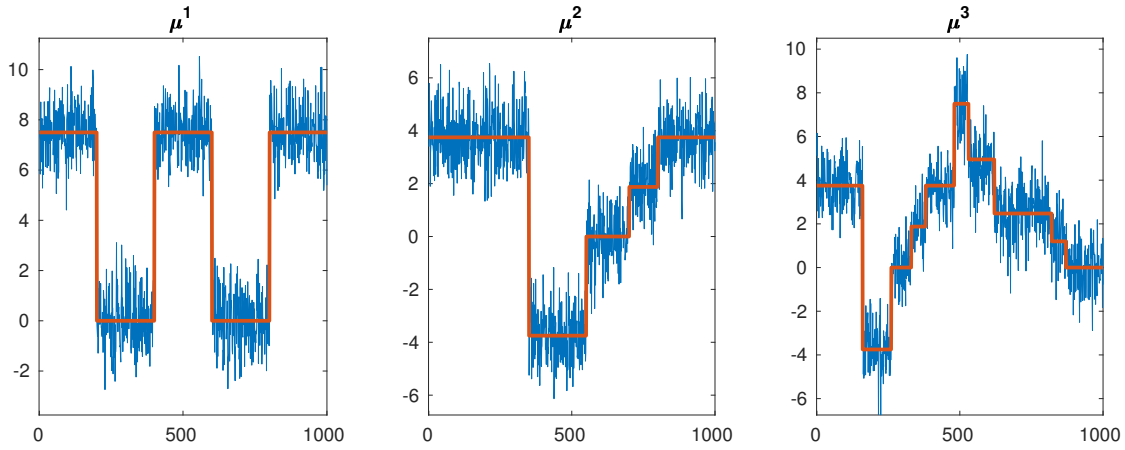

Figure 4-3 – In thick red lines, the three piecewise constant functions used in the simulations of Section 4.2. In lighter blue, a noisy version of these functions. Both μ^1 and μ^2 have 5 segments; μ^3 has 10 jumps.

segmentation and the estimated segmentation for each value of n. The error bars are $\pm\widehat{\sigma}/\sqrt{N}$, where $\widehat{\sigma}$ is the empirical standard deviation over $N = 10^3$ repetitions. We want to emphasize that, though these experiments illustrate our main result Theorem 3.1, they are carried out in a slightly different setting since the penalty

constant C is not chosen according to (3.3), but using the dimension jump heuristic as explained in Section 4.1.

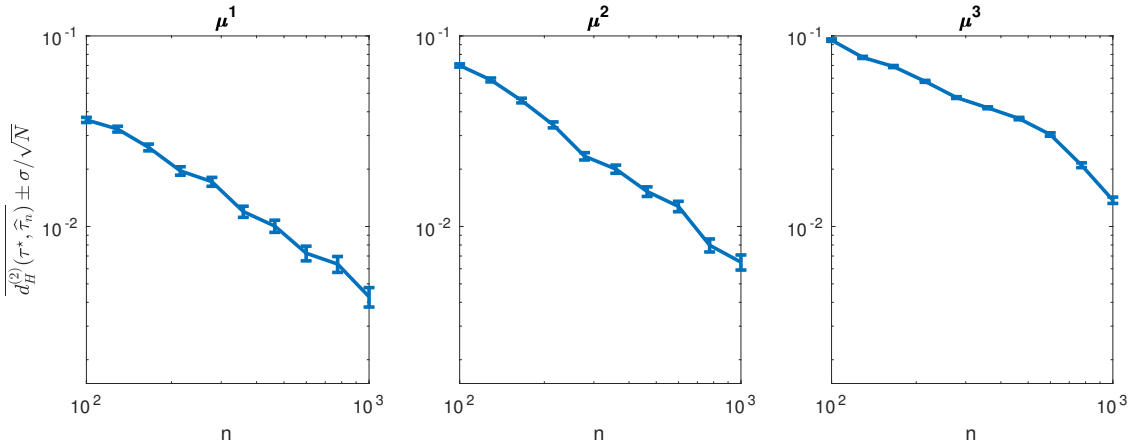

Figure 4-4 – Convergence of $\frac{1}{n}\,\mathrm{d}_{\mathrm{H}}^{(2)}\left(\tau^\star, \widehat{\tau}_n\right)$ towards 0 when the number of data points n is increasing. A linear regression between $\log(n)$ and $\frac{1}{n}\,\mathrm{d}_{\mathrm{H}}^{(2)}\left(\tau^\star, \widehat{\tau}_n\right)$ for $n \geq 300$ yields slope estimates -0.97, -1.04 and -1.00, respectively.

The three segmentation problems considered here are quite different in nature, but all lead to a linear convergence rate (slopes close to -1 on the graphs of Fig. 4-4) with different constants (different values for the intercept on the graphs of Fig. 4-4). Recall that Theorem 3.1 combined with Lemma 3.1 states that, with high probability,

$$\frac{1}{n}\,\mathrm{d}_{\mathrm{H}}^{(2)}\left(\tau^\star, \widehat{\tau}_n\right) \lesssim \widetilde{v}_1 = \frac{D^\star M^2}{\underline{\Delta}^2} \cdot \frac{\log(n)}{n}\,.$$

Hence, whenever $D^\star$, $\underline{\Delta}$ and M are fixed, $\frac{1}{n}\,\mathrm{d}_{\mathrm{H}}^{(2)}\left(\tau^\star, \widehat{\tau}_n\right)$ converges to 0 at rate at least $\log(n)/n$ when the number of data points increases. In our experimental setting, these quantities are fixed, and the observed convergence rate matches our theoretical upper bound. The performance of KCP still depends on the regression function μ^m experimentally, by a constant multiplicative factor, like the theoretical bound $\widetilde{v}_1$.

We want to emphasize that other choices of ν can lead to another ordering of the speeds of convergence observed in Fig. 4-4. To put it plainly, for another bandwidth, *e.g.*, $\nu = 1.0$, $\frac{1}{n}\,\mathrm{d}_{\infty}^{(1)}\left(\widehat{\tau}_n, \tau^\star\right)$ can converge more quickly to zero for μ^3 than for μ^1.

Detecting changes in the number of modes. Let us now consider observations $X_1, \ldots, X_n \in \mathbb{R}$ whose distribution vary only through the number of modes. Can we accurately detect such changes with the KCP procedure? The data are generated according to the following process for several n:

— Set $\tau_1^\star = \lfloor n/3 \rfloor$ and $\tau_2^\star = \lfloor 2n/3 \rfloor$;

— Draw $X_1, \ldots, X_{\tau_1^\star}, X_{\tau_2^\star+1}, \ldots, X_n$ according to a standard Gaussian distribution, and $X_{\tau_1^\star+1}, \ldots, X_{\tau_2^\star}$ according to a $(1/2, 1/2)$-mixture of Gaussian distributions $\mathcal{N}\left(\delta, 1-\delta^2\right)$ and $\mathcal{N}\left(-\delta, 1-\delta^2\right)$, with $\delta = 0.999$; the X_i are independent.

We test KCP with various kernels assuming that the number of change-points ($D^\star = 3$) is known; this simplification avoids possible artifacts linked to the choice of the penalty constant. Results are shown on Fig. 4-5. The X_i all have zero mean and unit variance, hence a classical penalized least-squares procedure —KCP with the linear kernel— is expected to detect poorly the changes in the distribution of the X_i, as confirmed by Fig. 4-5 (for instance, according to the right panel, it is not consistent). On the contrary, a Gaussian kernel with well-chosen bandwidth yields much better performance according to the middle and right panels of Fig. 4-5 (with a rate of order $1/n$).

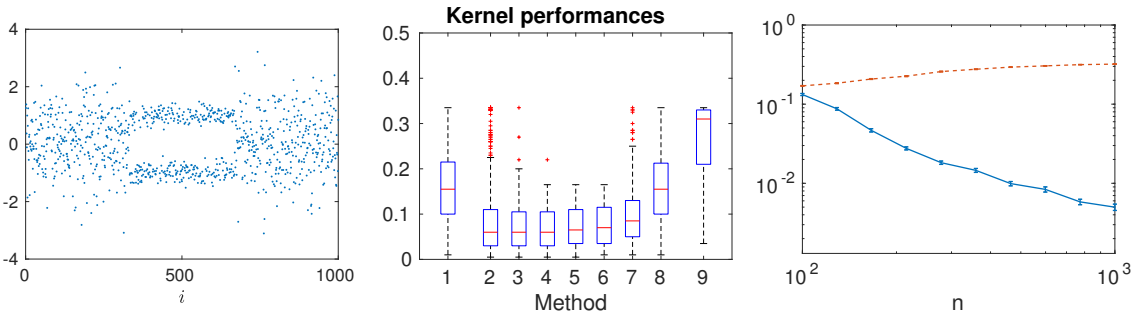

Figure 4-5 – **Left:** One sample $X_1, \ldots, X_n$ for $n = 10^3$. **Middle:** Performance of KCP with various kernels ($n = 200$). Methods 1 to 8: Gaussian kernel with bandwidth set *via* the median heuristic (method 1), or fixed equal to $0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1$ (methods $2, \ldots, 9$, respectively). Method 9: linear kernel. **Right:** Estimated values of $n^{-1}\, \mathrm{d}_{\mathrm{H}}^{(2)}\left(\tau^\star, \widehat{\tau}_n\right)$ vs. n in log scale, for KCP with a Gaussian kernel with bandwidth 0.01 (blue solid line; estimated slope -1.05) and with the linear kernel (red dashed line; estimated slope 0.16).

4.3 Translation-invariant kernels

Experiments for finite sample size in Arlot et al. [2012, Section 6] demonstrate how KCP can detect changes in the distribution of the data, even though the mean and variance are fixed. In this section, on the contrary, we focus on simple examples where the mean or variance of a sequence of Gaussian random variables vary, and we bring out the role of the kernel in KCP's ability to accurately detect changes.

Recall that Theorem 3.1 suggests that $\underline{\Delta}^2$ is a quantity of interest regarding the performances of KCP. In particular, Theorem 3.1 states that, with high probability,

$$c_{\max} \geq C_{\max} \approx \frac{\underline{\Delta}^2}{M^2}\,,$$

up to factors depending on $\tau^\star$, $\mu^\star$ and n. We restrict our study to translation-invariant kernels, since it is possible to compute $\underline{\Delta}^2$ in closed form for some kernels belonging to this class when the noise is Gaussian. In this setting, we show that $c_{\max} \propto \underline{\Delta}^2/M^2$ holds experimentally.

We first introduce translation-invariant kernels in Section 4.3.1, and show how to compute $\underline{\Delta}^2$ in this setting. In Section 4.3.2, we go through the details of theses computations when $k = k_{\mathrm{G}}$ and $k = k_{\mathrm{L}}$, with Gaussian observations. Finally, we use the results of these calculations in Section 4.3.3, where we demonstrate experimentally that $c_{\max} \propto \underline{\Delta}^2 / M^2$.

4.3.1 Introduction

We begin with a definition.

Definition 4.1. A translation-invariant kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a positive definite kernel such that there exists a function $\kappa : \mathbb{R}^d \to \mathbb{R}$ that satisfies

$$k(x, y) = \kappa(x - y).$$

In other words, $k(x, y)$ only depends on the *difference* between the vectors x and y. It is an attractive property, for instance when one wishes to detect changes in the mean of a signal: *a priori*, detecting a change between 0 and 5 is as important as detecting a change between 5 and 10.

If furthermore the function κ is continuous, then κ is called a *positive-definite function* [Berg et al., 1984]. We assume that k denotes a translation-invariant kernel such that $M = 1$ from now on. For instance, the Gaussian kernel and the Laplace kernel are both translation-invariant.

Bochner [1932, 1933] showed there is a deep correspondence between positive semi-definite functions and nonnegative measures on $\mathbb{R}$. More precisely, let us define the Fourier transform of any integrable function f as $\mathcal{F} f(\omega) := \int \mathrm{e}^{-\mathrm{i}x\omega} f(x) \,\mathrm{d}\, x$. Then

Theorem 4.1 (Bochner)**.** *A continuous function $\kappa : \mathbb{R}^d \to \mathbb{C}$ is positive semi-definite if, and only if, it is the Fourier transform of a finite nonnegative Borel measure ν on $\mathbb{R}^d$, that is*

$$\forall x \in \mathbb{R}^d, \qquad \kappa(x) = \mathcal{F}\nu(x) = \int \mathrm{e}^{-\mathrm{i}x^\top \omega} \,\mathrm{d}\,\nu(\omega).$$

It is possible to be more precise than Theorem 4.1 if we assume that κ is real-valued and *integrable*. More precisely, we have the following description of the RKHS associated to k.

Theorem 4.2 (Wendland [2005], Theorem 10.12)**.** *Let k be a translation-invariant kernel such that κ is integrable on $\mathbb{R}^d$ as well as its Fourier transform $\mathcal{F}\kappa$. The subset $\mathcal{H}$ of $L_2\left(\mathbb{R}^d\right)$ that consists of integrable and continuous functions f such that*

$$\|f\|_{\mathcal{H}}^2 := \frac{1}{(2\pi)^d} \int \frac{|\mathcal{F} f(\omega)|^2}{\mathcal{F}\kappa(\omega)} \,\mathrm{d}\,\omega < +\infty, \tag{4.3}$$

endowed with the inner product

$$\langle f, g \rangle_{\mathcal{H}} := \frac{1}{(2\pi)^d} \int \frac{\mathcal{F} f(\omega) \overline{\mathcal{F} g(\omega)}}{\mathcal{F}\kappa(\omega)} \,\mathrm{d}\,\omega,$$

is the RKHS associated to k.

Note that both the Gaussian kernel and the Laplace kernel satisfy the assumptions of Theorem 4.2.

We now explain how Theorem 4.2 allows to compute $\underline{\Delta}^2$ in certain cases. Let us restrict our study to real observations with changes in the mean and centered i.i.d. noise with a density on $\mathbb{R}$, that is

Assumption 4.1. There exist real numbers $L_1, \dots, L_n$ and i.i.d. centered real random variables β_i with densities ρ_i such that

$$\forall i \in \{1, \dots, n\}, \qquad X_i = L_i + \beta_i \,.$$

This setting corresponds to the experiments of Section 4.1 and 4.2. Recall that $\underline{\Delta}^2 = \min_{\mu_i^\star \neq \mu_{i+1}^\star} \left\|\mu_{i+1}^\star - \mu_i^\star\right\|_{\mathcal{H}}$, thus we have to compute $\left\|\mu_{i+1}^\star - \mu_i^\star\right\|_{\mathcal{H}}$ for every i such that $\mu_i^\star \neq \mu_{i+1}^\star$. It is the content of the next proposition, which is equivalent to Lemma 13 in Sriperumbudur et al. [2008].

Proposition 4.1. *Assume that k is a translation-invariant kernel such that κ and $\mathcal{F}\kappa$ are integrable. Suppose that Assumption 4.1 holds true. Let i, j be distinct elements of $\{1, \dots, n\}$. Then*

$$\|\mu_i^\star - \mu_j^\star\|_{\mathcal{H}}^2 = \frac{1}{2\pi} \int |\, \mathrm{e}^{-\mathrm{i}L_i\omega}\, \mathcal{F}\rho_i(\omega) - \mathrm{e}^{-\mathrm{i}L_j\omega}\, \mathcal{F}\rho_j(\omega)|^2\, \mathcal{F}\kappa(\omega)\, \mathrm{d}\,\omega \,.$$

Proof. Recall that the mean elements $\mu_i^\star$ satisfy

$$\forall g \in \mathcal{H}, \qquad \langle \mu_j^\star, g\rangle_{\mathcal{H}} = \mathbb{E}\left[g(X_j)\right] = \mathbb{E}\left[\langle Y_j, g\rangle_{\mathcal{H}}\right] \,.$$

Let y be a real number. Take $g = k\,(\cdot, y)$ in the previous display, we obtain

$$\mu_i^\star(y) = \mathbb{E}\left[k\,(X_i, y)\right] = \int \kappa(L_i + x - y)\rho_i(x)\, \mathrm{d}\,x \,.$$

Since κ is symmetric,

$$\mu_i^\star(y) = \int \kappa(y - L_i - x)\rho_i(x)\, \mathrm{d}\,x = \left((\kappa \star \rho_i) \circ \tau_{L_i}\right)(y) \,,$$

where $\star$ denotes the convolution operation and for any $a \in \mathbb{R}$, $\tau_a : x \mapsto x - a$. Elementary properties of the Fourier transform yields

$$\mathcal{F}\,\mu_i^\star(\omega) = \mathrm{e}^{-\mathrm{i}L_i\omega}\, \mathcal{F}\kappa(\omega) \cdot \mathcal{F}\,\rho_i(\omega) \,.$$

We conclude the proof by a straightforward application of Theorem 4.2. □

4.3.2 Computations

We now give a few concrete examples of $\underline{\Delta}^2$ computations under a Gaussian noise hypothesis. In the following, we give results for a single change-point, with the

convention that L_1 and ρ_1 are the mean and distribution before the jump, whereas L_2 and ρ_2 denote the mean and distribution after the jump. Note that stating these results can be straightforwardly adapted to an arbitrary number of change-points.

We first consider the Gaussian kernel, with a change in the mean.

Proposition 4.2 (Gaussian kernel, change in the mean)**.** *Set $k = k_G$. Suppose that Assumption 4.1 holds true with $L_2 - L_1 = \delta$ and $\beta_1, \beta_2 \sim \mathcal{N}(0, \sigma^2)$, where σ is a positive number. Then*

$$\underline{\Delta}^2 = \frac{2\nu}{\sqrt{\nu^2 + 2\sigma^2}} \left(1 - \exp\left(\frac{-\delta^2}{2\left(\nu^2 + 2\sigma^2\right)}\right)\right) .$$

Proof. The Gaussian kernel with bandwidth $\nu > 0$ is a translation-invariant kernel with

$$\kappa(z) = \exp\left(\frac{-z^2}{2\nu^2}\right) \qquad \text{and} \qquad \mathcal{F}\,\kappa(\omega) = \nu\sqrt{2\pi}\exp\left(\frac{-\omega^2\nu^2}{2}\right) .$$

Moreover, $\mathcal{F}\,\beta_1(\omega) = \mathcal{F}\,\beta_2(\omega) = \exp\left(-\sigma^2\omega^2/2\right)$. Prop. 4.1 then yields

$$\underline{\Delta}^2 = \frac{4\nu}{\sqrt{2\pi}} \int \sin^2\frac{\delta\omega}{2} \exp\left(-\left(\sigma^2 + \frac{\nu^2}{2}\right)\omega^2\right) \mathrm{d}\,\omega\,,$$

and a computation concludes the proof. □

It is possible to prove an analogous result for a change in the variance.

Proposition 4.3 (Gaussian kernel, change in the variance)**.** *Set $k = k_G$. Suppose that Assumption 4.1 holds true with $L_1 = L_2$, $\beta_1 \sim \mathcal{N}(0, \sigma_1^2)$, and $\beta_2 \sim \mathcal{N}(0, \sigma_2^2)$, where σ_1 and σ_2 are positive numbers. Then*

$$\underline{\Delta}^2 = \nu\left(\frac{1}{\sqrt{2\sigma_1^2 + \nu^2}} + \frac{1}{\sqrt{2\sigma_2^2 + \nu^2}} - \frac{2}{\sqrt{\sigma_1^2 + \sigma_2^2 + \nu^2}}\right) .$$

Proof. As in the proof of Prop. 4.2, $\mathcal{F}\,\kappa(\omega) = \nu\sqrt{2\pi}\exp\left(-\omega^2\nu^2/2\right)$, and $\mathcal{F}\,\beta_i(\omega) = \exp\left(-\sigma_i^2\omega^2/2\right)$ for $i \in \{1, 2\}$. Prop. 4.1 yields

$$\underline{\Delta}^2 = \frac{\nu}{\sqrt{2\pi}} \int \left(\mathrm{e}^{-\frac{\sigma_1^2\omega^2}{2}} - \mathrm{e}^{-\frac{\sigma_1^2\omega^2}{2}}\right)^2 \mathrm{e}^{\frac{-\omega^2\nu^2}{2}} \mathrm{d}\,\omega\,,$$

and a computation concludes the proof. □

We now turn to the Laplace kernel. A technical lemma is needed.

Lemma 4.1. *Let a, b and λ be positive numbers. Define*

$$f(\lambda) := \int \frac{\mathrm{e}^{-a\omega^2}\cos(\lambda\omega)}{1 + b\omega^2}\,\mathrm{d}\,\omega\,.$$

Then

$$f(\lambda) = \frac{\pi\,\mathrm{e}^{a/b}}{2\sqrt{b}}\left[\mathrm{e}^{\lambda/\sqrt{b}}\left(1-\operatorname{erf}\left(\frac{\lambda}{2\sqrt{a}}+\sqrt{\frac{a}{b}}\right)\right)+\mathrm{e}^{-\lambda/\sqrt{b}}\left(1+\operatorname{erf}\left(\frac{\lambda}{2\sqrt{a}}-\sqrt{\frac{a}{b}}\right)\right)\right].$$

In particular,

$$\int \frac{\mathrm{e}^{-a\omega^2}}{1+b\omega^2}\,\mathrm{d}\,\omega = \frac{\pi}{\sqrt{b}}\,\mathrm{e}^{\frac{a}{b}}\operatorname{erfc}\left(\sqrt{\frac{a}{b}}\right).$$

The proof of Lemma 4.1 is postponed to the end of this section. We are now able to prove analogous statements to Prop. 4.2 and 4.3.

Proposition 4.4 (Laplace kernel, change in the mean)**.** *Set $k = k_L$. Suppose that Assumption 4.1 holds true with $L_2 - L_1 = \delta$, and $\beta_1, \beta_2 \sim \mathcal{N}(0, \sigma^2)$, where σ is a positive number. Then*

$$\underline{\Delta}^2 = \mathrm{e}^{\frac{\sigma^2}{4\nu^2}}\left(2\operatorname{erfc}\left(\tfrac{\sigma}{2\nu}\right)-\mathrm{e}^{\delta/2\nu}\operatorname{erfc}\left(\tfrac{\delta}{2\sigma}+\tfrac{\sigma}{2\nu}\right)-\mathrm{e}^{-\delta/2\nu}\left(1+\operatorname{erf}\left(\tfrac{\delta}{2\sigma}-\tfrac{\sigma}{2\nu}\right)\right)\right).$$

Proof. The Laplace kernel with bandwidth $\nu > 0$ is a translation-invariant kernel with

$$\kappa(z) = \exp\left(\frac{-|z|}{2\nu}\right) \qquad \text{and} \qquad \mathcal{F}\,\kappa(\omega) = \frac{4\nu}{1+4\nu^2\omega^2}.$$

Moreover, $\mathcal{F}\,\beta_1(\omega) = \mathcal{F}\,\beta_2(\omega) = \exp(-\sigma^2\omega^2/2)$. Prop. 4.1 then yields

$$\underline{\Delta}^2 = \frac{4\nu}{\pi}\int \frac{(1-\cos(\delta\omega))}{1+4\nu^2\omega^2}\,\mathrm{e}^{-\sigma^2\omega^2}\,\mathrm{d}\,\omega,$$

and we conclude the proof with Lemma 4.1. □

Proposition 4.5 (Laplace kernel, change in the variance)**.** *Set $k = k_L$. Suppose that Assumption 4.1 holds true with $L_1 = L_2$, $\beta_1 \sim \mathcal{N}(0, \sigma_1^2)$, and $\beta_2 \sim \mathcal{N}(0, \sigma_2^2)$, where σ_1 and σ_2 are positive numbers. Then*

$$\begin{aligned}\underline{\Delta}^2 = {} & \mathrm{e}^{\frac{\sigma_1^2}{4\nu^2}}\operatorname{erf}(\nu)\left(\frac{1}{\sqrt{2\sigma_1^2+\nu^2}}+\frac{1}{\sqrt{2\sigma_2^2+\nu^2}}-\frac{2}{\sqrt{\sigma_1^2+\sigma_2^2+\nu^2}}\right)c\frac{\sigma_1}{2\nu} \\ & + \mathrm{e}^{\frac{\sigma_2^2}{4\nu^2}}\operatorname{erfc}\left(\frac{\sigma_2}{2\nu}\right)-2\,\mathrm{e}^{\frac{\sigma_1^2+\sigma_2^2}{8\nu^2}}\operatorname{erfc}\left(\sqrt{\frac{\sigma_1^2+\sigma_2^2}{8\nu^2}}\right). \qquad (4.4)\end{aligned}$$

Proof. As in the proof of Prop. 4.4, $\mathcal{F}\,\kappa(\omega) = \nu\sqrt{2\pi}\exp(-\omega^2\nu^2/2)$, and $\mathcal{F}\,\beta_i(\omega) = \exp(-\sigma_i^2\omega^2/2)$ for $i \in \{1, 2\}$. Prop. 4.1 yields

$$\underline{\Delta}^2 = \frac{2\nu}{\pi}\int \frac{\mathrm{e}^{-\sigma_1^2\omega^2}+\mathrm{e}^{-\sigma_2^2\omega^2}-2\,\mathrm{e}^{-\frac{\sigma_1^2+\sigma_2^2}{2}\omega^2}}{1+4\nu^2\omega^2}\,\mathrm{d}\,\omega,$$

we can conclude with Lemma 4.1. □

Remark 4.1. We have noted before that, in the single change-point setting, $\underline{\Delta}$ coincides with the MMD. As such, computations similar to Prop. 4.2, 4.3, 4.4 and 4.5 exist in the kernel two-sample test literature. In particular, Prop. 4.2 is consequence of Prop. 1 in Reddi et al. [2015], and the content of Prop. 4.3 appears in the proof of Prop. 3 in Reddi et al. [2015]. Prop. 4.4 and 4.5, however, are new up to the best of our knowledge.

Proof of Lemma 4.1. We note that f is well defined and twice differentiable. It is clear that

$$f''(\lambda) = -\int \frac{\omega^2 \, \mathrm{e}^{-a\omega^2} \cos(\lambda\omega)}{1 + b\omega^2} \, \mathrm{d}\,\omega \,,$$

hence f satisfies a second order linear differential equation, namely

$$f''(\lambda) - \frac{1}{b} f(\lambda) = \frac{-1}{b} \sqrt{\frac{\pi}{a}} \, \mathrm{e}^{-\lambda^2/(4a)} \,.$$

The solutions of this equation can be found with the variation of constants method, and we have, for c_1 and c_2 unknown constants,

$$\begin{aligned} f(\lambda) =& c_1 \, \mathrm{e}^{-\lambda/\sqrt{b}} + c_2 \, \mathrm{e}^{\lambda/\sqrt{b}} \\ & - \frac{\pi \, \mathrm{e}^{a/b}}{2\sqrt{b}} \left[\mathrm{e}^{\lambda/\sqrt{b}} \operatorname{erf}\left(\frac{\lambda}{2\sqrt{a}} + \sqrt{\frac{a}{b}} \right) - \mathrm{e}^{-\lambda/\sqrt{b}} \operatorname{erf}\left(\frac{\lambda}{2\sqrt{a}} - \sqrt{\frac{a}{b}} \right) \right] . \end{aligned} \tag{4.5}$$

Since

$$f(0) = \frac{\pi \, \mathrm{e}^{a/b}}{\sqrt{b}} \left(1 - \operatorname{erf}\left(\sqrt{\frac{a}{b}} \right) \right) \,,$$

we get $c_1 + c_2 = \pi \, \mathrm{e}^{a/b} / \sqrt{b}$. On the other side, we know that f is bounded for $\lambda \to \infty$ thus $c_2 = \pi \, \mathrm{e}^{a/b} / (2\sqrt{b})$, hence $c_1 = c_2 = \pi \, \mathrm{e}^{a/b} / (2\sqrt{b})$, and we conclude the proof. □

4.3.3 Empirical study

Let us use the theoretical results of Section 4.3.2 to study the connection between $c_{\max}$ and $\underline{\Delta}^2$.

The sample size is set to $n = 200$. We consider two regression functions:

— μ^1 such that for any $i \in \{1, \ldots, n/2\}$, $\mu^1(i) = 0.0$ and for any $i \in \{n/2 + 1, \ldots, n\}$, $\mu^1(i) = 5.0$;

— μ^2 constant equal to zero.

We generate samples X^1 by adding i.i.d. standard Gaussian noise to μ^1, and X^2 by adding standard Gaussian noise for $i \in \{1, \ldots, n/2\}$ and $\mathcal{N}(0, 5^2)$ for $i \in \{n/2 + 1, \ldots, n\}$. In a nutshell, X^1 and X^2 are semi-deterministic versions of the signals we used so far in this chapter — there is still randomness in the noise. For 100 values of ν in $[0, 20]$, we generate 100 samples of X^1 and X^2. For each sample, we compute $c_{\max}$ via algorithm 4.2, for X^1 using both the Gaussian and the Laplace kernel, for

X^2 using only the Laplace kernel. We also compute $\underline{\Delta}^2(\nu)$ thanks to Prop. 4.2, 4.3 and 4.4. The results are reported in Fig. 4-6.

Again, the bound given in Theorem 3.1 is accurate, and $\underline{\Delta}^2$ and $c_{\max}$ are linearly correlated — all R^2 coefficients we obtained for a linear regression of $c_{\max}$ versus $\underline{\Delta}^2$ are > 0.999. We find this result remarkable for two reasons.

— First, it shows that the analysis conducted in Chapter 3 brings out a relevant quantity, $\underline{\Delta}^2$. Note however that we got rid of the influence of the segment sizes in the present experiments. We think that the true quantity of interest is a trade-off between size of the jump in the RKHS and size of the segments adjacent to the jump. We refer to Eq. (3.27), the point in our analysis where we begin working with de-correlated jump size and segment size.

— Second, this link between $c_{\max}$ and $\underline{\Delta}$ gives us a route to choosing the kernel for KCP. By taking a kernel that maximizes $\underline{\Delta}$, we have a shot at maximizing $c_{\max}$. Of course $\underline{\Delta}$ depends on unknown quantities — for instance, the size of the jump that we are trying to detect —, but it is conceivable to find an estimator of $\underline{\Delta}$ which could act as a proxy for this goal.

Finally, note that we also report in Fig. 4-6 the bandwidth proposed by the median heuristic. This simple heuristic consist to pick $2\nu^2 = \text{Med}\{|X_i - X_j|^2\}$ for the Gaussian kernel, where Med is the sample median. As we can see, it is generally close to the global maximum of $\underline{\Delta}^2$, and we recommend the use of this heuristic for choosing the bandwidth of the Gaussian and Laplace kernels.

We study the median heuristic in more depth in the next chapter.

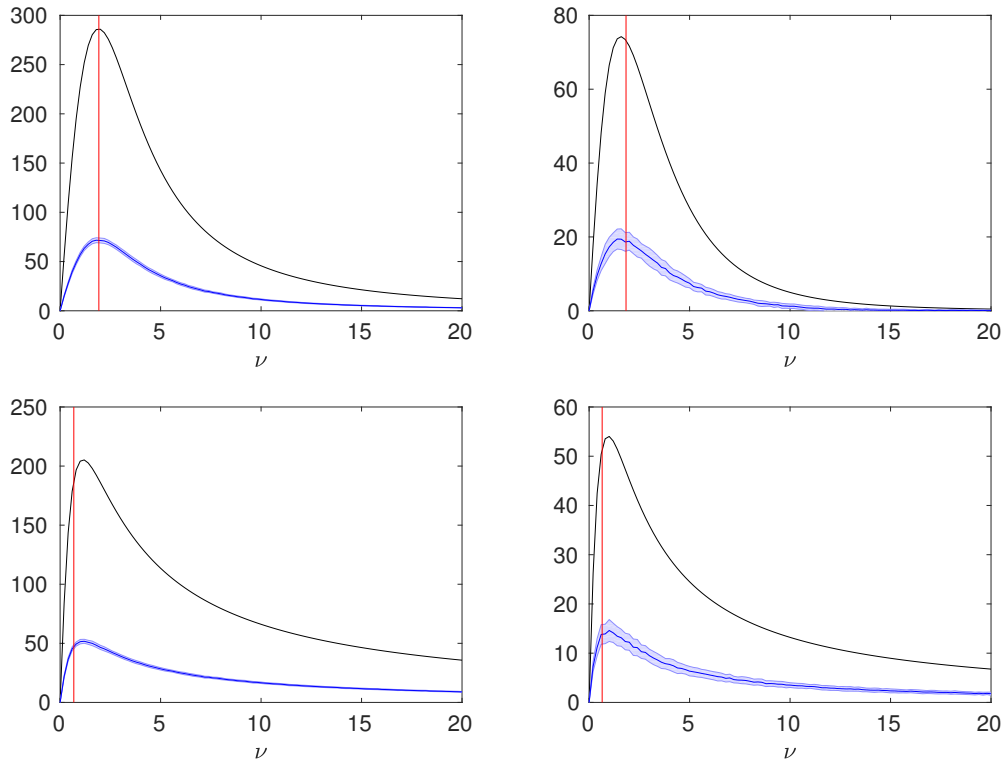

Figure 4-6 – In this figure, we plot $n \cdot \underline{\Delta}^2/M^2$ (in black) and $c_{\max}$ (in blue) as a function of ν in different scenarios. The error bars for $c_{\max}$ come from 100 repetitions of the experiment, the vertical red line is the average median heuristic bandwidth (standard deviation was very small and is not represented). *Upper left:* Gaussian kernel, change in the mean; *Upper right:* Gaussian kernel, change in the variance; *Bottom left:* Laplace kernel, change in the mean; *Bottom right:* Laplace kernel, change in the variance.

Chapter 5

The median heuristic

Abstract

The median heuristic is a popular tool to set the bandwidth of radial basis function kernels. While its empirical performances make it a safe choice under most circumstances, there is little theoretical understanding of why this is the case. For a large sample size, we show in this chapter that the median heuristic behaves approximately as the median of a distribution that we describe completely in the setting of kernel two-sample test and kernel change-point detection. More precisely, we show that the median heuristic is asymptotically normal around this value. We illustrate these findings when the underlying distributions are multivariate Gaussian distributions. This chapter is based upon the preprint Garreau [2017].

5.1 Introduction

Kernel methods form an important class of algorithms in machine learning and statistics. They make use of rich feature spaces that depend only on the kernel chosen by the user. Given a positive semi-definite kernel k and observations $x_1, \dots, x_n$, the first step of most kernel-based method is to compute the Gram matrix $\mathrm{K} = (k(x_i, x_j))_{1 \leq i,j \leq n}$. Thanks to the celebrated *kernel trick*, all ensuing computations need only the knowledge of K.

We are especially interested in data lying in a metric space $(\mathcal{X}, \mathrm{d})$. When this is the case, commonly used kernels are *radial basis function kernels*. They can be written

$$k(x, y) = f(\mathrm{d}(x, y)/\nu), \tag{5.1}$$

where $f : \mathbb{R}_+ \to \mathbb{R}_+$ and ν is a positive parameter called the *bandwidth*. In many applications, the space $\mathcal{X}$ is $\mathbb{R}^d$ and the distance d is derived from the Euclidean norm $\|x\| = \sqrt{\sum_i x_i^2}$, that is, $\mathrm{d}(x, y) = \|x - y\|$.

Among this class of kernel are found numerous kernels often used in practice. For instance, $f(x) = \exp(-x^2)$ corresponds to the Gaussian kernel [Aizerman et al., 1964], arguably the most popular positive definite kernel used in applications (see, for instance, Vert et al. [2004]). The function $f(x) = \exp(-x)$ yields the exponential

kernel – also called Laplace or Laplacian kernel–, whereas more exotic f give rise to less common kernels such as the rational quadratic kernel, the wave kernel or the Matérn kernel [see Genton, 2001, and references therein].

It is well-known that the performances of kernel methods depend highly on the kernel choice, see for instance Sriperumbudur et al. [2009]. However, the calibration of the bandwidth ν is perhaps even more important as the choice of f [Schölkopf and Smola, 2002, Sec. 4.4.5].

Since the Gram matrix depends only on the $\|x_i - x_j\| / \nu$ in this case, it is common sense to pick ν of the same order than the family of all pairwise distances $(\|x_i - x_j\|)_{1 \leq i,j \leq n}$. As an example, suppose that we settled for the Gaussian kernel k_{G}. Then when $\nu \to 0$, the Gram matrix K is the identity matrix, and when $\nu \to \infty$, the components of K are constant equal to 1. All relevant information about the data is lost in both these extreme cases. This is a general phenomenon, even though the values taken by K in the degenerate cases depend on the function f. Hence a reasonable middle-ground for choosing ν is to pick a value "in the middle range" of the $(\|x_i - x_j\|)_{1 \leq i,j \leq n}$, that is, an empirical quantile, which is often set to be the median. This strategy is called the median heuristic.

As noted in Flaxman et al. [2016], the origin of the median heuristic is quite unclear and does not appear in the monograph of Schölkopf and Smola [2002], while it has become the main reference for this heuristic. The earliest appearance of the median heuristic that we know of is in Sriperumbudur et al. [2009, Sec. 5]. Let us also mention Gretton et al. [2012a], which refers to Takeuchi et al. [2006] and Schölkopf et al. [1997] for similar heuristics.

Nevertheless, the median heuristic is extensively used in practice. In a supervised learning setting, *e.g.*, kernel SVM [Boser et al., 1992] or kernel ridge regression [Hoerl and Kennard, 1970], one can always resort to a grid-search for the choice of ν. The performance of each prescribed value for ν is then evaluated by cross-validation. But whenever cross-validation is too expensive, the median heuristic can provide a good alternative. It is for instance the default bandwidth choice in the kernel SVM implementation of the R language `kernlab` package [Karatzoglou et al., 2004].

In the context of hypothesis testing, a possibility is to try and find the bandwidth that maximizes the test power. This is for instance what is done in [Gretton et al., 2012a; Jitkrittum et al., 2016; Sutherland et al., 2017]. However, it can be challenging to compute the test power, and further to maximize it with respect to the bandwidth. In the unsupervised setting, the choice of the kernel and more specifically the choice of the bandwidth ν has no definitive answer. Hence heuristics as the median heuristic are among the first choice that comes to mind when one has to choose a bandwidth, and numerous authors report using the median heuristic in their experiments [Sriperumbudur et al., 2009; Arlot et al., 2012; Reddi et al., 2015; Zhang et al., 2017; Jitkrittum et al., 2016; Muandet et al., 2016; Sutherland et al., 2017]. Note that it is also possible to use other rule-of-thumbs, coming for instance from kernel density estimation where the problem of the bandwidth choice is also of the utmost importance – see, *e.g.*, Harchaoui and Cappé [2007] in the context of kernel change-point detection.

5.1.1 Related work

Despite its popularity, there is very little theoretical understanding of the median heuristic. To the best of our knowledge, the only work in this direction is contained in Reddi et al. [2015]. They observe that the median of all the pairwise distances has to be close to the mean pairwise distance $\mathbb{E}\|X_i - X_j\|$, the computation of which is then used to obtain the asymptotic of the median heuristic when the dimension of the data goes to infinity. This argument can be made rigorous by observing that, given a random variable X with a second order moment, the following inequality holds [Mallows, 1991]:

$$|\mathbb{E}\, X - \mathrm{med}(X)| \leq \sqrt{\mathrm{Var}\,(X)}\,.$$

Hence the observation of Reddi et al. [2015] is correct, up to a variance term. We will see in Section 5.4 that our results make this insight more precise.

5.1.2 Outline

Our goal in this chapter is to obtain a precise understanding of the median heuristic for a large sample size in the setting of kernel two-sample test and off-line kernel change-point detection. Our setting is made explicit in Section 5.2, and we show in the same section how it is relevant for these applications. In Section 5.3, we claim our main result: the median heuristic is asymptotically normal when the number of observations goes to ∞. In particular, the median heuristic converges towards the theoretical median of a target distribution that we describe completely. This result is obtained thanks to an auxiliary proposition that we think has an interest of its own, namely a central limit theorem for a certain class of U-statistics that we state and prove in the same section. We demonstrate with numerical experiments the validity of our claims.

5.2 Setting

Given any random variable Z, the notation Z' will stand for an independent copy of Z. Unless specified in subscript, the expected value is taken with respect to all the random variables that appear in the expression. For instance, $\mathbb{E}\,[h(X,Y)]$ means $\mathbb{E}_{X,Y}\,[h(X,Y)]$. We also denote by $L^2(P)$ the space of real functions such that $\mathbb{E}\,[f(X)^2] < +\infty$ where $X \sim P$.

In the following, we suppose that we are given a *triangular array* of independent $\mathbb{R}^d$-valued random variables. Namely, for each n, we suppose that the observations

are drawn from line n of the following scheme:

$$\begin{array}{cccccc} X_{1,1} & & & & & \\ X_{2,1} & X_{2,2} & & & & \\ \vdots & & \ddots & & & \\ X_{n,1} & \cdots & \cdots & X_{n,n} & & \\ \vdots & & & & \ddots & \end{array}$$

Let X (resp. Y) be a $\mathbb{R}^d$-valued random variable following the law P (resp. Q). Our main hypothesis on the distribution of the $X_{n,i}$ is the following:

Assumption 5.1. There exists $\alpha \in (0,1)$ such that $X_{n,i} \sim P$ for any $i \leq \alpha n$ and $X_{n,i} \sim Q$ otherwise.

We will assume from now on that αn is an integer. Everything that follows can be readily adapted by replacing αn with $\lfloor \alpha n \rfloor$ when it is needed. Assumption 5.1 means that our observations are split in two segments, $\{1, \ldots, \alpha n\}$ and $\{\alpha n + 1, \ldots, n\}$. On the left segment, they follow P and on the right segment they follow Q, as illustrated below:

$$\underbrace{X_{n,1} X_{n,2} \cdots X_{n,\alpha n}}_{\sim P} \Big| \underbrace{X_{n,\alpha n+1} \cdots X_{n,n}}_{\sim Q}$$

5.2.1 Connection with kernel two-sample test

Let us briefly recall the modus operandi of kernel two-sample test. Suppose that, for a given n, we sample observations $x_1, \ldots, x_M$ of X and observations $y_1, \ldots, y_N$ of Y. The goal of two-sample test is to decide whether $P = Q$ or $P \neq Q$ given these observations. Gretton et al. [2007] have proposed a kernel method for two-sample testing, that relies on the mean embeddings of P and Q inside the reproducing kernel Hilbert space $\mathcal{H}$ associated with k. Let us call μ_P and μ_Q these embeddings, then a good measure of proximity between the distributions P and Q is the so called *maximum mean discrepancy* (MMD), that can be written $\|\mu_P - \mu_Q\|_{\mathcal{H}}$. It is also possible to write the (squared) MMD as

$$\mathrm{MMD}^2(p,q) = \mathbb{E}\left[k\left(x, x'\right)\right] - 2\,\mathbb{E}\left[k\left(x, y\right)\right] + \mathbb{E}\left[k\left(y, y'\right)\right] .$$

It is proven in Gretton et al. [2012a] that an unbiased estimate of this quantity is

$$\begin{aligned} \widehat{\mathrm{MMD}}^2(P,Q) = &\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{j \neq i}^{M} k\left(x_i, x_j\right) \\ &+ \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j \neq i}^{N} k\left(y_i, y_j\right) - \frac{2}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} k\left(x_i, y_j\right) . \end{aligned}$$

The setting that we described in Section 5.2 corresponds to the kernel two-sample test setting when we take $M = \alpha n$, $N = (1-\alpha)n$ and the x_i are realizations of the $X_{n,i}$. Letting n grow to infinity will correspond to let both M and N grow to infinity with the ratio M/N constant equal to $\alpha/(1-\alpha)$.

5.2.2 Connection with kernel change-point detection

We restrict here to the case when there is a single change-point, thus both the approaches of Harchaoui and Cappé [2007] and Arlot et al. [2012] coincide and may be summarized as follow: suppose that we are given observations $x_1, \ldots, x_n \in \mathcal{X}$ drawn from $X_1, \ldots, X_n$ such that $P_{X_\cdot}$ is a step function. That is, there exists an *unknown* change-point $1 \leq \tau_1 \leq n$ such that the X_i share a common distribution for $1 \leq i \leq \tau_1$, say P, and another for $\tau_1 < i \leq n$, say $Q \neq P$. The kernel change-point detection procedures then consider the minimization of the *kernel least-squares criterion*

$$\underset{1 \leq \tau_1 \leq n}{\text{Minimize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} k(x_i, x_i) - \frac{1}{n} \left[\frac{1}{\tau_1} \sum_{i,j=1}^{\tau_1} k(x_i, x_j) + \frac{1}{n - \tau_1} \sum_{i,j=\tau_1+1}^{n} k(x_i, x_j) \right] \right\}$$

to estimate τ_1. This corresponds to our setting when we let the x_i be realizations of $X_{n,i}$ for any $1 \leq i \leq n$ and set $\tau_1 = \alpha n$.

Remark 5.1. Our setting can be easily adapted to multiple change-points, that is, when there are more than one change-point. Indeed, set D the –possibly unknown– number of segments, and $\alpha_1, \ldots, \alpha_D$ positive numbers such that $\sum_\ell \alpha_\ell = 1$ the lengths of the segments. In addition, set $\alpha_0 = \alpha_{D+1} = 0$ and let $X_{n,i}$ follow the distribution P_ℓ if $\alpha_0 + \cdots + \alpha_{\ell-1} n < i \leq \alpha_0 + \cdots + \alpha_\ell n$. Then the single change-point case corresponds to $D = 2$ and $\alpha_1 = \alpha$. Note that this is the asymptotic interpretation of off-line change-point detection presented in Chapter 1: taking $n \to \infty$ corresponds to letting the number of observations on each segments grow to infinity while the unknown change-points remains constant.

5.2.3 The median heuristic

Suppose that we use a kernel that has the form (5.1) for a fixed f. Both for kernel two-sample test and kernel change-point detection, the choice of the kernel thus boils down to the choice of the bandwidth ν. The power of the kernel two-sample test procedure that we recalled in Sec. 5.2.1 is known to have maximum power when the kernel maximizes the MMD divided by its standard deviation [Gretton et al., 2012b]. This is approximately true for the quadratic MMD as well [Sutherland et al., 2017]. A similar situation occurs in the kernel change-point setting of Sec. 5.2.2: as we have seen in Section 4.3.1, $\underline{\Delta}^2$ is a key quantity in kernel change-point detection. Thus it would be very interesting to know whether the median heuristic picks a bandwidth that achieves these goals.

We first define

$$H_n = \text{Med}\left\{ \|X_{n,i} - X_{n,j}\|^2 \mid 1 \leq i < j \leq n \right\}, \tag{5.2}$$

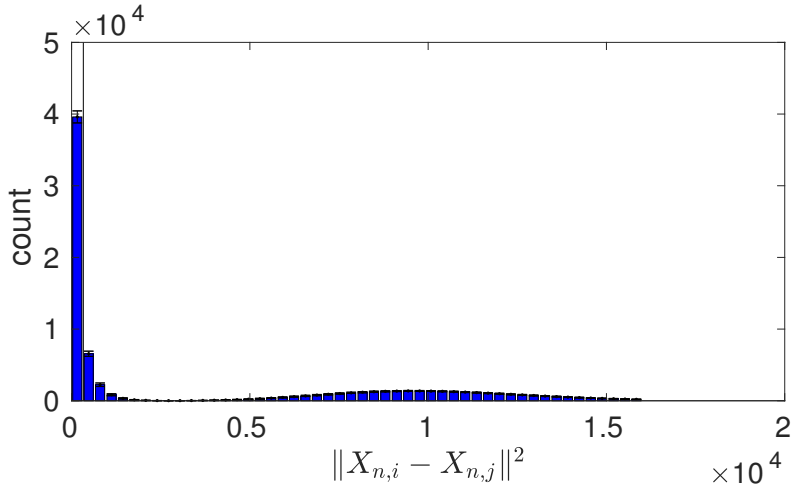

Figure 5-1 – Histogram of the $\|X_{n,i} - X_{n,j}\|^2$ for Gaussian distributions in dimension $d = 100$. Namely, $X \sim \mathcal{N}(0, \mathrm{I}_d)$, $Y \sim \mathcal{N}(10d\mathbb{1}, \mathrm{I}_d)$, $\alpha = .25$ and $n = 400$. We repeated the experiment 10^3 times, counting each time how many observations were into the bins, error bars are standard deviations over these repetitions; the vertical black line is the average sample median.

where Med is the empirical median. That is (i) order the $\|X_{n,i} - X_{n,j}\|^2$ in increasing order, (ii) output the central element if $n(n-1)/2$ is odd ($n \equiv 2, 3 \mod 4$) and the mean of the two most central elements if $n(n-1)/2$ is even ($n \equiv 0$ or $1 \mod 4$). We call median heuristic the choice $\nu = \sqrt{H_n}$. Note that some authors choose $\nu = \sqrt{H_n/2}$.

In order to investigate the asymptotic properties of H_n, rather than using (5.2), we are going to define H_n *via* the empirical cumulative distribution function of the $\|X_{n,i} - X_{n,j}\|^2$. Namely, for any $t \in \mathbb{R}$, we let

$$\widehat{F}_n(t) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \mathbb{1}_{\|X_{n,i} - X_{n,j}\|^2 \le t} \tag{5.3}$$

For any $p \in (0,1)$, we define the generalized inverse of $\widehat{F}_n$ by

$$\widehat{F}_n^{-1}(p) = \inf\left\{t \in \mathbb{R} \,|\, \widehat{F}(t) \ge p\right\}.$$

We choose to define H_n as

$$H_n = \widehat{F}_n^{-1}\left(\frac{1}{2}\right). \tag{5.4}$$

Notice that definitions (5.2) and (5.4) differ whenever $n \equiv 0$ or $1 \pmod 4$.

Remark 5.2. It is tempting to define other empirical quantiles as $\widehat{F}_n^{-1}(p)$ for $p \in (0,1)$, and indeed such quantiles are used in practice, often for $p = 0.1$ and $p = 0.9$. Though we are mainly concerned with H_n, we will see that our main result still holds for

arbitrary p.

5.3 Main results

5.3.1 The empirical distribution function

Under Assumption 5.1, there are three possibilities for an element of the family

$$T_n = \left\{ \|X_{n,i} - X_{n,j}\|^2 \mid 1 \leq i < j \leq n \right\}.$$

Namely

1. $i \leq \alpha n$ and $j \leq \alpha n$: Then $\|X_{n,i} - X_{n,j}\|^2$ has the distribution of $\|X - X'\|^2$, which we call T_{XX};
2. $i > \alpha n$ and $j > \alpha n$: Then $\|X_{n,i} - X_{n,j}\|^2$ has the distribution of $\|Y - Y'\|^2$, which we call T_{YY};
3. $i \leq \alpha n$ and $j > \alpha n$: Then $\|X_{n,i} - X_{n,j}\|^2$ has the distribution of $\|X - Y\|^2$, which we call T_{XY};

There are $\alpha n(\alpha n - 1)/2$ occurrences of case (i). Suppose that we make $n \to \infty$, then case (i) occurs with proportion α^2. Similarly, case (ii) occurs with proportion $(1-\alpha)^2$ and case (iii) with proportion $2\alpha(1-\alpha)$.

Define a mixture distribution $T \sim T_{XX}$, $T \sim T_{YY}$ and $T \sim T_{XY}$ with weights α^2, $(1-\alpha)^2$ and $2\alpha(1-\alpha)$ respectively. Thereafter, we will call T the *target* distribution and denote by F its cumulative distribution function. The non-rigorous reasoning above suggests that when $n \to \infty$, T_n behaves like a n-sample of the target distribution T. Indeed, a specialization of a result stated in the next paragraph shows that

$$\forall t \in \mathbb{R}, \qquad \widehat{F}_n(t) \xrightarrow{\mathbb{P}} F(t) \,. \tag{5.5}$$

Fig. 5-1 illustrates this phenomenon. For large n, if we plot the histogram of the $\|X_{n,i} - X_{n,j}\|^2$ for $1 \leq i < j \leq n$, then the "two bumps" behavior depicted in Fig. 5-1 is typical. The left mode of the empirical distribution corresponds to T_{XX} and T_{YY}, close to zero by definition, whereas the right mode corresponds to T_{XY} that can be arbitrarily far from 0. Eq. (5.5) is already a step in the comprehension of the median heuristic, since we are now able to think about H_n "approximately" as the theoretical median of the target distribution T.

It turns out that (5.5) is a trivial consequence of a much stronger statement. Indeed, $\widehat{F}_n(t)$ can be seen as a sum of three dependent U-statistic with kernel $h(x,y) = \mathbb{1}_{\|x-y\|^2 \leq t}$, and the following result shows that it follows a central limit theorem. We refer to classical textbooks [Lee, 1990; Korolyuk and Borovskich, 2013] for an introduction to the theory of U-statistics.

Proposition 5.1 (CLT for non-identically distributed triangular array U-statistic)**.** *Consider $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that $h \in L^2(P) \cap L^2(Q) \times L^2(P) \cap L^2(Q)$, and suppose*

that the $X_{n,i}$ satisfy Assumption 5.1. Define

$$U_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} h(X_{n,i}, X_{n,j}) \,,$$

and set

$$\theta = \alpha^2 \, \mathbb{E}\left[h(X, X')\right] + 2\alpha(1-\alpha) \, \mathbb{E}\left[h(X, Y)\right] + (1-\alpha)^2 \, \mathbb{E}\left[h(Y, Y')\right] \,.$$

Then

$$\sqrt{n}(U_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right) \,, \tag{5.6}$$

where $\sigma = \sigma(h, P, Q)$ is defined in Eq. (5.9).

We make the following remarks.

Central limit theorems for U-statistics are known since the fundamental article of Hoeffding [1948]. Prop. 5.1 is in the line of such results. An asymptotic normality result also exists in the non-identically distributed case, see Hoeffding [1948, Th. 8.1]. However, this result was not applicable in our triangular array setting. The material in Jammalamadaka and Janson [1986] covers the case of a triangular array scheme, but does not cover the non-identically distributed setting. Results regarding *two-sample* U-statistics are closest in spirit but not directly applicable, see van der Vaart [1998, Sec. 12.2] for an introduction and Dehling and Fried [2012] for recent developments. With our notations, the two-sample statistic is written

$$\frac{1}{\alpha n (1-\alpha) n} \sum_{i=1}^{\alpha n} \sum_{j=\alpha n + 1}^{n} h(X_{n,i}, X_{n,j}) \,.$$

The sole difference is the absence of "intra-segment" interactions: the previous display does not contain terms in $h(X_{n,i}, X_{n,j})$ with i and j in the same segment. It is the appearance of these terms in our case which complicates the analysis.

Finally, suppose that h is degenerate, that is, $\mathbb{E}\, h(X, y) = \mathbb{E}\, h(x, Y) = 0$. Then the variance term in Eq. (5.6) is zero, and Prop. 5.1 remains true in the following sense: $\sqrt{n}(U_n - \theta)$ converges towards the constant 0, which is a degenerate Gaussian distribution $\mathcal{N}(0, 0)$. In this case, we believe that the convergence will be faster, but not toward a Gaussian distribution, c.f. Lee [1990, Section 3.2.2] for results in this direction.

Remark 5.3. It is possible to prove a version of Prop. 5.1 for the multiple change-point setting introduced in Rem. 5.1. The proof follows the lines of Sec. 5.3.2, with an additional technical difficulty due to the numerous inter-segment interactions – we only deal with one in the present work.

5.3.2 Proof of Prop. 5.1

The idea of the proof is the following: (i) split U_n in three terms depending on the relative position of the indices; (ii) write down the Hoeffding decomposition of

each of these terms; (iii) show that the remainders are negligible, and (iv) conclude thanks to the central limit theorem for triangular arrays.

We begin by decomposing U_n. To this extent, define

$$A_n = \binom{\alpha n}{2}^{-1} \sum_{1 \le i < j \le \alpha n} h(X_{n,i}, X_{n,j}) \,,$$

$$B_n = \binom{(1-\alpha) n}{2}^{-1} \sum_{\alpha n < i < j \le n} h(X_{n,i}, X_{n,j}) \,,$$

and

$$C_n = \frac{1}{\alpha n (1-\alpha) n} \sum_{\substack{1 \le i \le \alpha n \\ \alpha n < j \le n}} h(X_{n,i}, X_{n,j}) \,.$$

Note that in C_n there are no terms $h(X_{n,i}, X_{n,j})$ with $1 \le j \le \alpha n$ and $\alpha n < i \le n$, since the sum in U_n is prescribed to $i < j$. Simple algebra shows that

$$\begin{aligned} U_n = {} & \alpha^2 A_n + (1-\alpha)^2 B_n + 2\alpha(1-\alpha) C_n \\ & - \frac{\alpha(1-\alpha)}{n-1} A_n - \frac{\alpha(1-\alpha)}{n-1} B_n - \frac{2\alpha(1-\alpha)}{n-1} C_n \,. \end{aligned} \tag{5.7}$$

According to Lemma 5.1 – see Section 5.6–, the variance of A_n, B_n and C_n is $\mathrm{O}\left(1/n\right)$. Hence the second line of Eq. (5.7) converges in probability to 0 with speed at least $\sqrt{n}$. Therefore we can focus on the first line of (5.7).

The next step is to obtain the H-decomposition [Lee, 1990, Sec. 1.6] of A_n and B_n. Let us detail this process for A_n. We set $\theta_A = \mathbb{E}\left[h(X, X')\right]$, $h_A(x) = \mathbb{E}\left[h(x, X')\right] - \theta_A$ and $g_A(x, y) = h(x, y) - h_A(x) - h_A(y) - \theta_A$. Then it is possible to write [Lee, 1990, Th. 1]

$$A_n = \theta_A + L_A + R_A \,,$$

where

$$L_A = \frac{2}{\alpha n} \sum_{1 \le i \le \alpha n} h_A(X_{n,i}) \,,$$

and

$$R_A = \binom{\alpha n}{2}^{-1} \sum_{1 \le i < j \le \alpha n} g_A(X_{n,i}, X_{n,j}) \,.$$

A totally analogous statement holds for B_n. We decompose C_n in the same fashion, the only difference is the appearance of a second term in the linear part. Namely, set $\theta_C = \mathbb{E}\left[h(X, Y)\right]$, $h_{C,1}(x) = \mathbb{E}\left[h(x, Y)\right]$, $h_{C,2}(y) = \mathbb{E}\left[h(X, y)\right]$ and $g_C(x, y) = h(x, y) - h_{C,1}(x) - h_{C,2}(y) - \theta_C$. Then $C_n = \theta_C + L_C + R_C$, with

$$L_C = \frac{1}{\alpha n} \sum_{i=1}^{\alpha n} h_{C,1}(X_{n,i}) + \frac{1}{(1-\alpha) n} \sum_{i=\alpha n+1}^{n} h_{C,2}(X_{n,i}) \,,$$

and

$$R_C = \frac{1}{\alpha n(1-\alpha)n} \sum_{\substack{1\le i\le \alpha n \\ \alpha n < j \le n}} g_C(X_{n,i}, X_{n,j})\,.$$

Note that $\theta = \alpha^2\theta_A + (1-\alpha)^2\theta_B + 2\alpha(1-\alpha)\theta_C$. According to Lemma 5.2, the variance of R_A, R_B and R_C is of order n^{-2}, thus

$$\sqrt{n}(U_n - \theta) = \sqrt{n}\left[\alpha^2 L_A + (1-\alpha)^2 L_B + 2\alpha(1-\alpha)L_C\right] + r_n\,, \tag{5.8}$$

with $r_n \xrightarrow{\mathbb{P}} 0$.

We now regroup the terms in (5.8) that belong to the same segment. For any $1 \le i \le \alpha n$, define

$$Z_{n,i}^{(1)} = \alpha h_A(X_{n,i}) + (1-\alpha)h_{C,1}(X_{n,i})\,.$$

Since $h \in L^2(P)\cap L^2(Q) \times L^2(P)\cap L^2(Q)$ and the $Z_{n,i}^{(1)}$ are identically distributed,

$$\mathrm{Var}\left(Z_{n,i}^{(1)}\right) = \mathrm{Var}\left(\alpha h_A(X) + (1-\alpha)h_{C,1}(X)\right)$$

is finite and does not depend on i. Let us put $\sigma_1^2 := \mathrm{Var}\left(Z_{n,i}^{(1)}\right)$. The Lindeberg condition is satisfied, hence we can apply the central limit theorem for triangular arrays of independent random variables Billingsley [2012, Th. 27.2]. Thus

$$\frac{1}{\sqrt{\alpha n}} \sum_{i=1}^{\alpha n} Z_{n,i}^{(1)} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma_1^2\right)\,.$$

In a similar fashion, set $Z_{n,i}^{(2)} := \alpha h_{C,2}(X_{n,i}) + (1-\alpha)h_B(X_{n,i})$ and $\sigma_2^2 := \mathrm{Var}\left(Z_{n,i}^{(2)}\right)$. Then

$$\frac{1}{\sqrt{(1-\alpha)n}} \sum_{i=\alpha n+1}^{n} Z_{n,i}^{(2)} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma_2^2\right)\,.$$

The two previous sums are independent, thus by Lévy's theorem

$$\sqrt{n}\left[\alpha^2 L_A + (1-\alpha)^2 L_B + 2\alpha(1-\alpha)L_C\right] \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2\right)\,,$$

with

$$\sigma^2 = \sigma_1^2 + \sigma_2^2\,. \tag{5.9}$$

Since the remainder term converges in probability to 0, we can conclude *via* Slutsky's Lemma.

5.3.3 Asymptotic normality of H_n

We now turn to the statement of our main result. In the previous section, we only obtained the convergence of the empirical distribution function. It is well-known that such a result implies the convergence of the empirical quantiles towards the

theoretical quantiles of the target distribution if the convergence of the empirical distribution function is "strong enough" [van der Vaart, 1998, Chapter 21]. More precisely, if this convergence is uniform or follows a CLT – as in our case.

Proposition 5.2 (Asymptotic normality of H_n)**.** *Suppose that Assumption 5.1 holds, and define T as in Sec. 5.3.1. Define $m = \mathrm{med}(T)$ the theoretical median of the target distribution. Suppose that F has a non-zero derivative at m, and define $\sigma = \sigma(\ell, P, Q)$ as in Prop. 5.1, where*

$$\ell(x, y) = \mathbb{1}_{\|x-y\|^2 \leq m} \ .$$

Then

$$\sqrt{n}(H_n - m) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\sigma^2}{F'(m)^2}\right) . \tag{5.10}$$

We illustrate Prop. 5.2 in Fig. 5-2. Before we turn to the proof of Prop. 5.2, we make a few remarks.

Note that we made very few assumptions on the distribution of the $X_{n,i}$s, hence Prop. 5.2 can be applied in a wide range of situations. Of course, when met with an experimental problem, one does not know P and Q. Nevertheless, Prop. 5.2 suggests that H_n is concentrated around some deterministic value depending on the data at hand for large n. If more information is available regarding P and Q, we will demonstrate in the next section that it is possible to transfer this information to $\mathrm{med}(T)$, hence to H_n.

Once again, this result should not come as a shock to the knowledgeable reader since empirical U-quantiles are known to satisfy asymptotic normality since Serfling [1980] in the i.i.d. case. Though a lot of work has been done to relax the independence assumption, to the best of our knowledge there is no result regarding the non-identically distributed case. In the two-sample setting, some results exist, both in the independent case [Lehmann, 1951] and with some dependence structure [Dehling and Fried, 2012]. However, as noted before, in our setting it is necessary to consider the intra-segment interactions.

Note that, as it can be seen for instance in Fig. 5-1, observations lying between $\max(\mathbb{E}\,T_{XX}, \mathbb{E}\,T_{YY})$ and $\mathbb{E}\,T_{XY}$ are quite scarce. Therefore, it is possible for $F'(m)$ to be small, leading to a large variance term in (5.10). Note that $F'(m) \neq 0$ does not hold for arbitrary continuous distributions.

5.3.4 Proof of Prop. 5.2

Set $t \in \mathbb{R}$. The general idea of the proof is to rewrite statements about the event $\left\{\sqrt{n}\,(H_n - m) \leq t\right\}$ as statements about a sum of U-statistics. We will then control these U-statistics with Prop. 5.1 for conveniently chosen h, and conclude with Slutsky's Lemma. Throughout this proof, we only suppose that $p \in (0, 1)$ to emphasize that Prop. 5.2 can be extended to *any* quantile, not only the median.

We use the property of the generalized inverse to obtain

$$\left\{\sqrt{n}\,(H_n - m) \leq t\right\} = \left\{p \leq \widehat{F}_n\left(m + \frac{t}{\sqrt{n}}\right)\right\},$$

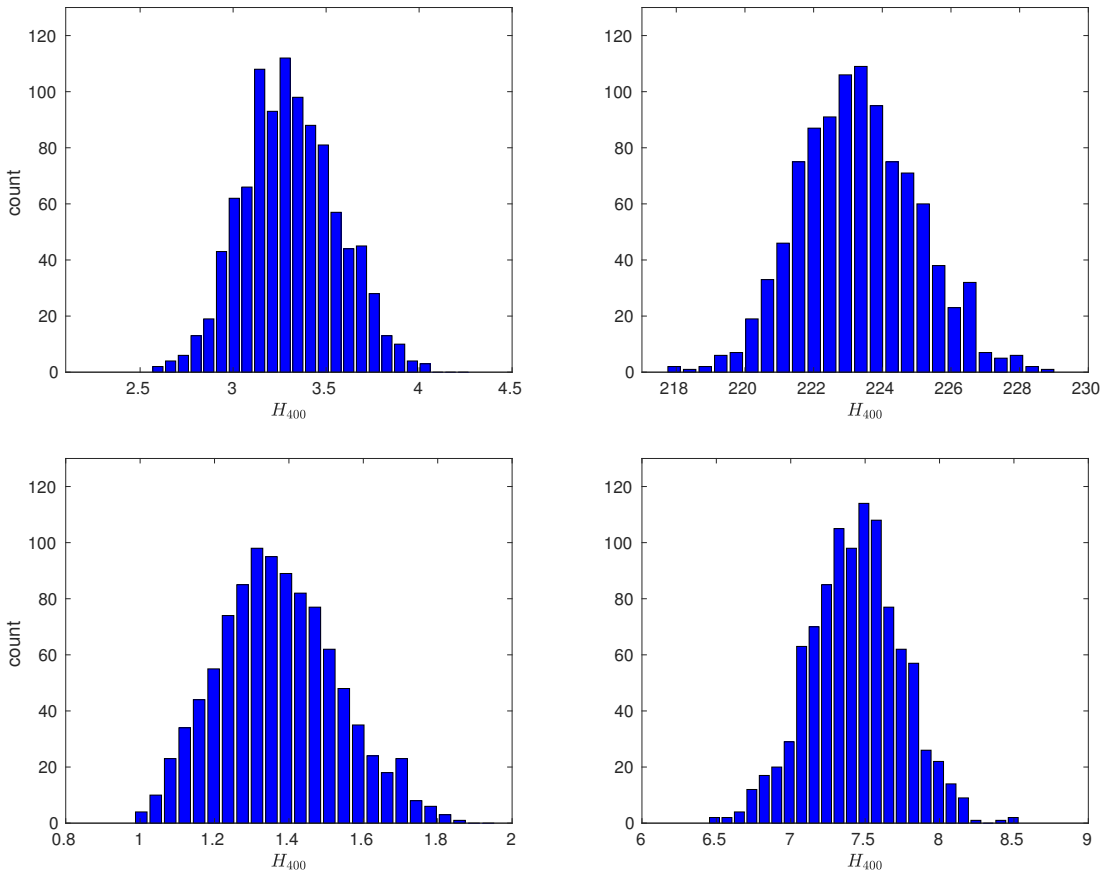

Figure 5-2 – Empirical distribution of H_{400} for different distributions P and Q over 10^3 repetitions. *Upper left:* $P \sim \mathcal{N}(0,1)$ and $Q \sim \mathcal{N}(10,1)$; *Upper right:* $P \sim \mathcal{N}(0, \mathrm{I}_{100})$ and $Q \sim \mathcal{N}(10\,\mathbb{1}_{100}, \mathrm{I}_{100})$; *Bottom left:* $P \sim \Gamma(2,2)$ and $Q \sim \mathcal{E}(1)$; *Bottom right:* both P and Q are Gaussian mixtures $\mathcal{N}(0,1)$ and $\mathcal{N}(5,1)$ with equal weight. The results of Kolmogorov-Smirnov tests were non-significant, indicating that H_{400} is normally distributed in each case.

and rewrite this event as

$$\left\{\sqrt{n}\left(\widehat{F}_n\left(m+\frac{t}{\sqrt{n}}\right)-F\left(m+\frac{t}{\sqrt{n}}\right)\right)\geq\sqrt{n}\left(p-F\left(m+\frac{t}{\sqrt{n}}\right)\right)\right\}. \quad (5.11)$$

Since F is differentiable in m, the right-hand side of (5.11) converges towards $-tF'(m)$ by Taylor expansion. From Prop. 5.1, it is also true that

$$\sqrt{n}\left(\widehat{F}_n(m)-F\left(m\right)\right)\xrightarrow{\mathcal{L}}\mathcal{N}\left(0,\sigma^2\right).$$

Therefore, if we manage to prove that

$$\sqrt{n}\left[\left(\widehat{F}_n\left(m+\frac{t}{\sqrt{n}}\right)-\widehat{F}_n(m)\right)-\left(F\left(m+\frac{t}{\sqrt{n}}\right)-F\left(m\right)\right)\right]\xrightarrow{\mathbb{P}}0, \quad (5.12)$$

Eq. (5.10) will follow by Slutsky's Lemma.

Define $h(x,y)=\mathbb{1}_{m<\|x-y\|^2\leq m+\frac{t}{\sqrt{n}}}$. Then, with the notations used in the proof of Prop. 5.1, Eq. (5.12) reads

$$\sqrt{n}(U_n-\theta)\xrightarrow{\mathbb{P}}0.$$

We dispose of the remainder terms as in the proof of Prop. 5.1, thus we are left to show that

$$\sqrt{n}\left[\alpha^2(A_n-\theta_A)+(1-\alpha)^2(B_n-\theta_B)+2\alpha(1-\alpha)(C_n-\theta_C)\right]\xrightarrow{\mathbb{P}}0. \quad (5.13)$$

Let us focus on the first term of the previous display, which can be written

$$\alpha^2\sqrt{n}\left(A_n-\theta_A\right)=\alpha^2\sqrt{n}\left(L_A+R_A\right).$$

Once again, we use Lemma 5.2 to get rid of R_A. The linear term is slightly more tedious to analyze. Recall that $\mathbb{E}\left[h_A(X)\right]=0$. Thanks to Jensen's inequality,

$$\begin{aligned}\mathrm{Var}\left(h_A(X)\right)&=\mathbb{E}_X\left[\left(\mathbb{E}_{X'}\left[\mathbb{1}_{m<\|X-X'\|^2\leq m+\frac{t}{\sqrt{n}}}\right]^2\right)\right]\\&\leq\mathbb{P}\left(m<\|X-X'\|^2\leq m+\frac{t}{\sqrt{n}}\right).\end{aligned}$$

We recognize

$$F\left(m+\frac{t}{\sqrt{n}}\right)-F\left(m\right),$$

which goes to 0 when $n\to\infty$, since we assumed F to have a derivative in m. Furthermore, by independence of the $X_{n,i}$,

$$\mathrm{Var}\left(\sqrt{n}L_A\right)=4\,\mathrm{Var}\left(h_A(X)\right)\underset{n\to\infty}{\longrightarrow}0.$$

A similar reasoning applies to the other terms in Eq. (5.13), and the proof is concluded.

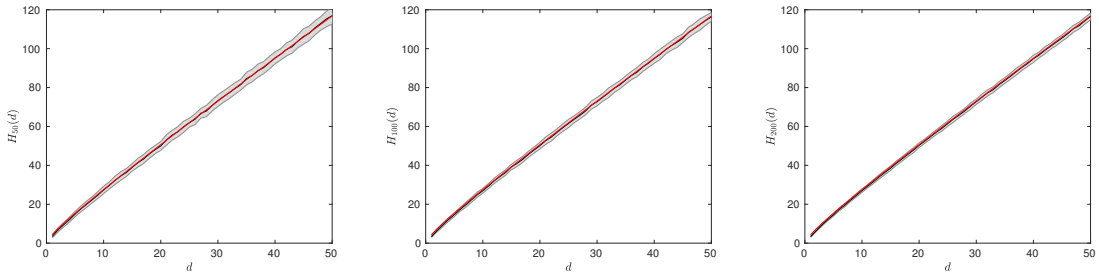

Figure 5-3 – A plot of $H_n(d)$ for $X \sim \mathcal{N}(0, \mathrm{I}_d)$ and $Y \sim \mathcal{N}(10d\,\mathbb{1}, \mathrm{I}_d)$ for $n \in \{50, 100, 200\}$. For each d, we repeated the experiment 10^2 times. The shaded area is within standard deviation from these experiments. We plot in red the approximation given by Eq. (5.14).

5.4 An example

In this section, we investigate the consequences of Prop. 5.2 when P and Q are known to be multivariate Gaussian distributions.

Before, let us recall that a sum of d independent squared standard Gaussian random variables is said to follow the *chi-squared* distribution with d degrees of freedom, denoted by χ^2_d. We also define the *non-central* chi-squared central distribution with the law of $\sum_{i=1}^d \mathcal{N}(\mu_i, 1)^2$, with *non-centrality* parameter $\lambda = \sum_{i=1}^d \mu_i^2$.

Now suppose that $X \sim \mathcal{N}(0, \mathrm{I}_d)$ and $Y \sim \mathcal{N}(\mu, \mathrm{I}_d)$ with $\mu \in \mathbb{R}^d$. This is the situation depicted in Fig. 5-1. We choose to tackle the unit variance case for convenience, as everything that follows can be adapted for a covariance matrix $\Sigma = \gamma\,\mathrm{I}_d$. With the notations introduced in Section 5.3, a quick computation shows that

$$\begin{cases} T_{XX} = 2\chi^2_d, \\ T_{XY} = 2\chi^2_d(\lambda), \qquad \text{with} \quad \lambda = \frac{1}{2}\|\mu\|^2, \\ T_{YY} = 2\chi^2_d. \end{cases}$$

According to Prop. 5.2, H_n is close to $m = \mathrm{med}(T)$ for large n. Can we hope for some insights on the value of m? The following proposition shows that it is possible in the high-dimensional regime.

Proposition 5.3. *Let T be as before. Take $\alpha \in (0, 1/2)$. Suppose that $\|\mu\|^2 = \mathrm{O}(d)$ when d goes to infinity. Define*

$$\kappa_\alpha = 2\sqrt{2}\Phi^{-1}\left(\frac{1}{2\left(\alpha^2 + (1-\alpha)^2\right)}\right) > 0\,,$$

where Φ is the repartition function of the standard Gaussian random variable. Then

$$m = 2d + \kappa_\alpha\sqrt{d} + \mathrm{o}\left(\sqrt{d}\right)\,, \tag{5.14}$$

when $d \to +\infty$.

The accuracy of Eq. (5.14) is illustrated in Fig. 5-3. We observe that, even though Prop. 5.3 is an asymptotic statement, it is valid with good accuracy for small values of d. Furthermore, we see the variance (represented by the shaded area) thinning when the number of data-points increases, as predicted by Prop. 5.2.

We make a few comments.

At first sight, it appears that m behaves as if only T_{XX} and T_{YY} contribute, in the sense that Eq. (5.14) does not depend on μ. There is a simple explanation to this observation. Indeed, a careful inspection of Fig. 5-1 shows that the left part of the histogram, corresponding to the contribution of T_{XX} and T_{YY}, is much larger than the right part, corresponding to T_{XY}. It holds since, for any $\alpha \in (0, 1)$,

$$\alpha^2 + (1-\alpha)^2 \geq 2\alpha(1-\alpha)\,.$$

Thus if $\mathbb{E}\,[T_{XX}]$ and $\mathbb{E}\,[T_{YY}]$ are both small with respect to $\mathbb{E}\,[T_{XY}]$, this is a general phenomenon and H_n will be close to $\max\left(\mathbb{E}\,[T_{XX}]\,, \mathbb{E}\,[T_{YY}]\right)$. Hence the median heuristic will select a bandwidth according to the maximum variance, since $\mathbb{E}\,[T_{XX}] = 2\,\mathrm{Var}\,(X)$ and $\mathbb{E}\,[T_{YY}] = 2\,\mathrm{Var}\,(Y)$. As noted in Gretton et al. [2012b, Sec. 5], if the variance of X and Y is much higher than the scale of the changes one aims to detect, the median heuristic will thus fail completely to select an appropriate bandwidth.

In this special case, we recover the setting of Reddi et al. [2015, Sec. 4.1 (A)] as in Sec. 5.1.1, with the exception that $\|\mu\|^2 = \mathrm{O}\,(d)$. Following the reasoning in Sec. 5.1.1, we would write

$$\mathbb{E}\,T = 2d + 2\alpha(1-\alpha)\,\|\mu\|^2\,,$$

which is no longer dominated by the first term as d increases. Yet, according to Prop. 5.3, $H_n \sim 2d$. We have thus obtained a rigorous and more precise result, though we acknowledge that the order of magnitude of H_n stays the same, that is, $H_n = \mathrm{O}\,(d)$.

5.4.1 Proof of Prop. 5.3

We are going to show that

$$\mathbb{P}\,(T \leq m) \xrightarrow[d\to\infty]{} \frac{1}{2}\,,$$

and our claim will follow.

We first note that $\mathbb{P}\,(T \leq m)$ can be decomposed as

$$\left(\alpha^2 + (1-\alpha)^2\right)\mathbb{P}\left(2\chi^2_d \leq m\right) + 2\alpha(1-\alpha)\,\mathbb{P}\left(2\chi^2_d(\lambda) \leq m\right)\,. \tag{5.15}$$

Let us show that, with the prescribed choice of κ_α, the left-hand side of Eq. (5.15) converges to $1/2$ and (5.11) converges to 0.

First note that, by the definition of m,

$$\begin{aligned}\mathbb{P}\left(2\chi^2_d \leq m\right) &= \mathbb{P}\left(\chi^2_d \leq d + \frac{\kappa_\alpha}{2}\sqrt{d}\right) \\ &= \mathbb{P}\left(\frac{\chi^2_d - d}{\sqrt{2d}} \leq \frac{\kappa_\alpha}{2\sqrt{2}}\right) .\end{aligned}$$

A direct application of Lemma 5.3 (see Section 5.6) yields,

$$\mathbb{P}\left(\frac{\chi^2_d - d}{\sqrt{2d}} \leq \frac{\kappa_\alpha}{2\sqrt{2}}\right) \to \Phi\left(\frac{\kappa_\alpha}{2\sqrt{2}}\right) ,$$

which is exactly $1/(2(\alpha^2 + (1-\alpha)^2))$ according to the definition of κ_α.

Now we turn to Eq. (5.11). By the same manipulations, we obtain

$$\mathbb{P}\left(2\chi^2_d(\lambda) \leq m\right) = \mathbb{P}\left(\frac{\chi^2_d(\lambda) - (d+\lambda)}{\sqrt{2(d+\lambda)}} \leq \frac{\frac{\kappa_\alpha}{2}\sqrt{d} - \lambda}{\sqrt{2(d+2\lambda)}}\right) .$$

Since $\lambda = \mathrm{O}\left(d\right)$, clearly

$$\frac{\frac{\kappa_\alpha}{2}\sqrt{d} - \lambda}{\sqrt{2(d+2\lambda)}} \to -\infty .$$

According to Lemma 5.3,

$$\frac{\chi^2_d(\lambda) - (d+\lambda)}{\sqrt{2(d+\lambda)}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0,1\right) ,$$

and we can conclude. □

Remark 5.4. It is also possible to prove a version of Prop. 5.3 for an arbitrary number of segments. With the notations of Remark 5.1, suppose that μ_p is the average on segment p. Assume $\|\mu_p - \mu_q\| = \lambda_{p,q} d$ and order the $\lambda_{p,q}$ in increasing order, and set $a_0 = \sum_p \alpha_p^2$ and $a_i = 2\alpha_{p_i}\alpha_{q_i}$. The median will then be close to $\lambda_{p_{i^\star},q_{i^\star}} d$, where $i^\star$ is such that $a_0 + \cdots + a_{i^\star} > 1/2$.

5.5 Conclusion and future directions

In this chapter, we partly explained the behavior of the median heuristic for a large sample size. We believe that it opens the door to more rigorous statements regarding the optimality of bandwidth choice in kernel two-sample test and kernel change-point detection, at least for some specific distributions.

As a future direction for research, we believe that it would be interesting to obtain a non-asymptotic version of Prop. 5.2. Indeed, as it is often the case in kernel methods, both kernel two-sample test and kernel change-point detection run in quadratic time – even though linear time approximations are available. Hence these methods, and consequently the median heuristic, are frequently used with a sample size that does

not exceed a few hundreds.

We would also like to improve the results given in Prop. 5.3. Namely, a version of Prop. 5.3 with non-identity covariance matrices for P and Q seems out of reach at the moment. Indeed, obtaining asymptotic behavior for T_{XX}, T_{YY} and T_{XY} is much harder. Extending Prop. 5.3 to the case where P and Q are mixtures of Gaussian distributions with non-identity covariance matrices could yield some precious insights on situations where the median heuristic is known to fail empirically [Gretton et al., 2012b, Fig. 1].

5.6 Additional proofs

In this section, we state and prove the technical results that are needed in the proofs of this chapter. Recall that we denote by $|A|$ the cardinality of any finite set A.

Lemma 5.1. *Let A_n, B_n and C_n be defined as in the proof of Prop. 5.1. Then* $\mathrm{Var}\,(A_n)$, $\mathrm{Var}\,(B_n)$ *and* $\mathrm{Var}\,(C_n)$ *are* $\mathrm{O}\,(n^{-1})$.

The proof is standard in U-statistics [Lee, 1990].

Proof. We set $m = \alpha n$ in this proof. Recall that

$$A_n = \binom{m}{2}^{-1} \sum_{1 \leq i < j \leq m} h(X_{n,i}, X_{n,j})\,,$$

Define $h_{i,j} = h(X_{n,i}, X_{n,j})$, thus $\mathbb{E}\, h_{i,j} = \theta_A$ and $\mathbb{E}\, A_n = \theta_A$. Let us turn to the computation of $\mathbb{E}\,[A_n^2]$, that is,

$$\mathbb{E}\left[A_n^2\right] = \binom{m}{2}^{-2} \sum_{a<b} \sum_{c<d} \mathbb{E}\left[h_{a,b} h_{c,d}\right]\,,$$

where a, b, c and d range from 1 to m. There are three separate cases in the sum that we detail below.

— The indices a, b, c and d are all distincts. There are $\binom{m}{4}\binom{4}{2} = \frac{m^4}{4} + \mathrm{O}\,(m^3)$ ways to choose such indices, that is, $\binom{m}{4}$ ways to choose the location of the 4 indices among the m possible locations, then $\binom{4}{2}$ choices for, say, $a < b$, and only one possibility left for $c < d$.

— One of the indices is common, that is, $|\{a, b\} \cap \{c, d\}| = 1$. There are $6\binom{m}{3} = \mathrm{O}\,(m^3)$ ways to do so.

— Both indices are equal, that is, $a = c$ and $b = d$. There are $\binom{m}{2} = \mathrm{O}\,(m^2)$ ways to do so.

Note that when a, b, c and d are all distinct, $\mathbb{E}\left[h_{a,b}h_{c,d}\right] = \mathbb{E}\left[h_{a,b}\right]\mathbb{E}\left[h_{c,d}\right]$ by independence of the $X_{n,i}$s. Thus

$$\mathbb{E}\left[A_n^2\right] = \binom{m}{2}^{-2} \sum_{\substack{a<b \\ c<d \\ \neq}} \mathbb{E}\left[h_{a,b}\right]\mathbb{E}\left[h_{c,d}\right] + \mathrm{O}\left(m^{-1}\right) ,$$

where the summation is on distinct indices. On the other hand,

$$\mathbb{E}\left[A_n\right]^2 = \binom{m}{2}^{-2} \sum_{\substack{a<b \\ c<d}} \mathbb{E}\, h_{a,b}\, \mathbb{E}\, h_{c,d} .$$

By the same combinatorial argument, the terms corresponding to intersecting sets of indices are at most $\mathrm{O}\left(n^3\right)$ and we have

$$\mathbb{E}\left[A_n\right]^2 = \binom{m}{2}^{-2} \sum_{\substack{a<b \\ c<d \\ \neq}} \mathbb{E}\, h_{a,b}\, \mathbb{E}\, h_{c,d} + \mathrm{O}\left(m^{-1}\right) .$$

Since $m = \alpha n$, $\mathrm{O}\left(m^{-1}\right) = \mathrm{O}\left(n^{-1}\right)$ and we can conclude for A_n:

$$\mathrm{Var}\left(A_n\right) = \mathbb{E}\left[A_n^2\right] - \mathbb{E}\left[A_n\right]^2 = \mathrm{O}\left(n^{-1}\right) .$$

The same proof transfers readily for B_n and C_n. □

Lemma 5.2. *Let R_A, R_B and R_C be as in the proof of Prop. 5.2. Then* $\mathrm{Var}\left(R_A\right)$, $\mathrm{Var}\left(R_B\right)$ *and* $\mathrm{Var}\left(R_C\right)$ *are of order* $\mathrm{O}\left(n^{-2}\right)$.

Proof. Recall that

$$R_A = \binom{m}{2}^{-1} \sum_{1\le i<j\le m} g_{i,j} ,$$

with $g_{i,j} = g_A(X_{n,i}, X_{n,j})$. By definition of g_A, it holds that $\mathbb{E}\, g_{i,j} = 0$ for any $i \neq j$, thus $\mathbb{E}\, R_A = 0$. Hence to control the variance of R_A, we just need to compute $\mathbb{E}\left[R_A^2\right]$. As in the proof of Lemma 5.1, we have

$$\mathbb{E}\left[R_A^2\right] = \binom{m}{2}^{-2} \sum_{a<b} \sum_{c<d} \mathbb{E}\left[g_{a,b}g_{c,d}\right] .$$

Note that $\mathbb{E}\left[g_{a,b}g_{c,d}\right] = 0$ whenever a, b, c and d are all distinct. But a straightforward computation also shows that $\mathbb{E}\left[g_{a,b}g_{a,c}\right] = 0$ for any distinct a, b, c. Thus the previous display reduces to

$$\mathbb{E}\left[R_A^2\right] = \binom{m}{2}^{-2} \sum_{\substack{a<b \\ c<d \\ \star}} \mathbb{E}\left[g_{a,b}g_{c,d}\right] ,$$

where $\star$ denotes that we sum on indices such that $|\{a,b\} \cap \{c,d\}| \geq 2$. As we have seen in the proof of Lemma 5.1, there are only $\mathrm{O}(m^2)$ such possibilities, and we can conclude. □

Our last result is a central limit theorem for a non-central chi-squared distributed random variable.

Lemma 5.3. *Let Y be a $\chi^2_d(\lambda)$ distributed random variable, with $\lambda > 0$ possibly depending on d. Then Y satisfies a central limit theorem, namely*

$$\frac{Y - (d+\lambda)}{\sqrt{2(d+2\lambda)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \,.$$

Proof. The characteristic function of Y is

$$t \mapsto \varphi_Y(t) = \frac{\exp\left(\frac{\mathrm{i}\lambda t}{1-2t}\right)}{(1-2\mathrm{i}t)^{d/2}} \,.$$

Define $S_d := (Y - (d+\lambda))/\sqrt{2(d+\lambda)}$. We show that $\varphi_{S_d}(t) \to \exp(-t^2/2)$, and conclude with Lévy's continuity's theorem.

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\mathrm{i}t\frac{Y-(d+\lambda)}{\sqrt{2(d+2\lambda)}}\right)\right] &= \mathbb{E}\left[\exp\left(\mathrm{i}\frac{t}{\sqrt{2(d+2\lambda)}}\right)\right]\exp\left(\frac{-\mathrm{i}t(d+\lambda)}{\sqrt{2(d+2\lambda)}}\right) \\
&= \exp\left(\frac{\mathrm{i}\lambda\frac{t}{\sqrt{2(d+2\lambda)}}}{1-\frac{2\mathrm{i}t}{\sqrt{2(d+2\lambda)}}}\right)\cdot\exp\left(\frac{-\mathrm{i}t(d+\lambda)}{\sqrt{2(d+2\lambda)}}\right) \\
&\qquad\cdot\left(1-2\mathrm{i}\frac{t}{\sqrt{2(d+2\lambda)}}\right)^{-d/2} \\
&= \exp\left(\frac{\mathrm{i}\lambda t}{\sqrt{2(d+2\lambda)}}\cdot\frac{1}{1-\frac{2\mathrm{i}t}{\sqrt{2(d+2\lambda)}}} - \frac{\mathrm{i}t(d+\lambda)}{\sqrt{2(d+2\lambda)}}\right. \\
&\qquad\left.-\frac{d}{2}\log\left(1-2\mathrm{i}\frac{t}{\sqrt{2(d+2\lambda)}}\right)\right) \\
&= \exp\left(\frac{\mathrm{i}\lambda t}{\sqrt{2(d+2\lambda)}}\left(1+\frac{2\mathrm{i}t}{\sqrt{2(d+2\lambda)}}\right) - \frac{\mathrm{i}t(d+\lambda)}{\sqrt{2(d+2\lambda)}}\right. \\
&\qquad\left.+\frac{d}{\cancel{2}}\frac{\cancel{2}\mathrm{i}t}{\sqrt{2(d+2\lambda)}} + \frac{d}{\cancel{2}}\frac{1}{\cancel{2}}\frac{-\cancel{4}t^2}{2(d+2\lambda)} + \mathrm{o}\left(t^2\right)\right) \\
&= \exp\left(\frac{\mathrm{i}\lambda t}{\cancel{\sqrt{2(d+2\lambda)}}} - \frac{\cancel{2}\lambda t^2}{\cancel{2}(d+2\lambda)} - \frac{\mathrm{i}t\lambda}{\cancel{\sqrt{2(d+2\lambda)}}}\right. \\
&\qquad\left.-\frac{dt^2}{2(d+2\lambda)} + \mathrm{o}\left(t^2\right)\right)
\end{aligned}$$

$$= \exp\left(-t^2 + \mathrm{o}\left(t^2\right)\right) .$$

□

Chapter 6

Conclusion and future work

6.1 Summary of the thesis

In this thesis, we mostly focused on a method for detecting abrupt changes in a sequence of independent observations belonging to an arbitrary set $\mathcal{X}$ on which a positive semi-definite kernel k is defined. That method, kernel change-point detection, is a kernelized version of a penalized least-squares procedure.

Our main contribution is to show that, for any kernel satisfying some reasonably mild hypotheses, the KCP procedure outputs a segmentation close to the true segmentation with high probability. This result is obtained under a bounded assumption on the kernel, in Theorem 3.1 for a linear penalty, and Theorem 3.2 for another penalty function, coming from model selection.

The proofs rely on a concentration result for bounded random variables in Hilbert spaces, and we prove a less powerful result under relaxed hypotheses — a finite variance assumption — in Theorem 3.3.

Up to now, it seemed difficult to combine the "change-point estimation / asymptotic" with the "model selection / non-asymptotic" point of view. The proofs of Theorems 3.1 and 3.3 show how they can be reconciled. Moreover, the structure of these proofs is modular, so that one can easily adapt them to different sets of assumptions.

In the asymptotic setting, we show that we recover the minimax rate $\log(n)/n$ for the change-point locations without additional hypothesis on the segment sizes. We provide empirical evidence supporting these claims.

Another contribution of this thesis is the detailed presentation of the different notions of distances between segmentations. Additionally, we prove a result showing these different notions coincide for sufficiently close segmentations.

From a practical point of view, a contribution of this thesis is to demonstrate how the so-called dimension jump heuristic can be a reasonable choice of penalty constant when using kernel change-point detection with a linear penalty.

We also show how a key quantity depending on the kernel that appears in our theoretical results, $\underline{\Delta}$, influences the performance of KCP in the case of a single change-point. When the kernel is translation-invariant and parametric assumptions

are made, it is possible to compute $\underline{\Delta}$ in closed-form. Thanks to these computations, some of them novel, we are able to study precisely the link between the maximal penalty constant and $\underline{\Delta}$. We show that, as suggested by Theorem 3.1, our theoretical upper bound on the penalty constant is proportional to $\underline{\Delta}^2$ all other things being equal.

Our last main contribution is a study of the median heuristic, a popular tool to set the bandwidth of radial basis function kernels. For a large sample size, we show that the median heuristic behaves approximately as the median of a distribution that we describe completely in the setting of kernel two-sample test and kernel change-point detection. More precisely, we show that the median heuristic is asymptotically normal around this value.

6.2 Perspectives

A number of questions remain unanswered at the end of this manuscript.

— In our theoretical study, we essentially provide an upper bound for $C_{\max}$, that is, $C_{\max} \lesssim \underline{\Delta}^2$ with high probability, up to constants that do not depend on the kernel. We believe that a lower bound exists. Proving this lower bound would achieve two goals. First, it would explain our empirical findings, and second, it would provide an important argument in favor of keeping up the study of $\underline{\Delta}$ as a criterion for kernel choice.

— In all our main results, we impose that $\widehat{D} = D^\star$. This may be a little too demanding, and as such it translates into very tight bounds on the possible constants for the penalty functions. Maybe it would be preferable not to work on an event such that $\widehat{D} = D^\star$, but rather such that $\left|\widehat{D} - D^\star\right| \leq \eta$, with η a threshold. As another technical improvement, we believe that it is possible to take advantage of the structure of our proofs to account for dependency between observations — a very natural hypothesis when dealing with time series. The only missing piece is a concentration inequality for dependent Hilbert-valued random variables.

— In the main concentration result used in our proofs, the deviations are of order M^2, which yields the $\underline{\Delta}^2/M^2$ term that intervenes in both the expression of $C_{\max}$ and the speed of convergence. We believe that this result is not optimal. In the real case, for instance, we know for a fact that M^2 can be replaced by a variance term. To prove a concentration result where the kernel bound is replaced by a variance term would provide a much more general version of Theorem 3.1. In particular, this new version would hold in the archetypal change-point problem — linear kernel, change in the mean of Gaussian observations —, which is not covered by the present result.

— The penalty function $\mathrm{pen}_{\mathrm{L}}$ showed its superiority for detecting the changes in $\mathbb{R}$ [Lebarbier, 2005] and is a natural choice of penalty function for KCP as well. So far, we have not managed to calibrate properly the constants of $\mathrm{pen}_{\mathrm{L}}$. In particular, in the real case, both these constants are of order σ^2 — the variance

of the data — for which we have no natural substitute in the kernel setting. Is it possible to provide a method for the calibration of the constants in $\mathrm{pen}_{\mathrm{L}}$ in the setting of KCP?

— The link uncovered between $\underline{\Delta}^2$ and the maximal penalty constant suggests a simple method of kernel choice for KCP: maximizing $\underline{\Delta}^2$ with respect to the kernel. Unfortunately, even in the simplest case, $\underline{\Delta}^2$ depends on unknown quantities, in particular the size of the jump we are trying to detect. Is it possible to estimate $\underline{\Delta}^2$ accurately, and thus to build a data-driven method for choosing the kernel for KCP?

— The results we obtained regarding the asymptotic behavior of the median heuristic are more precise than what was already known. It would be interesting to use this new insight in simple situations. For instance, suppose that we want to detect a single change in the mean of Gaussian observations. In this situation, is the median heuristic close to the global maximum of $\underline{\Delta}^2$?

— Finally, we think that the KCP algorithm can be applied successfully to real data in cases where d-dimensional features are not easily accessible but a positive semi-definite kernel is available. An example of such a situation is the study of granular material, for instance cereal grains in a silo. It is possible to summarize such an environment to the weighted graph of forces between particles, and even further to the persistence diagram associated to this graph — see Oudot [2015] for an introduction to persistence. The time-evolution of the granular material is then described by a sequence of such diagrams. The study of abrupt changes in the distribution of the forces between particles, as for instance in Gutiérrez et al. [2015], needs a change-point detection method for persistence diagrams. They are complicated objects, but positive semi-definite kernels do exist on the space of persistence diagrams [Kusano et al., 2016].

Bibliography

Abou-Elailah, A., Gouet-Brunet, V., and Bloch, I. (2015). Detection of abrupt changes in spatial relationships in video sequences. In *Pattern Recognition: Applications and Methods*, pp. 89–106. Springer.

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, **34**(2):584–653.

Aizerman, M. A., Braverman, E. M., and Rozonoer, L. I. (1964). Theoretical foundations of the potential function method in pattern recognition. *Automation and Remote Control*, **25**:917–936.

Allen, S., Madras, D., Ye, Y., and Zanotti, G. (2016). Change-point detection methods for body-worn video. Preprint. Available at `https://arxiv.org/abs/1610.06453v1`.

Arias-Castro, E., Candes, E. J., and Durand, A. (2011). Detection of an anomalous cluster in a network. *The Annals of Statistics*, **39**(1):278–304.

Arlot, S. (2007). *Resampling and model selection*. PhD thesis, Université Paris-Sud.

Arlot, S. and Celisse, A. (2011). Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, **21**(4):613–632.

Arlot, S., Celisse, A., and Harchaoui, Z. (2012). A kernel multiple change-point algorithm via model selection. Preprint. Available at `https://arxiv.org/abs/1202.3878v2`.

Aronszajn, N. (1943). La théorie générale des noyaux reproduisants et ses applications. *Proceedings of the Cambridge Philosophical Society*, **39**(3):133–153.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, **68**(3):337–404.

Bai, J. (1994). Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis*, **15**(5):453–472.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, **66**(1):47–78.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems*. PhD thesis, Australian National University, Camberra.

Basseville, M. (1981). Edge detection using sequential methods for change in level–Part II: Sequential detection of change in mean. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **29**(1):32–50.

Basseville, M. and Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application*, Prentice Hall Information and System Sciences Series. Prentice Hall, Englewood Cliffs (NJ).

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, **22**(2):455–470.

Bejjani, B. A. and Shaffer, L. G. (2006). Application of array-based comparative genomic hybridization to clinical diagnostics. *The Journal of Molecular Diagnostics*, **8**(5):528–533.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. *Communications of the ACM*, **4**(6):284.

Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic analysis on semigroups*, Graduate Texts in Mathematics, **100**. Springer-Verlag.

Bernstein, S. (1924). On a modification of chebyshev's inequality and of the error formula of laplace. *Annales Scientifiques des Institutions Mathématiques Savantes de l'Ukraine*, **1**(4):38–49.

Billingsley, P. (2012). *Probability and Measure*, Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken (NJ), third edition.

Birgé, L. and Massart, P. (2001). Gaussian model selection. *Journal of the European Mathematical Society*, **3**(3):203–268.

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, **138**(1-2):33–73.

Bleakley, K. and Vert, J.-P. (2011). The group fused lasso for multiple change-point detection. Preprint. Available at `https://arxiv.org/abs/1106.4199v1.pdf`.

Bochner, S. (1932). *Vorlesungen über Fouriersche Integrale*, Mathematik und ihre Anwendungen, **12**. Leipzig Akademische Verlagsgesellschaft.

Bochner, S. (1933). Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, **108**(1):378–410.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual workshop on Computational Learning Theory*, pp. 144–152. ACM Press.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

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

Brodsky, B. E. and Darkhovsky, B. S. (2013). *Nonparametric methods in change point problems*, Mathematics and Its Applications, **243**. Springer.

Brunel, V.-E. (2014). Convex set detection. Preprint. Available at `https://arxiv.org/abs/1404.6224v1`.

Camp, B. H. (1922). A new generalization of Tchebycheff's statistical inequality. *Bulletin of the American Mathematical Society*, **28**(9):427–432.

Carlstein, E. (1988). Nonparametric change-point estimation. *The Annals of Statistics*, **16**(1):188–197.

Celisse, A., Marot, G., Pierre-Jean, M., and Rigaill, G. (2016). New efficient algorithms for multiple change-point detection with kernels. Preprint. Available at `https://hal.inria.fr/hal-01413230`.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time. *The Annals of Mathematical Statistics*, **35**(3):999–1018.

Comte, F. and Rozenholc, Y. (2004). A new algorithm for fixed design regression and denoising. *Annals of the Institute of Statistical Mathematics*, **56**(3):449–473.

Cotsaces, C., Nikolaidis, N., and Pitas, I. (2006). Video shot boundary detection and condensed representation: a review. *IEEE Signal Processing Magazine*, **23**(2): 28–37.

Dehling, H. and Fried, R. (2012). Asymptotic distribution of two-sample empirical U-quantiles with applications to robust tests for shifts in location. *Journal of Multivariate Analysis*, **105**(1):124–140.

Deshayes, J. and Picard, D. (1985). Off-line statistical analysis of change-point models using non parametric and likelihood methods. In *Detection of Abrupt Changes in Signals and Dynamical Systems*, pp. 103–168. Springer.

Desobry, F., Davy, M., and Doncarli, C. (2005). An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, **53**(8):2961–2974.

Diestel, J. and Uhl, J. J. (1977). *Vector measures*, Mathematical Surveys and Monographs, **15**. American Mathematical Society.

Dieuleveut, A. and Bach, F. R. (2016). Nonparametric stochastic approximation with large step-sizes. *The Annals of Statistics*, **44**(4):1363–1399.

Fearnhead, P. and Liu, Z. (2007). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **69**(4): 589–605.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Association*, **53**(284):789–798.

Flaxman, S., Sejdinovic, D., Cunningham, J. P., and Filippi, S. (2016). Bayesian learning of kernel embeddings. In *Proceedings of the 32th Conference on Uncertainty in Artificial Intelligence*, pp. 182–191.

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

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, **5**:73–99.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2008). Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pp. 489–496. Curran Associates, Inc.

Gardner, L. A. (1969). On detecting changes in the mean of normal variates. *The Annals of Mathematical Statistics*, **40**(1):116–126.

Garreau, D. (2017). Asymptotic normality of the median heuristic. Preprint. Available at `https://arxiv.org/abs/1707.07269`.

Garreau, D. and Arlot, S. (2016). Consistent change-point detection with kernels. Preprint. Available at `http://arxiv.org/abs/1612.04740v3`.

Gärtner, T., Flach, P., and Wrobel, S. (2003). On graph kernels: Hardness results and efficient alternatives. In *Proceedings of the 16th Conference on Learning Theory and 7th Kernel Workshop*, pp. 129–143. Springer.

Genton, M. G. (2001). Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, **2**(12):299–312.

Giné, E. and Nickl, R. (2015). *Mathematical foundations of infinite-dimensional statistical models*, Cambridge Series in Statistical and Probabilistic Mathematics, **40**. Cambridge University Press.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19*, pp. 513–520. MIT Press.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012a). A kernel two-sample test. *Journal of Machine Learning Research*, **13**(1):723–773.

Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. (2012b). Optimal kernel choice for large-scale two-sample tests. In *Advances in Neural Information Processing Systems 25*, pp. 1205–1213. Curran Associates, Inc.

Gutiérrez, G., Colonnello, C., Boltenhagen, P., Darias, J. R., Peralta-Fabi, R., Brau, F., and Clément, E. (2015). Silo collapse under granular discharge. *Physical Review Letters*, **114**(1):018001.

Hájek, J. and Rényi, A. (1955). Generalization of an inequality of Kolmogorov. *Acta Mathematica Hungarica*, **6**(3-4):281–283.

Hall, P., Kay, J. W., and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**(3):521–528.

Harchaoui, Z. and Cappé, O. (2007). Retrospective multiple change-point estimation with kernels. In *Proceedings of the 14th IEEE Workshop on Statistical Signal Processing*, pp. 768–772.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, **105**(492): 1480–1493.

Harchaoui, Z., Moulines, E., and Bach, F. R. (2009). Kernel change-point analysis. In *Advances in Neural Information Processing Systems 21*, pp. 609–616. Curran Associates, Inc.

Hawkins, D. M. (1977). Testing a sequence of observations for a shift in location. *Journal of the American Statistical Association*, **72**(357):180–186.

He, H. and Severini, T. A. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli*, **16**(3):759–779.

Hilbert, D. (1904). Grundzüge einer allgemeinen Theorie der linaren Integralrechnungen. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse*, pp. 49–91.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika*, **56**(3):495–504.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, **57**(1):1–17.

Hinkley, D. V. (1971). Inference about the change-point from cumulative sum tests. *Biometrika*, **58**(3):509–523.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, **19**(3):293–325.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1):55–67.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2** (1):193–218.

James, B., James, K. L., and Siegmund, D. (1987). Tests for a change-point. *Biometrika*, **74**(1):71–83.

Jammalamadaka, S. R. and Janson, S. (1986). Limit theorems for a triangular scheme of U-statistics with applications to inter-point distances. *The Annals of Probability*, pp. 1347–1358.

Jitkrittum, W., Szabó, Z., and Gretton, A. (2016). An adaptive test of independence with analytic kernel embeddings. Preprint. Available at `http://arxiv.org/abs/1610.04782v1`.

Kallioniemi, A., Kallioniemi, O.-P., Sudar, D., Rutovitz, D., Gray, J. W., Waldman, F., and Pinkel, D. (1992). Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science*, **258**(5083):818–821.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, **11**(9):1–20.

Kim, A. Y., Marzban, C., Percival, D. B., and Stuetzle, W. (2009). Using labeled data to evaluate change detectors in a multivariate streaming environment. *Signal Processing*, **89**(12):2529–2536.

Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, **33**(1):82–95.

Kolmogorov, A. N. (1928). Über die Summen durch den Zufall bestimmten unabhängigen größen. *Mathematische Annalen*, **99**:484–488.

Kolmogorov, A. N. (1941). Stationary sequences in Hilbert space. *Bulletin of the Moscow State University*, **2**(6).

Korolyuk, V. S. and Borovskich, Y. V. (2013). *Theory of U-statistics*, Mathematics and Its Applications, **273**. Springer Science & Business Media.

Korostelev, A. P. (1988). On minimax estimation of a discontinuous signal. *Theory of Probability and its Applications*, **32**(4):727–730.

Korostelev, A. P. and Tsybakov, A. B. (2012). *Minimax theory of image reconstruction*, Lecture Notes in Statistics, **82**. Springer Science & Business Media.

Kusano, G., Hiraoka, Y., and Fukumizu, K. (2016). Persistence weighted gaussian kernel for topological data analysis. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 2004–2013.

Lai, T. L. (1974). Control charts based on weighted sums. *The Annals of Statistics*, **2**(1):134–147.

Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. *Bioinformatics*, **21**(19):3763–3770.

Lajugie, R., Arlot, S., and Bach, F. R. (2014). Large-margin metric learning for constrained partitioning problems. In *Proceedings of The 31st International Conference on Machine Learning*, pp. 297–305.

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, **83**(1):79–102.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem. *Signal Processing*, **85**(8):1501–1510.

Lavielle, M. and Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, **21**(1):33–59.

Lavielle, M. and Teyssiere, G. (2006). Detection of multiple change-points in multivariate time series. *Lithuanian Mathematical Journal*, **46**(3):287–306.

Lebarbier, É. (2002). *Quelques approches pour la détection de ruptures à horizon fini*. PhD thesis, Université Paris 11.

Lebarbier, É. (2005). Detecting multiple change-points in the mean of a gaussian process by model selection. *Signal Processing*, **85**(4):717–736.

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*, Classics in Mathematics, **23**. Springer Science & Business Media.

Lee, C.-B. (1995). Estimating the number of change points in a sequence of independent normal random variables. *Statistics & probability letters*, **25**(3):241–248.

Lee, C.-B. (1997). Estimating the number of change points in exponential families distributions. *Scandinavian Journal of Statistics*, **24**(2):201–210.

Lee, J. (1990). *U-statistics: Theory and Practice*, Statistics: Textbooks and Monographs, **110**. Marcel Dekker, Inc.

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *The Annals of Mathematical Statistics*, **22**(2):165–179.

Li, S., Xie, Y., Dai, H., and Song, L. (2015). M-statistic for kernel change-point detection. In *Advances in Neural Information Processing Systems 28*, pp. 3366–3374. Curran Associates, Inc.

Lin, Y. and Zhang, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, **34**(5):2272–2297.

Liu, J., Wu, S., and Zidek, J. V. (1997). On segmented multivariate regression. *Statistica Sinica*, **7**(2):497–525.

Liu, S., Suzuki, T., Relator, R., Sese, J., Sugiyama, M., and Fukumizu, K. (2017). Support consistency of direct sparse-change learning in markov networks. *The Annals of Statistics*, **45**(3):959–990.

Lorden, G. (1971). Procedures for reacting to a change in distribution. *The Annals of Mathematical Statistics*, **45**(6):1897–1908.

Mallows, C. L. (1991). Another comment on O'Cinneide. *The American Statistician*, **45**(3):257.

Massart, P. (2007). *Concentration inequalities and model selection*, Lecture Notes in Mathematics, **1896**. Springer, Berlin.

Meidell, B. (1918). Note sur quelques inégalités et formules d'approximation. *Scandinavian Actuarial Journal*, **1918**(1):180–198.

Mercer, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A*, **209**:415–446.

Mikosch, T. and Starica, C. (2004). Changes of structure in financial time series and the garch model. *Revstat Statistical Journal*, **2**(1):41–73.

Moore, E. H. (1916). On properly positive Hermitian matrices. *Bulletin of the American Mathematical Society*, **23**(59):66–67.

Moore, E. H. (1935). General analysis (part I). *Memoirs of the American Philosophical Society*.

Moore, E. H. (1939). General analysis (part II). *Memoirs of the American Philosophical Society*.

Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *The Annals of Statistics*, **14**(4):1379–1387.

Muandet, K., Sriperumbudur, B., Fukumizu, K., Gretton, A., and Schölkopf, B. (2016). Kernel mean shrinkage estimators. *Journal of Machine Learning Research*, **17**(48):1–41.

Muandet, K., Fukumizu, K., Sriperumbudur, B., and Schölkopf, B. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, **10**(1-2):1–141.

Oudot, S. Y. (2015). *Persistence theory: from quiver representations to data analysis*, Mathematical Surveys and Monographs, **209**. American Mathematical Society.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41**(1–2):100–115.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**(3–4):523–527.

Page, E. S. (1957). On problems in which a change in a parameter occurs at an unknown point. *Biometrika*, **44**(1–2):248–252.

Pepelyshev, A. and Polunchenko, A. S. (2015). Real-time financial surveillance via quickest change-point detection methods. Preprint. Available at `https://arxiv.org/abs/1509.01570v2`.

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

Pinelis, I. F. and Sakhanenko, A. I. (1986). Remarks on inequalities for large deviation probabilities. *Theory of Probability and its Applications*, **30**(1):143–148.

Reddi, S. J., Ramdas, A., Poczos, B., Singh, A., and Wasserman, L. (2015). On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pp. 3571–3577.

Ritov, Y., Raz, A., and Bergman, H. (2002). Detection of onset of neuronal activity by allowing for heterogeneity in the change points. *Journal of neuroscience methods*, **122**(1):25–42.

Roberts, L. G. (1963). *Machine Perception of Three Dimensional Solids*. PhD thesis, Massachusetts Institute of Technology.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Adaptive Computation and Machine Learning. MIT Press, Cambridge (MA).

Schölkopf, B., Smola, A., and Müller, K.-R. (1997). Kernel principal component analysis. In *Proceedings of the 7th International Conference on Artificial Neural Networks*, pp. 583–588. Springer.

Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**(5):1299–1319.

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

Sen, A. and Srivastava, M. S. (1975a). On tests for detecting change in mean. *The Annals of Statistics*, **3**(1):98–108.

Sen, A. and Srivastava, M. S. (1975b). Some one-sided tests for change in level. *Technometrics*, **17**(1):61–64.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*, Wiley Series in Probability and Statistics, **162**. John Wiley & Sons.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, **7** (2):221–242.

Sharipov, O., Tewes, J., and Wendler, M. (2016). Sequential block bootstrap in a Hilbert space with application to change point analysis. *The Canadian Journal of Statistics. La Revue Canadienne de Statistique*, **44**(3):300–322.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. ASQ Quality Press.

Shiryaev, A. N. (1961). The problem of the most rapid detection of a disturbance in a stationary process. In *Soviet Mathematics Doklady*, pp. 795–799.

Shiryaev, A. N. (1963). On optimum methods in quickest detection problems. *Theory of Probability and its Applications*, **8**(1):22–46.

Shiryaev, A. N. (1965). Some exact formulas in a "disorder" problem. *Theory of Probability and its Applications*, **10**(2):348–354.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, A. K., Huey, B., and Kimura, K. (2001). Assembly of microarrays for genome-wide measurement of dna copy number. *Nature Genetics*, **29**(3):263.

Spokoiny, V. (2009). Multiscale local change point detection with applications to value-at-risk. *The Annals of Statistics*, **37**(3):1405–1436.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Lanckriet, G. R. G., and Schölkopf, B. (2009). Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems 22*, pp. 1750–1758. Curran Associates, Inc.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Lanckriet, G. R. G., and Schölkopf, B. (2008). Injective hilbert space embeddings of probability measures. In *COLT*, pp. 111–122.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. G. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, **11**:1517–1561.

Srivastava, M. S. and Worsley, K. J. (1986). Likelihood ratio tests for a change in the multivariate normal mean. *Journal of the American Statistical Association*, **81** (393):199–204.

Straumann, D. (2005). *Estimation in Conditionnally Heteroscedastic Time Series Models*, Lecture Notes in Statistics, **181**. Springer-Verlag Berlin Heidelberg, first edition.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *Proceedings of the 5th International Conference on Learning Representations*.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, **7**(7):1231–1264.

Tartakovsky, A., Nikiforov, I. V., and Basseville, M. (2014). *Sequential Analysis: Hypothesis Testing and Changepoint Detection*, Monographs on Statistics and Applied Probability, **136**. Chapman and Hall/CRC, Boca Raton (FL).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **58**(1):267–288.

Tsay, R. S. (2005). *Analysis of financial time series*, Wiley Series in Probability and Statistics, **543**. John Wiley & Sons.

van der Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge Series in Statistical and Probabilistic Mathematics, **3**. Cambridge University Press, Cambridge.

Vert, J.-P., Tsuda, K., and Schölkopf, B. (2004). A primer on kernel methods. In *Kernel Methods in Computational Biology*, pp. 35–70. MIT Press.

Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R., and Borgwardt, K. M. (2010). Graph kernels. *Journal of Machine Learning Research*, **11**:1201–1242.

Vogt, M. and Dette, H. (2015). Detecting gradual changes in locally stationary processes. *The Annals of Statistics*, **43**(2):713–740.

Vostrikova, L. Y. (1981). Detecting disorder in multidimensional random process. In *Soviet Mathematics Doklady*, pp. 55–59.

Wald, A. (1947). *Sequential analysis*. Dover Publications, Inc. (NY).

Wang, T. and Samworth, R. J. (2016). High-dimensional changepoint estimation via sparse projection. Preprint. Available at `https://arxiv.org/abs/1606.06246v2`.

Weiss, M. M., Kuipers, E. J., Meuwissen, S. G. M., van Diest, P. J., and Meijer, G. A. (2003). Comparative genomic hybridisation as a supportive tool in diagnostic pathology. *Journal of Clinical Pathology*, **56**(7):522–527.

Wendland, H. (2005). *Scattered data approximation*, Cambridge Monographs on Applied and Computational Mathematics, **17**. Cambridge University Press.

Worsley, K. J. (1979). On the likelihood ratio test for a shift in location of normal populations. *Journal of the American Statistical Association*, **74**(366a):365–367.

Worsley, K. J. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variables. *Biometrika*, **73**(1):91–104.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statistics & Probability Letters*, **6**(3):181–189.

Yao, Y.-C. and Au, S.-T. (1989). Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, **51**(2):370–381.

Zhang, Q., Filippi, S., Gretton, A., and Sejdinovic, D. (2017). Large-scale kernel methods for independence testing. *Statistics and Computing*, pp. 1–18.

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

List of Figures

Résumé

Dans cette thèse, nous nous intéressons à une méthode de détection des ruptures dans une suite d'observations appartenant à un ensemble muni d'un noyau semi-défini positif. Cette procédure est une version « à noyaux » d'une méthode des moindres carrés pénalisés.

Notre principale contribution est de montrer que, pour tout noyau satisfaisant des hypothèses raisonnables, cette méthode fournit une segmentation proche de la véritable segmentation avec grande probabilité. Ce résultat est obtenu pour un noyau borné et une pénalité linéaire, ainsi qu'une autre pénalité venant de la sélection de modèles. Les preuves reposent sur un résultat de concentration pour des variables aléatoires bornées à valeurs dans un espace de Hilbert, et nous obtenons une version moins précise de ce résultat lorsque l'on suppose seulement que la variance des observations est finie.
Dans un cadre asymptotique, nous retrouvons les taux minimax usuels en détection de ruptures lorsqu'aucune hypothèse n'est faite sur la taille des segments. Ces résultats théoriques sont confirmés par des simulations.
Nous étudions également de manière détaillée les liens entre différentes notions de distances entre segmentations. En particulier, nous prouvons que toutes ces notions coïncident pour des segmentations suffisamment proches.

D'un point de vue pratique, nous montrons que l'heuristique du « saut de dimension » pour choisir la constante de pénalisation est un choix raisonnable lorsque celle-ci est linéaire.
Nous montrons également qu'une quantité clé dépendant du noyau et qui apparaît dans nos résultats théoriques influe sur les performances de cette méthode pour la détection d'une unique rupture. Dans un cadre paramétrique, et lorsque le noyau utilisé est invariant par translation, il est possible de calculer cette quantité explicitement. Grâce à ces calculs, nouveaux pour plusieurs d'entre eux, nous sommes capable d'étudier précisément le comportement de la constante de pénalité maximale.

Pour finir, nous traitons de l'heuristique de la médiane, un moyen courant de choisir la largeur de bande des noyaux à base de fonctions radiales. Dans un cadre asymptotique, nous montrons que l'heuristique de la médiane se comporte à la limite comme la médiane d'une distribution que nous décrivons complètement dans le cadre du test à deux échantillons à noyaux et de la détection de ruptures. Plus précisément, nous montrons que l'heuristique de la médiane est approximativement normale centrée en cette valeur.

Mots Clés

Détection de ruptures, méthodes à noyaux, moindres carrés pénalisés, heuristique de la médiane.

Abstract

In this thesis, we focus on a method for detecting abrupt changes in a sequence of independent observations belonging to an arbitrary set on which a positive semi-definite kernel is defined. That method, kernel change-point detection, is a kernelized version of a penalized least-squares procedure.

Our main contribution is to show that, for any kernel satisfying some reasonably mild hypotheses, this procedure outputs a segmentation close to the true segmentation with high probability. This result is obtained under a bounded assumption on the kernel for a linear penalty and for another penalty function, coming from model selection.
The proofs rely on a concentration result for bounded random variables in Hilbert spaces and we prove a less powerful result under relaxed hypotheses — a finite variance assumption.
In the asymptotic setting, we show that we recover the minimax rate for the change-point locations without additional hypothesis on the segment sizes. We provide empirical evidence supporting these claims.
Another contribution of this thesis is the detailed presentation of the different notions of distances between segmentations. Additionally, we prove a result showing these different notions coincide for sufficiently close segmentations.

From a practical point of view, we demonstrate how the so-called dimension jump heuristic can be a reasonable choice of penalty constant when using kernel change-point detection with a linear penalty.
We also show how a key quantity depending on the kernel that appears in our theoretical results influences the performance of kernel change-point detection in the case of a single change-point. When the kernel is translation-invariant and parametric assumptions are made, it is possible to compute this quantity in closed-form. Thanks to these computations, some of them novel, we are able to study precisely the behavior of the maximal penalty constant.

Finally, we study the median heuristic, a popular tool to set the bandwidth of radial basis function kernels. For a large sample size, we show that it behaves approximately as the median of a distribution that we describe completely in the setting of kernel two-sample test and kernel change-point detection. More precisely, we show that the median heuristic is asymptotically normal around this value.

Keywords

Change-point detection, kernel methods, penalized least-squares, median heuristic.